ELECTRONIC DESIGNERS' HANDBOOK

LANDEE
DAVIS
ALBRECHT

McGRAW-HILL
BOOK COMPANY

ELECTRONIC DESIGNERS' HANDBOOK

McGRAW-HILL HANDBOOKS

ELECTRONIC DESIGNERS' HANDBOOK

ROBERT W. LANDEE

Director of Research and Development
Space and Airborne Data Systems, Western Division
Collins Radio Company, Burbank, California

DONOVAN C. DAVIS

Assistant Director of Engineering
Gilfillan Bros. Inc., Los Angeles, California

ALBERT P. ALBRECHT

General Manager
Space Electronics Corporation

ELECTRONIC DESIGNERS' HANDBOOK

VIII

36110

Preface

This is a handbook for students and practicing engineers in the electronic field and is intended to provide adequate technical discussions, design data, and basic design procedures for the solution of many design problems. It has been the authors' opinion that handbooks often have limited value since the presentation is frequently so concise that the material presented has little value unless the reader has had previous experience with the subject. Consequently, an attempt has been made to overcome this limitation by making the text as lucid as possible and by including design examples which illustrate the application of the material to specific design problems.

The "Electronic Designers' Handbook" contains over 1,100 figures, 1,400 equations, and 140 examples. In many cases it is believed that significant contributions have been made in the form of new design data and/or design techniques. Also, the degree to which frequently used circuits and practices have been integrated into the material should contribute appreciably to the value of the handbook. Specific examples of material which are of special interest include fundamentals of statistical and probability theory (Sec. 23), graphical methods as an aid in the determination of inverse Laplace transforms (Sec. 23), a comparatively complete treatment of the design of low-power iron-core transformers and chokes (Sec. 14), design data for the optimum number of RC and LC ladder types of filter sections (Sec. 15), graphical presentation of the amplitude and phase characteristics of RC parallel-T and RC bridged-T circuits (Sec. 16), graphical performance data for constant-k and m-derived filter sections having dissipation (Sec. 16), extensive coverage and analysis of both voltage and power amplifier fundamentals (Secs. 3 and 4), a detailed discussion of receivers and associated subjects such as noise figure and characteristics of different mixer types (Sec. 7), an extensive presentation of feedback fundamentals and the principal methods of stability analysis (Secs. 18 and 19), the analysis of complex waveforms (Sec. 22), etc.

As is the case with all handbooks of this type, it has been necessary to omit a large amount of valuable design material. The electronic field

has become so vast and is growing at such a rate that the authors of such publications can only include the material which they believe has the broadest application and then, with the publisher, attempt to attain a publication date before the material becomes either obsolescent or seriously limited in scope. For example, in the case of the "Electronic Designers' Handbook," we regret that we did not have sufficient time to include certain digital computer techniques, magnetic memory circuits, and a much broader treatment of transistor applications.

It is believed that the "Electronic Designers' Handbook" should find extensive use as a text even though the equation derivations are not always given. Before publication, this material was used as a text for undergraduate students, graduate students, and practicing engineers with very gratifying results. With reference to these classes, the assignments which included the derivation of equations often proved to be especially educational. In general, the material is intended to be self-explanatory and, considering the practical design examples, should be of value in teaching the student to take a practical approach to design problems.

To those who have contributed in the preparation of this handbook we are especially grateful. Specifically, we wish to single out the names of Messrs. Nathan Patrusky, Ray W. Sanders, Robert T. Johnson, Robert E. Hull, and Warren E. Wilke who enthusiastically participated in all aspects of this effort and who generously donated much of their time and made many significant contributions, suggestions, and constructive criticisms. We are equally appreciative of the many valuable suggestions and criticisms made by D. C. Arnold, S. Benson, Miss G. R. Brown, P. M. Brown, C. W. Chandler, D. J. Green, M. L. Ingalsbe, T. J. Johnson, E. S. Klotz, J. J. Nesler, G. M. Salamonovich, and A. G. Woolfries.

We also wish to express thanks to Mrs. Lois Van Allen, Miss Liana Pucelli, and Miss Barbara McArdell for their aid in typing and preparing many of the original figures. In addition, we shall be forever grateful for the indirect but real contribution made by our families and friends who patiently awaited the completion of this handbook.

It is our sincere hope that we have successfully prepared a combination handbook and text which will be of value to our fellow engineers.

<div style="text-align: right">

ROBERT W. LANDEE
DONOVAN C. DAVIS
ALBERT P. ALBRECHT

</div>

Contents

Index follows Section 23.

General Design Data

TABLE 1.1. DECIBEL TABLE

Db	Gain Voltage ratio	Gain Power ratio	Loss Voltage ratio	Loss Power ratio
0	1.00	1.00	1.00	1.00
0.1	1.01	1.02	0.989	0.977
0.2	1.02	1.05	0.977	0.955
0.3	1.04	1.07	0.966	0.933
0.4	1.05	1.10	0.955	0.912
0.5	1.06	1.12	0.944	0.891
0.6	1.07	1.15	0.933	0.871
0.7	1.08	1.17	0.923	0.851
0.8	1.10	1.20	0.912	0.832
0.9	1.11	1.23	0.902	0.813
1.0	1.12	1.26	0.891	0.794
1.1	1.14	1.29	0.881	0.776
1.2	1.15	1.32	0.871	0.759
1.3	1.16	1.35	0.861	0.741
1.4	1.17	1.38	0.851	0.724
1.5	1.19	1.41	0.841	0.708
1.6	1.20	1.45	0.832	0.692
1.7	1.22	1.48	0.822	0.676
1.8	1.23	1.51	0.813	0.661
1.9	1.24	1.55	0.804	0.646
2.0	1.26	1.58	0.794	0.631
2.2	1.29	1.66	0.776	0.603
2.4	1.32	1.74	0.759	0.575
2.6	1.35	1.82	0.741	0.550
2.8	1.38	1.91	0.724	0.525
3.0	1.41	2.00	0.708	0.501
3.2	1.45	2.09	0.692	0.479
3.4	1.48	2.19	0.676	0.457
3.6	1.51	2.29	0.661	0.436
3.8	1.55	2.40	0.646	0.417
4.0	1.58	2.51	0.631	0.398
4.2	1.62	2.63	0.617	0.380
4.4	1.66	2.75	0.603	0.363
4.6	1.70	2.88	0.589	0.347
4.8	1.74	3.02	0.575	0.331
5.0	1.78	3.16	0.562	0.316
5.5	1.88	3.55	0.531	0.282

Db	Gain Voltage ratio	Gain Power ratio	Loss Voltage ratio	Loss Power ratio
6.0	2.00	3.98	0.501	0.251
6.5	2.11	4.47	0.473	0.224
7.0	2.24	5.01	0.447	0.200
7.5	2.37	5.62	0.422	0.178
8.0	2.51	6.31	0.398	0.158
8.5	2.66	7.08	0.376	0.141
9.0	2.82	7.94	0.355	0.126
9.5	2.99	8.91	0.335	0.112
10.0	3.16	10.00	0.316	0.100
12.0	3.98	15.8	0.251	0.063
14.0	5.01	25.1	0.200	0.040
16.0	6.31	39.8	0.158	0.025
18.0	7.94	63.1	0.126	0.016
20.0	10.00	100.0	0.100	0.010
25.0	17.8	3.16×10^2	0.056	3.16×10^{-3}
30.0	31.6	10^3	0.032	10^{-3}
35.0	56.2	3.16×10^3	0.018	3.16×10^{-4}
40.0	100.0	10^4	0.010	10^{-4}
45.0	178.0	3.16×10^4	0.006	3.16×10^{-5}
50.0	316	10^5	0.003	10^{-5}
55.0	562	3.16×10^5	0.002	3.16×10^{-6}
60.0	1,000	10^6	0.001	10^{-6}
65.0	1,780	3.16×10^6	0.0006	3.16×10^{-7}
70.0	3,160	10^7	0.0003	10^{-7}
75.0	5,620	3.16×10^7	0.0002	3.16×10^{-8}
80.0	10,000	10^8	0.0001	10^{-8}
85.0	17,800	3.16×10^8	0.00006	3.16×10^{-9}
90.0	31,600	10^9	0.00003	10^{-9}
95.0	56,200	3.16×10^9	0.00002	3.16×10^{-10}
100.0	100,000	10^{10}	0.00001	10^{-10}
105.0	178,000	3.16×10^{10}	0.000006	3.16×10^{-11}
110.0	316,000	10^{11}	0.000003	10^{-11}
115.0	562,000	3.16×10^{11}	0.000002	3.16×10^{-12}
120.0	1,000,000	10^{12}	0.000001	10^{-12}
130.0	3.16×10^6	10^{13}	3.16×10^{-7}	10^{-13}
140.0	10^7	10^{14}	10^{-7}	10^{-14}
150.0	3.16×10^7	10^{15}	3.16×10^{-8}	10^{-15}

TABLE 1.2. RELATIONS BETWEEN UNITS
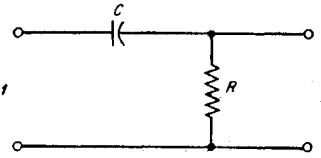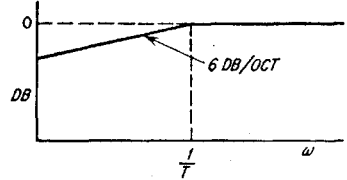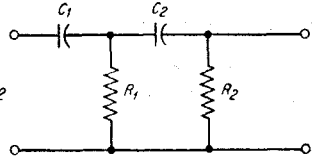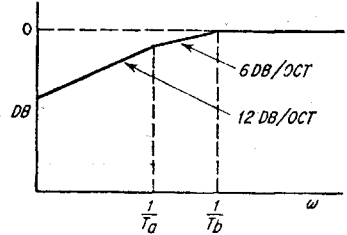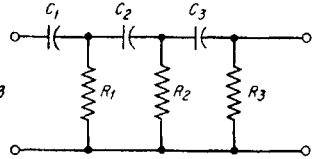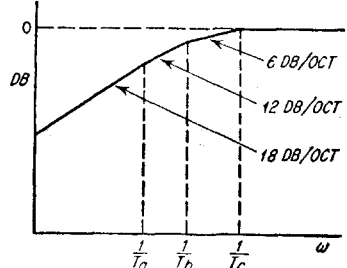
Quantity		Relation
Capacitance	1 farad	$= 10^{-9}$ abfarad
		$= 9 \times 10^{11}$ statfarads
Inductance	1 henry	$= 10^9$ abhenrys
		$= 1.11 \times 10^{-12}$ stathenry
Resistance	1 ohm	$= 10^9$ abohms
		$= 1.11 \times 10^{-12}$ statohm
Emf	1 volt	$= 10^8$ abvolts
		$= 3.33 \times 10^{-3}$ statvolt
Current	1 ampere	$= 10^{-1}$ abampere
		$= 3 \times 10^9$ statamperes
Charge	1 coulomb	$= 10^{-1}$ abcoulomb
		$= 3 \times 10^9$ statcoulombs
Energy	1 joule	$= 10^7$ ergs
Power	1 watt	$= 10^7$ ergs/second
		$= \frac{1}{746}$ horsepower

Energy stored in an inductor $= \dfrac{LI^2}{2}$ joules

Energy stored in a capacitor $= \dfrac{CV^2}{2}$ joules

where L is in henrys, C is in farads, I is in amperes, and V is in volts.

TABLE 1.3. WINDING DATA FOR RANDOM-WOUND COILS

B. & S. gauge	Turns per square inch				B. & S. gauge	Turns per square inch			
	Single enamel or single formvar	Double formvar	Single cotton-covered	Double cotton-covered		Single enamel or single formvar	Double formvar	Single cotton-covered	Double cotton-covered
8	56				25	2,580	2,380	1,890	1,360
9	70				26	3,250	2,970	2,280	1,600
10	88	85	82	75	27	4,060	3,670	2,720	1,840
11	111	107	104	95	28	5,140	4,520	3,250	2,130
12	139	134	129	117	29	6,380	5,540	3,810	2,420
13	173	168	160	143					
14	217	210	199	175	30	8,150	7,060	4,520	2,780
					31	10,100	8,780	5,290	3,140
15	272	262	246	214	32	12,600	10,800	6,080	3,490
16	341	328	305	261	33	16,000	13,500	7,060	3,900
17	428	409	375	316	34	20,000	16,900	8,150	4,340
18	536	511	463	383					
19	669	638	568	461	35	24,700	21,200	9,310	4,780
					36	31,400	26,400	11,700	5,620
20	832	794	694	552	37	39,600	32,600	13,150	6,080
21	1,040	989	847	658	38	49,100	41,200	14,800	6,600
22	1,300	1,240	1,030	808	39	62,500	53,900	16,900	7,180
23	1,640	1,530	1,250	951	40	77,600	65,800	18,800	7,711
24	2,100	1,910	1,510	1,120					

TABLE 1.4. *RC* TRANSFER FUNCTIONS

Network	Transfer function	Approx. amp. response (ω on log scale)
#1	$$\frac{s}{s + \dfrac{1}{T}}$$ where $T = RC$	
#2	$$\frac{s^2}{s^2 + Bs + D} = \frac{s^2}{\left(s + \dfrac{1}{T_a}\right)\left(s + \dfrac{1}{T_b}\right)}$$ where $B = \dfrac{R_1 C_1 + R_1 C_2 + R_2 C_2}{R_1 R_2 C_1 C_2}$ $D = \dfrac{1}{R_1 R_2 C_1 C_2}$	
#3	$$\frac{s^3}{s^3 + Bs^2 + Ds + E} = \frac{s^3}{\left(s + \dfrac{1}{T_a}\right)\left(s + \dfrac{1}{T_b}\right)\left(s + \dfrac{1}{T_c}\right)}$$ where $$B = \frac{\begin{aligned}R_1 R_2 C_1 C_2 + R_1 R_3 C_1 C_3 + R_2 R_3 C_2 C_3 + R_1 R_3 C_2 C_3 \\ + R_1 R_2 C_2 C_3 + R_1 R_2 C_1 C_3\end{aligned}}{R_1 R_2 R_3 C_1 C_2 C_3}$$ $$D = \frac{R_1 C_1 + R_2 C_2 + R_3 C_3 + R_1 C_2 + R_2 C_3}{R_1 R_2 R_3 C_1 C_2 C_3}$$ $$E = \frac{1}{R_1 R_2 R_3 C_1 C_2 C_3}$$	

#4
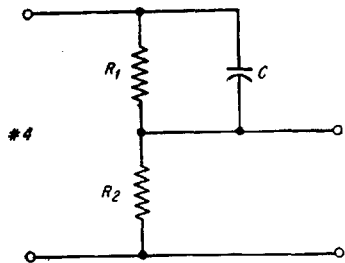
$$\frac{s + \dfrac{1}{T_a}}{s + \dfrac{1}{T_b}}$$

where $T_a = R_1 C$

$$T_b = \frac{R_1 R_2 C}{R_1 + R_2}$$

$$A'(\text{db}) = 20 \log_{10} \frac{R_2}{R_1 + R_2}$$

#5
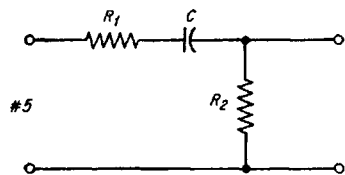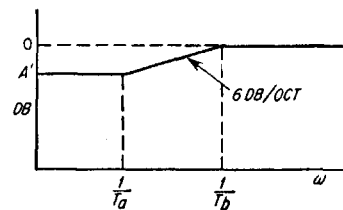
$$\frac{Bs}{s + \dfrac{1}{T}}$$

where $B = \dfrac{R_2}{R_1 + R_2}$

$$T = (R_1 + R_2)C$$

$$A'(\text{db}) = 20 \log_{10} \frac{R_2}{R_1 + R_2}$$

#6
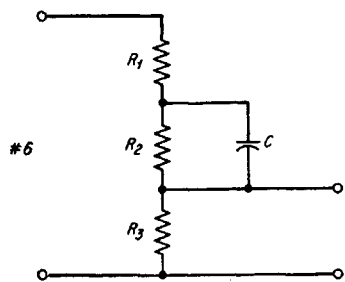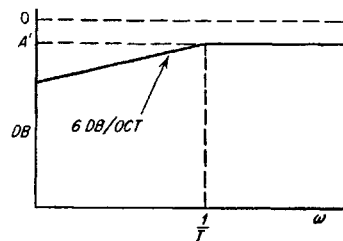
$$\frac{B\left(s + \dfrac{1}{T_a}\right)}{s + \dfrac{1}{T_b}}$$

where $B = \dfrac{R_3}{R_1 + R_3}$

$$T_a = R_2 C$$

$$T_b = \frac{R_2(R_1 + R_3)C}{R_1 + R_2 + R_3}$$

$$A'(\text{db}) = 20 \log_{10} \frac{R_3}{R_1 + R_2 + R_3}$$
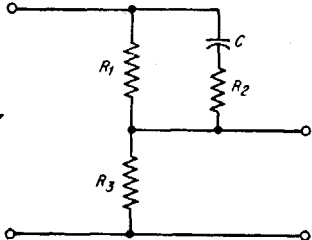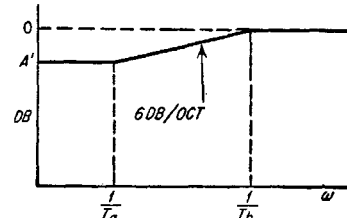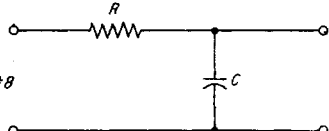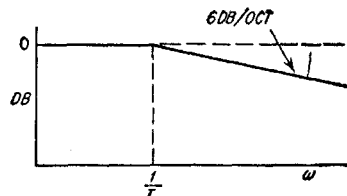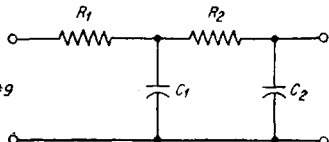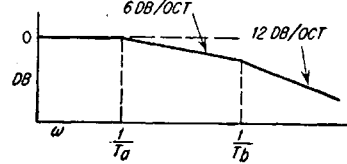
TABLE 1.4. *RC* TRANSFER FUNCTIONS (*Continued*)

1-6

Network	Transfer function	Approx. amp. response (ω on log scale)
#7	$$\dfrac{B\left(s+\dfrac{1}{T_a}\right)}{s+\dfrac{1}{T_b}}$$ where $B = \dfrac{R_1R_3 + R_2R_3}{R_1R_2 + R_1R_3 + R_2R_3}$ $T_a = (R_1 + R_2)C$ $T_b = \dfrac{(R_1R_2 + R_1R_3 + R_2R_3)C}{R_1 + R_3}$ $A'(\text{db}) = 20\log_{10}\dfrac{R_3}{R_1 + R_3}$	
#8	$$\dfrac{1}{s+\dfrac{1}{T}}$$ where $T = RC$	
#9	$$\dfrac{1}{s^2 + Bs + D} = \dfrac{1}{\left(s+\dfrac{1}{T_a}\right)\left(s+\dfrac{1}{T_b}\right)}$$ where $B = \dfrac{R_1C_1 + R_2C_2 + R_1C_2}{R_1R_2C_1C_2}$ $D = \dfrac{1}{R_1R_2C_1C_2}$	

#10
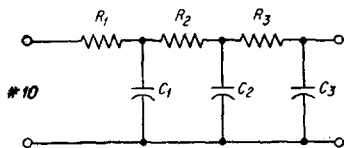
R_1 R_2 R_3

C_1 C_2 C_3

$$\frac{1}{B(s^3 + Ds^2 + Es + F)} = \frac{1}{B\left(s + \dfrac{1}{T_a}\right)\left(s + \dfrac{1}{T_b}\right)\left(s + \dfrac{1}{T_c}\right)}$$

where

$$B = R_1R_2R_3C_1C_2C_3$$

$$D = \frac{R_1R_2C_1C_2 + R_1R_2C_1C_3 + R_1R_3C_1C_3 + R_2R_3C_2C_3 + R_1R_3C_2C_3}{R_1R_2R_3C_1C_2C_3}$$

$$E = \frac{R_1C_1 + R_2C_2 + R_1C_2 + R_1C_3 + R_2C_3 + R_3C_3}{R_1R_2R_3C_1C_2C_3}$$

$$F = \frac{1}{R_1R_2R_3C_1C_2C_3}$$

#11

R_1

C R_2
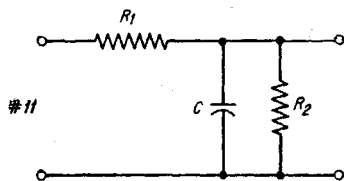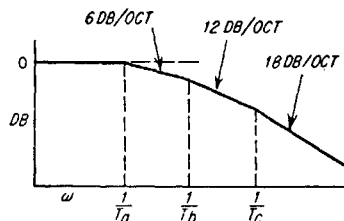
$$\frac{1}{R_1C\left(s + \dfrac{1}{T}\right)}$$

where $T = \dfrac{R_1R_2C}{R_1 + R_2}$

$$A'(\text{db}) = 20\log_{10}\frac{R_2}{R_1 + R_2}$$

#12

R_1

R_2

C
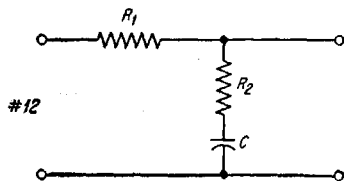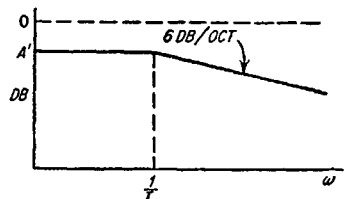
$$\frac{R_2\left(s + \dfrac{1}{T_b}\right)}{(R_1 + R_2)\left(s + \dfrac{1}{T_a}\right)}$$

where $T_a = (R_1 + R_2)C$
$T_b = R_2C$
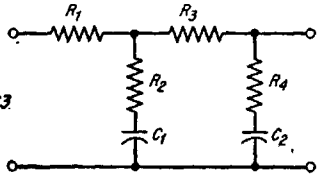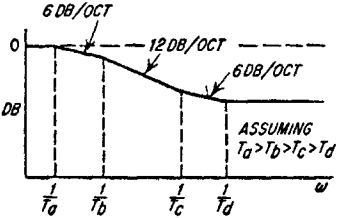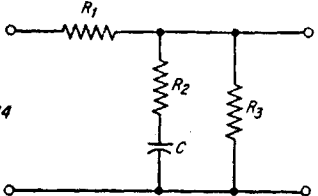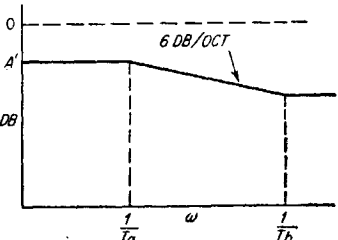
TABLE 1.4. *RC* TRANSFER FUNCTIONS (*Continued*)

Network	Transfer function	Approx. amp. response (ω on log scale)
#13 R_1 R_3 R_2 R_4 C_1 C_2	$\dfrac{B(s^2 + Ds + E)}{(s^2 + Fs + G)} = \dfrac{B\left(s + \dfrac{1}{T_c}\right)\left(s + \dfrac{1}{T_d}\right)}{\left(s + \dfrac{1}{T_a}\right)\left(s + \dfrac{1}{T_b}\right)}$ where $B = \dfrac{R_2 R_4}{R_1 R_2 + R_1 R_3 + R_1 R_4 + R_2 R_3 + R_2 R_4}$ $D = \dfrac{R_2 C_1 + R_4 C_2}{R_2 R_4 C_1 C_2} \qquad E = \dfrac{1}{R_2 R_4 C_1 C_2}$ $F = \dfrac{R_1 C_1 + R_1 C_2 + R_2 C_1 + R_3 C_2 + R_4 C_2}{C_1 C_2 (R_1 R_2 + R_1 R_3 + R_1 R_4 + R_2 R_3 + R_2 R_4)}$ $G = \dfrac{1}{C_1 C_2 (R_1 R_2 + R_1 R_3 + R_1 R_4 + R_2 R_3 + R_2 R_4)}$	 6 DB/OCT, 12 DB/OCT, 6 DB/OCT ASSUMING $T_a > T_b > T_c > T_d$ $\frac{1}{T_a}$ $\frac{1}{T_b}$ $\frac{1}{T_c}$ $\frac{1}{T_d}$
#14 R_1 R_2 R_3 C	$\dfrac{R_2 R_3 \left(s + \dfrac{1}{T_b}\right)}{(R_2 R_3 + R_1 R_2 + R_1 R_3)\left(s + \dfrac{1}{T_a}\right)}$ where $T_a = \dfrac{(R_1 R_2 + R_1 R_3 + R_2 R_3) C}{R_1 + R_3}$ $T_b = R_2 C$ $A'(\text{db}) = 20 \log_{10} \dfrac{R_3}{R_1 + R_3}$	 6 DB/OCT $\frac{1}{T_a}$ ω $\frac{1}{T_b}$

#15
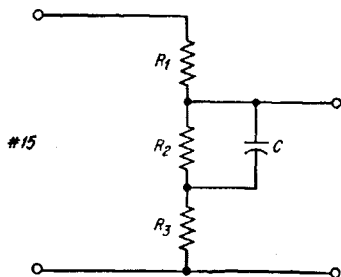
$$\frac{R_3\left(s + \frac{1}{T_b}\right)}{(R_2 + R_3)\left(s + \frac{1}{T_a}\right)}$$

where $T_a = \dfrac{R_2(R_1 + R_3)C}{R_1 + R_2 + R_3}$

$\quad\quad T_b = \dfrac{R_2 R_3 C}{R_2 + R_3}$

$\quad A'(\text{db}) = 20\log_{10}\dfrac{R_2 + R_3}{R_1 + R_2 + R_3}$

#16
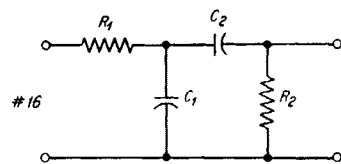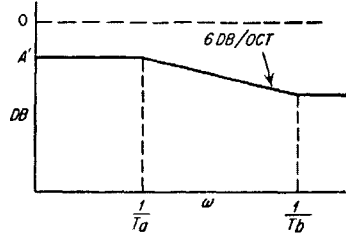
$$\frac{s}{B(s^2 + Ds + E)} = \frac{s}{B\left(s + \frac{1}{T_a}\right)\left(s + \frac{1}{T_b}\right)}$$

where $B = R_1 C_1$

$\quad\quad D = \dfrac{R_1 C_1 + R_2 C_2 + R_1 C_2}{R_1 R_2 C_1 C_2}$

$\quad\quad E = \dfrac{1}{R_1 R_2 C_1 C_2}$

#17
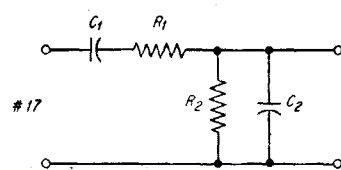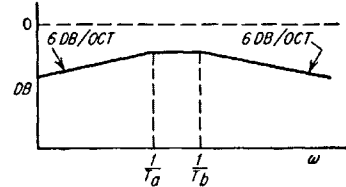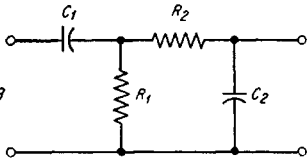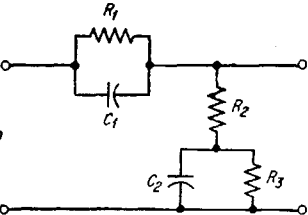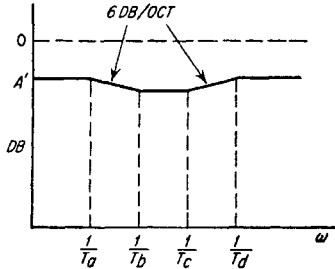
$$\frac{s}{B(s^2 + Ds + E)} = \frac{s}{B\left(s + \frac{1}{T_a}\right)\left(s + \frac{1}{T_b}\right)}$$

where $B = R_1 C_2$

$\quad\quad D = \dfrac{R_1 C_1 + R_2 C_2 + R_2 C_1}{R_1 R_2 C_1 C_2}$

$\quad\quad E = \dfrac{1}{R_1 R_2 C_1 C_2}$

Same as for #16

TABLE 1.4. *RC* TRANSFER FUNCTIONS *(Continued)*

Network	Transfer function	Approx. amp. response (ω on log scale)
#18	$$\frac{s}{B(s^2 + Ds + E)} = \frac{s}{B\left(s + \frac{1}{T_a}\right)\left(s + \frac{1}{T_b}\right)}$$ where $B = R_2 C_2$ $$D = \frac{R_1 C_1 + R_2 C_2 + R_1 C_2}{R_1 R_2 C_1 C_2}$$ $$E = \frac{1}{R_1 R_2 C_1 C_2}$$	Same as for #16
#19	$$\frac{s^2 + Bs + D}{s^2 + Es + F} = \frac{\left(s + \frac{1}{T_b}\right)\left(s + \frac{1}{T_c}\right)}{\left(s + \frac{1}{T_a}\right)\left(s + \frac{1}{T_d}\right)}$$ where $B = \frac{R_1 R_2 C_1 + R_2 R_3 C_2 + R_1 R_3 C_1}{R_1 R_2 R_3 C_1 C_2}$ $$D = \frac{R_2 + R_3}{R_1 R_2 R_3 C_1 C_2}$$ $$E = \frac{R_1 R_2 C_1 + R_1 R_3 C_2 + R_2 R_3 C_2 + R_1 R_3 C_1}{R_1 R_2 R_3 C_1 C_2}$$ $$F = \frac{R_1 + R_2 + R_3}{R_1 R_2 R_3 C_1 C_2}$$ $$A'(\text{db}) = 20 \log_{10} \frac{R_2 + R_3}{R_1 + R_2 + R_3}$$	

#20

$$\frac{s^2 + Bs + D}{s^2 + Es + D} = \frac{\left(s + \dfrac{1}{T_b}\right)\left(s + \dfrac{1}{T_c}\right)}{\left(s + \dfrac{1}{T_a}\right)\left(s + \dfrac{1}{T_d}\right)}$$

where $B = \dfrac{R_1 C_1 + R_2 C_2}{R_1 R_2 C_1 C_2}$

$D = \dfrac{1}{R_1 R_2 C_1 C_2}$

$E = \dfrac{R_1 C_1 + R_2 C_2 + R_1 C_2}{R_1 R_2 C_1 C_2}$

#21
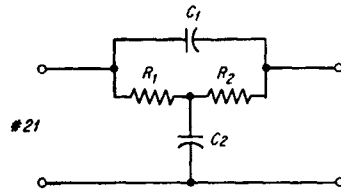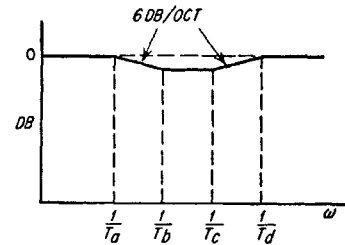
$$\frac{s^2 + Bs + D}{s^2 + Es + D} = \frac{\left(s + \dfrac{1}{T_b}\right)\left(s + \dfrac{1}{T_c}\right)}{\left(s + \dfrac{1}{T_a}\right)\left(s + \dfrac{1}{T_d}\right)}$$

where $B = \dfrac{R_1 C_1 + R_2 C_1}{R_1 R_2 C_1 C_2}$

$D = \dfrac{1}{R_1 R_2 C_1 C_2}$

$E = \dfrac{R_1 C_1 + R_1 C_2 + R_2 C_1}{R_1 R_2 C_1 C_2}$

Same as for #20

#22
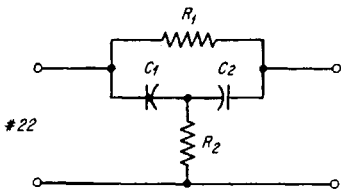
$$\frac{s^2 + Bs + D}{s^2 + Es + D} = \frac{\left(s + \dfrac{1}{T_b}\right)\left(s + \dfrac{1}{T_c}\right)}{\left(s + \dfrac{1}{T_a}\right)\left(s + \dfrac{1}{T_d}\right)}$$

where $B = \dfrac{R_2 C_1 + R_2 C_2}{R_1 R_2 C_1 C_2}$

$D = \dfrac{1}{R_1 R_2 C_1 C_2}$

$E = \dfrac{R_1 C_2 + R_2 C_1 + R_2 C_2}{R_1 R_2 C_1 C_2}$

Same as for #20

TABLE 1.5. RC TRANSFER IMPEDANCES FOR USE AS INPUT AND FEEDBACK IMPEDANCES IN OPERATIONAL AMPLIFIERS*

(See Fig. 3.81)

See explanatory note at end of table on page 1-16.

Transfer impedance function	Network	Relations	Inverse relations
A		$A = R$	$R = A$
$\dfrac{A}{1 + sT}$		$A = R$ $T = RC$	$R = A$ $C = \dfrac{T}{A}$
$A(1 + sT)$		$A = 2R$ $T = \dfrac{RC}{2}$	$R = \dfrac{A}{2}$ $C = \dfrac{4T}{A}$
$A\left(\dfrac{1 + s\theta T}{1 + sT}\right)$ $\theta < 1$		$A = R_1 + R_2$ $T = R_2 C$ $\theta = \dfrac{R_1}{R_1 + R_2}$	$R_1 = A\theta$ $R_2 = A(1 - \theta)$ $C = \dfrac{T}{A(1 - \theta)}$
		$A = R_1$ $T = (R_1 + R_2)C$ $\theta = \dfrac{R_2}{R_1 + R_2}$	$R_1 = A$ $R_2 = \dfrac{A\theta}{1 - \theta}$ $C = \dfrac{T(1 - \theta)}{A}$
$A\left(\dfrac{1 + sT}{1 + s\theta T}\right)$ $\theta < 1$		$A = \dfrac{2R_1 R_2}{2R_1 + R_2}$ $T = \dfrac{R_1 C}{2}$ $\theta = \dfrac{2R_1}{2R_1 + R_2}$	$R_1 = \dfrac{A}{2(1 - \theta)}$ $R_2 = \dfrac{A}{\theta}$ $C = \dfrac{4T(1 - \theta)}{A}$
		$A = 2R_1$ $T = \left(R_2 + \dfrac{R_1}{2}\right)C$ $\theta = \dfrac{2R_2}{2R_2 + R_1}$	$R_1 = \dfrac{A}{2}$ $R_2 = \dfrac{A\theta}{4(1 - \theta)}$ $C = \dfrac{4T(1 - \theta)}{A}$
		$A = 2R$ $T = \dfrac{R}{2}(C_1 + C_2)$ $\theta = \dfrac{2C_2}{C_1 + C_2}$	$R = \dfrac{A}{2}$ $C_1 = \dfrac{2T(2 - \theta)}{A}$ $C_2 = \dfrac{2T\theta}{A}$
$\dfrac{1}{sB}\left[\dfrac{(1 + sT_1)(1 + sT_3)}{1 + sT_2}\right]$ $T_1 < T_2 < T_3$		$B = C_1$ $T_2 = (R_1 + R_2)C_2$ $T_1 T_3 = R_1 R_2 C_1 C_2$ $T_1 + T_3 = R_1 C_1 + R_2 C_2 + R_1 C_2$	$R_1 = \dfrac{T_1 + T_3 - T_2}{B}$ $R_2 = \dfrac{T_1 T_3 (T_1 + T_3 - T_2)}{B(T_3 - T_2)(T_2 - T_1)}$ $C_1 = B$ $C_2 = \dfrac{B(T_3 - T_2)(T_2 - T_1)}{(T_1 + T_3 - T_2)^2}$
		$B = C_1 + C_2$ $T_2 = R_2\left(\dfrac{C_1 C_2}{C_1 + C_2}\right)$ $T_1 T_3 = R_1 R_2 C_1 C_2$ $T_1 + T_3 = R_1 C_1 + R_2 C_2 + R_1 C_2$	$R_1 = \dfrac{T_1 T_3}{BT_2}$ $R_2 = \dfrac{(T_1 T_2 + T_2 T_3 - T_1 T_3)^2}{BT_2(T_3 - T_2)(T_2 - T_1)}$ $C_1 = \dfrac{BT_2^2}{T_1 T_2 + T_2 T_3 - T_1 T_3}$ $C_2 = \dfrac{B(T_3 - T_2)(T_2 - T_1)}{T_1 T_2 + T_2 T_3 - T_1 T_3}$

* Reprinted from F. R. Bradley and R. McCoy, Driftless D-C Amplifier, *Electronics*, pp. 144–148, April, 1952.

TABLE 1.5. *RC* TRANSFER IMPEDANCES FOR USE AS INPUT AND FEEDBACK IMPEDANCES IN OPERATIONAL AMPLIFIERS (*Continued*)

Transfer impedance function	Network	Relations	Inverse relations
$\dfrac{1}{sB}\left[\dfrac{(1+sT_1)(1+sT_3)}{1+sT_2}\right]$ $T_1 < T_2 < T_3$		$B = C_1$ $T_2 = R_2C_2$ $T_1T_3 = R_1R_2C_1C_2$ $T_1 + T_3 = R_1C_1 + R_2C_2 + R_2C_1$	$R_1 = \dfrac{T_1T_3}{BT_2}$ $R_2 = \dfrac{(T_3 - T_2)(T_2 - T_1)}{BT_2}$ $C_1 = B$ $C_2 = \dfrac{BT_2^2}{(T_3 - T_2)(T_2 - T_1)}$
		$B = C_1$ $T_2 = R_2C_2$ $T_1T_3 = R_1R_2C_1C_2$ $T_1 + T_3 = R_1C_1 + R_2C_2 + R_1C_2$	$R_1 = \dfrac{T_1T_3}{BT_2}$ $R_2 = \dfrac{T_1T_2T_3}{B(T_3 - T_2)(T_2 - T_1)}$ $C_1 = B$ $C_2 = \dfrac{B(T_3 - T_2)(T_2 - T_1)}{T_1T_3}$
$\dfrac{1}{sB}(1 + sT_1)(1 + sT_2)$ $T_1 \neq T_2$		$B = C_2$ $T_1T_2 = R_1R_2C_1C_2$ $T_1 + T_2 = R_1C_1 + R_2C_2 + R_1C_2$	$R_1 = \dfrac{(\sqrt{T_1} - \sqrt{T_2})^2}{B}$ $R_2 = \dfrac{\sqrt{T_1T_2}}{B}$ $C_1 = \dfrac{B\sqrt{T_1T_2}}{(\sqrt{T_1} - \sqrt{T_2})^2}$ $C_2 = B$
$\dfrac{1}{sB}\left[\dfrac{(1+sT_1)(1+sT_2)}{s\sqrt{T_1T_2}}\right]$ $T_1 \neq T_2$		$B = C_2$ $T_1T_2 = R_1R_2C_1C_2$ $T_1 + T_2 = R_1C_1 + R_2C_2 + R_1C_2$	$R = \dfrac{(\sqrt{T_1} - \sqrt{T_2})^2}{B}$ $R_2 = \dfrac{\sqrt{T_1T_2}}{B}$ $C_1 = \dfrac{B\sqrt{T_1T_2}}{(\sqrt{T_1} - \sqrt{T_2})^2}$ $C_2 = B$
$\dfrac{1}{sB}\left[\dfrac{(1+sT_1)(1+sT_2)}{s^2T_1T_2}\right]$ $T_1 < T_2$		$B = \dfrac{C_1C_2}{C_1 + 2C_2}$ $T_1 = RC_1$ $T_2 = R(C_1 + 2C_2)$	$R = \dfrac{T_1(T_2 - T_1)}{2BT_2}$ $C_1 = \dfrac{2BT_2}{T_2 - T_1}$ $C_2 = \dfrac{BT_2}{T_1}$
$A\left(\dfrac{1 + sT_1}{1 + s^2T_1T_2}\right)$	$R_1C_1 = 4R_2C_2$	$A = 2R_1$ $T_1 = \dfrac{R_1C_1}{2} = 2R_2C_2$ $T_2 = R_1C_2$	$R_1 = \dfrac{A}{2}$ $R_2 = \dfrac{AT_1}{4T_2}$ $C_1 = \dfrac{4T_1}{A}$ $C_2 = \dfrac{2T_2}{A}$
$\dfrac{1}{sB}$		$B = C$	$C = B$
$\dfrac{1}{sB}(1 + sT)$		$B = C$ $T = RC$	$R = \dfrac{T}{B}$ $C = B$
$\dfrac{1}{sB}\left(\dfrac{1 + sT}{sT}\right)$		$B = \dfrac{C}{2}$ $T = 2RC$	$R = \dfrac{T}{4B}$ $C = 2B$

TABLE 1.5. RC TRANSFER IMPEDANCES FOR USE AS INPUT AND FEEDBACK IMPEDANCES IN OPERATIONAL AMPLIFIERS (*Continued*)

Transfer impedance function	Network	Relations	Inverse relations
$\dfrac{1}{sB}\left(\dfrac{1+sT}{1+s\theta T}\right)$ $\theta < 1$	*(network diagram)*	$B = C_1$ $T = R(C_1 + C_2)$ $\theta = \dfrac{C_2}{C_1 + C_2}$	$R = \dfrac{T(1-\theta)}{B}$ $C_1 = B$ $C_2 = \dfrac{B\theta}{1-\theta}$
	(network diagram)	$B = C_1 + C_2$ $T = RC_2$ $\theta = \dfrac{C_1}{C_1 + C_2}$	$R = \dfrac{T}{B(1-\theta)}$ $C_1 = B\theta$ $C_2 = B(1-\theta)$
$\dfrac{1}{sB}\left(\dfrac{(1+sT_1)(1+sT_3)}{1+sT_2}\right)$ $T_1 < T_2 < T_3$	*(network diagram)*	$B = C_1 + C_2$ $T_1 = R_1 C_1$ $T_2 = (R_1 + R_2)\left(\dfrac{C_1 C_2}{C_1 + C_2}\right)$ $T_3 = R_2 C_2$	$R_1 = \dfrac{T_1(T_3 - T_1)}{B(T_2 - T_1)}$ $R_2 = \dfrac{T_3(T_3 - T_1)}{B(T_3 - T_2)}$ $C_1 = \dfrac{B(T_2 - T_1)}{T_3 - T_1}$ $C_2 = \dfrac{B(T_3 - T_2)}{T_3 - T_1}$
$A\left(\dfrac{1+sT_1}{1+sT_1+s^2 T_1 T_2}\right)$	*(network diagram)*	$A = R_2$ $T_1 = 2R_1 C$ $T_2 = \dfrac{R_2 C}{2}$	$R_1 = \dfrac{A T_1}{4 T_2}$ $R_2 = A$ $C = \dfrac{2T_2}{A}$
$A\left(\dfrac{1+sT_2}{1+sT_1+s^2 T_1 T_2}\right)$	*(network diagram)*	$A = 2R$ $T_1 = 2RC_2$ $T_2 = \dfrac{RC_1}{2}$	$R = \dfrac{A}{2}$ $C_1 = \dfrac{4T_2}{A}$ $C_2 = \dfrac{T_1}{A}$
	(network diagram)	$A = \dfrac{2R_1 R_2}{2R_1 + R_2}$ $T_1 = \dfrac{R_1(R_1 C_1 + 2R_2 C_2)}{2R_1 + R_2}$ $T_2 = \dfrac{R_1 R_2 C_1 C_2}{R_1 C_1 + 2R_2 C_2}$ $T_3 = \dfrac{R_1 C_1}{2}$	$R_1 = \dfrac{A T_3^2}{2[T_3^2 - \frac{1}{4}T_1(T_3 - T_2)]}$ $R_2 = \dfrac{A T_3^2}{T_1(T_3 - T_2)}$ $C_1 = \dfrac{4[T_3^2 - T_1(T_3 - T_2)]}{A T_3}$ $C_2 = \dfrac{T_1 T_2}{A T_3}$
$A\left(\dfrac{1+sT_3}{1+sT_1+s^2 T_1 T_2}\right)$ $T_2 > \dfrac{T_1}{4}\binom{\text{Complex}}{\text{roots}}$ $T_3 > T_2$	*(network diagram)*	$A = 2R_1$ $T_1 = R_2 C_1 + 2R_1 C_2$ $T_2 = \dfrac{R_1(R_1 + 2R_2)C_1 C_2}{R_2 C_1 + 2R_1 C_2}$ $T_3 = \left(R_2 + \dfrac{R_1}{2}\right)C_1$	$R_1 = \dfrac{A}{2}$ $R_2 = \dfrac{A T_1(T_3 - T_2)}{4[T_3^2 - T_1(T_3 - T_2)]}$ $C_1 = \dfrac{4[T_3^2 - T_1(T_3 - T_2)]}{A T_3}$ $C_2 = \dfrac{T_1 T_2}{A T_3}$
	(network diagram)	$A = 2R$ $T_1 = R(C_2 + 2C_3)$ $T_2 = \dfrac{RC_3(C_1 + C_2)}{C_2 + 2C_3}$ $T_3 = \dfrac{R}{2}(C_1 + C_2)$	$R = \dfrac{A}{2}$ $C_1 = \dfrac{2[2T_3^2 - T_1(T_3 - T_2)]}{A T_3}$ $C_2 = \dfrac{2T_1(T_3 - T_2)}{A T_3}$ $C_3 = \dfrac{T_1 T_2}{A T_3}$

TABLE 1.5. *RC* TRANSFER IMPEDANCES FOR USE AS INPUT AND FEEDBACK
IMPEDANCES IN OPERATIONAL AMPLIFIERS (*Continued*)

Transfer impedance function	Network	Relations	Inverse relations
$A\left(\dfrac{1 + sT_3}{1 + sT_1 + s^2 T_1 T_2}\right)$ $T_2 > \dfrac{T_1}{4}\left(\begin{array}{c}\text{Complex}\\\text{roots}\end{array}\right)$ $T_3 < T_1$		$A = R_2$ $T_1 = 2R_1C_1 + R_2C_2$ $T_2 = \dfrac{R_1R_2C_1(C_1 + 2C_2)}{2R_1C_1 + R_2C_2}$ $T_3 = 2R_1C_1$	$R_1 = \dfrac{AT_3^2}{4[T_1T_2 - T_3(T_1 - T_3)]}$ $R_2 = A$ $C_1 = \dfrac{2[T_1T_2 - T_3(T_1 - T_3)]}{AT_3}$ $C_2 = \dfrac{(T_1 - T_3)}{A}$
		$A = R_2$ $T_1 = \dfrac{C_1(2R_1C_2 + R_2C_1)}{2C_1 + C_2}$ $T_2 = \dfrac{R_1R_2C_1C_2}{2R_1C_2 + R_2C_1}$ $T_3 = \dfrac{2R_1C_1C_2}{2C_1 + C_2}$	$R_1 = \dfrac{AT_3^2}{4[T_1T_2 - T_3(T_1 - T_3)]}$ $R_2 = A$ $C_1 = \dfrac{2T_1T_2}{AT_3}$ $C_2 = \dfrac{4T_1T_2[T_1T_2 - T_3(T_1 - T_3)]}{AT_3^2(T_1 - T_3)}$
		$A = R_3$ $T_1 = \dfrac{R_1(2R_2 + R_3)C}{R_1 + R_2}$ $T_2 = \dfrac{R_2R_3C}{2R_2 + R_3}$ $T_3 = \dfrac{2R_1R_2C}{R_1 + R_2}$	$R_1 = \dfrac{AT_3^2}{2[2T_1T_2 - T_3(T_1 - T_3)]}$ $R_2 = \dfrac{AT_3}{2(T_1 - T_3)}$ $R_3 = A$ $C = \dfrac{2T_1T_2}{AT_3}$
$A(1 + sT_1)(1 + sT_2)$ $T_1 < T_2$		$A = 2R_1 + R_2$ $T_1 = \left(\dfrac{R_1R_2}{2R_1 + R_2}\right)C$ $T_2 = R_1C$	$R_1 = A\left(\dfrac{T_2 - T_1}{2T_2}\right)$ $R_2 = A\dfrac{T_1}{T_2}$ $C = \dfrac{2T_2^2}{A(T_2 - T_1)}$
$\dfrac{1}{sB}\left(\dfrac{1 + s\theta T}{1 + sT}\right)$ $\theta < 1$		$B = C_2$ $T = RC_1\left(\dfrac{2C_2 + C_1}{C_2}\right)$ $\theta = \dfrac{2C_2}{2C_2 + C_1}$	$R = \dfrac{T\theta^2}{4B(1 - \theta)}$ $C_1 = \dfrac{2B(1 - \theta)}{\theta}$ $C_2 = B$
		$B = \dfrac{C_1^2}{2C_1 + C_2}$ $T = RC_2$ $\theta = \dfrac{2C_1}{2C_1 + C_2}$	$R = \dfrac{T\theta^2}{4B(1 - \theta)}$ $C_1 = \dfrac{2B}{\theta}$ $C_2 = \dfrac{4B(1 - \theta)}{\theta^2}$
		$B = \left(\dfrac{R_1}{R_1 + R_2}\right)C$ $T = R_2C$ $\theta = \dfrac{2R_1}{R_1 + R_2}$	$R_1 = \dfrac{T\theta^2}{2B(2 - \theta)}$ $R_2 = \dfrac{T\theta}{2B}$ $C = \dfrac{2B}{\theta}$
$A\left[\dfrac{1 + sT_2}{(1 + sT_1)(1 + sT_3)}\right]$ $T_1 < T_2 < T_3$		$A = R_1 + R_2$ $T_1 = R_1C_1$ $T_2 = \left(\dfrac{R_1R_2}{R_1 + R_2}\right)(C_1 + C_2)$ $T_3 = R_2C_2$	$R_1 = \dfrac{A(T_2 - T_1)}{T_3 - T_1}$ $R_2 = \dfrac{A(T_3 - T_2)}{T_3 - T_1}$ $C_1 = \dfrac{T_1(T_3 - T_1)}{A(T_2 - T_1)}$ $C_2 = \dfrac{T_3(T_3 - T_1)}{A(T_3 - T_2)}$

TABLE 1.5. RC TRANSFER IMPEDANCES FOR USE AS INPUT AND FEEDBACK IMPEDANCES IN OPERATIONAL AMPLIFIERS (*Continued*)

Transfer impedance function	Network	Relations	Inverse relations
$A\left[\dfrac{1+sT_2}{(1+sT_1)(1+sT_3)}\right]$ $T_1 < T_2 < T_3$		$A = R_2$ $T_2 = R_1C_1$ $T_1T_3 = R_1R_2C_1C_2$ $T_1 + T_3 = R_1C_1 + R_2C_2 + R_2C_1$	$R_1 = \dfrac{AT_2{}^2}{(T_3 - T_2)(T_2 - T_1)}$ $R_2 = A$ $C_1 = \dfrac{(T_3 - T_2)(T_2 - T_1)}{AT_2}$ $C_2 = \dfrac{T_1T_3}{AT_2}$
		$A = R_1 + R_2$ $T_2 = \left(\dfrac{R_1R_2}{R_1 + R_2}\right)C_2$ $T_1T_3 = R_1R_2C_1C_2$ $T_1 + T_3 = R_1C_1 + R_2C_2 + R_2C_1$	$R_1 = \dfrac{AT_2{}^2}{T_1T_2 + T_2T_3 - T_1T_3}$ $R_2 = \dfrac{A(T_3 - T_2)(T_2 - T_1)}{T_1T_2 + T_2T_3 - T_1T_3}$ $C_1 = \dfrac{T_1T_3}{AT_2}$ $C_2 = \dfrac{(T_1T_2 + T_2T_3 - T_1T_3)^2}{AT_2(T_3 - T_2)(T_2 - T_1)}$
		$A = R_1$ $T_2 = R_2(C_1 + C_2)$ $T_1T_3 = R_1R_2C_1C_2$ $T_1 + T_3 = R_1C_1 + R_2C_2 + R_2C_1$	$R_1 = A$ $R_2 = \dfrac{A(T_3 - T_2)(T_2 - T_1)}{(T_1 + T_3 - T_2)^2}$ $C_1 = \dfrac{T_1 + T_3 - T_2}{A}$ $C_2 = \dfrac{T_1T_3(T_1 + T_3 - T_2)}{A(T_3 - T_2)(T_2 - T_1)}$
$A\left[\dfrac{1+sT_2}{(1+sT_1)(1+sT_3)}\right]$ $T_2 \gtrless T_1 \gtrless T_3$		$A = 2R_1 + \dfrac{R_1{}^2}{R_2}$ $T_1 = R_1C_1$ $T_2 = \left(\dfrac{R_1R_2}{R_1 + 2R_2}\right)(C_1 + C_2)$ $T_3 = R_1C_2$	$R_1 = \dfrac{AT_2}{T_1 + T_3}$ $R_2 = \dfrac{AT_2{}^2}{(T_1 + T_3)(T_1 + T_3 - 2T_2)}$ $C_1 = \dfrac{T_1(T_1 + T_3)}{AT_2}$ $C_2 = \dfrac{T_3(T_1 + T_3)}{AT_2}$
$A\left[\dfrac{1+sT_2}{(1+sT_1)(1+sT_3)}\right]$ $T_1 \gtrless T_3 \gtrless T_2$		$A = R_1 + R_2$ $T_1 = R_1C_1$ $T_2 = \dfrac{R_1R_2}{R_1 + R_2}(2C_1 + C_2)$ $T_3 = R_2C_1$	$R_1 = \dfrac{AT_1}{T_1 + T_3}$ $R_2 = \dfrac{AT_3}{T_1 + T_3}$ $C_1 = \dfrac{T_1 + T_3}{A}$ $C_2 = \dfrac{(T_1 + T_3)}{A}\left(\dfrac{T_2}{T_1} + \dfrac{T_2}{T_3} - 2\right)$

NOTE: In order that the operational amplifier generate a specific $f(s)$, make Z_f/Z_i equal to $f(s)$ where Z_f and Z_i are each in a form corresponding to a function in the left-hand column. Relationships between the network values, gain and time constants can be established from the adjacent columns. In the case of either an input or a feedback impedance, the right-hand terminal of the network must be connected to the junction between Z_i and Z_f.
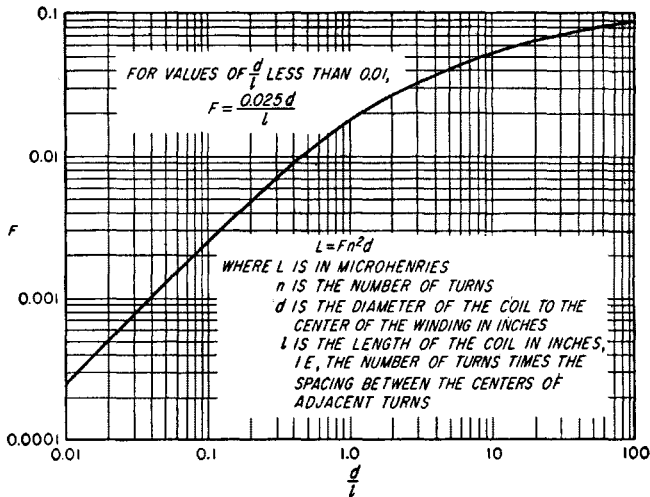
FIG. 1.1. Inductance of a single-layer air-core coil.

$$f_0 \approx \frac{0.45}{\tau} \qquad C = \frac{\tau}{Z_0} = \frac{m}{\pi f_0 Z_0}$$

$$T = \frac{m}{\pi f_0} \qquad L = \frac{m Z_0}{\pi f_0}$$
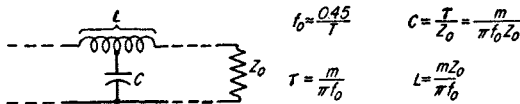
WHERE τ = OUTPUT RISE TIME IN SECONDS BETWEEN 10 AND 90 PERCENT
  POINTS FOR A STEP FUNCTION INPUT
  f_0 = CUTOFF FREQUENCY IN CYCLES PER SECOND
  T = TIME DELAY PER SECTION IN SECONDS
  Z_0 = CHARACTERISTIC IMPEDANCE OF DELAY LINE IN OHMS
PROCEDURE:
  1. CHOOSE THE VALUE OF m FOR THE DESIRED DELAY VS. FREQUENCY
    CHARACTERISTIC
  2. CALCULATE THE REQUIRED VALUES OF L AND C FOR THE DESIRED
    IMPEDANCE, DELAY AND RISE TIME
  3. DETERMINE THE COIL FORM FACTOR $\frac{d}{l}$ FROM FIG. 1.3 FOR THE
    REQUIRED VALUE OF m
  4. OBTAIN THE COIL WINDING DATA FROM FIG. 1.1 FOR THE DESIRED
    TOTAL INDUCTANCE L AND FORM FACTOR $\frac{d}{l}$
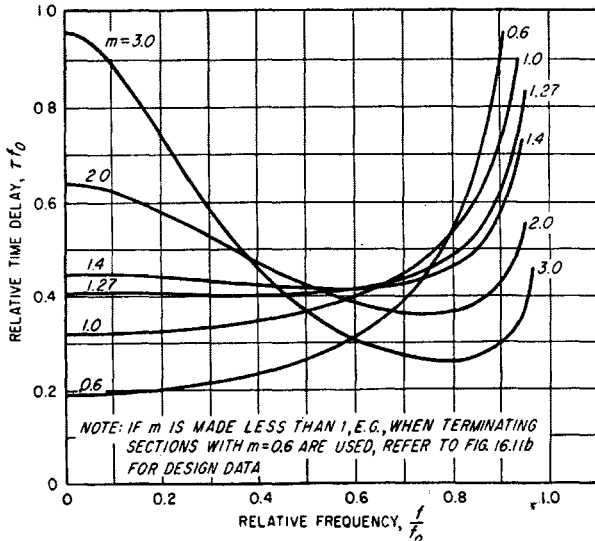
FIG. 1.2. Design data for delay-line sections.
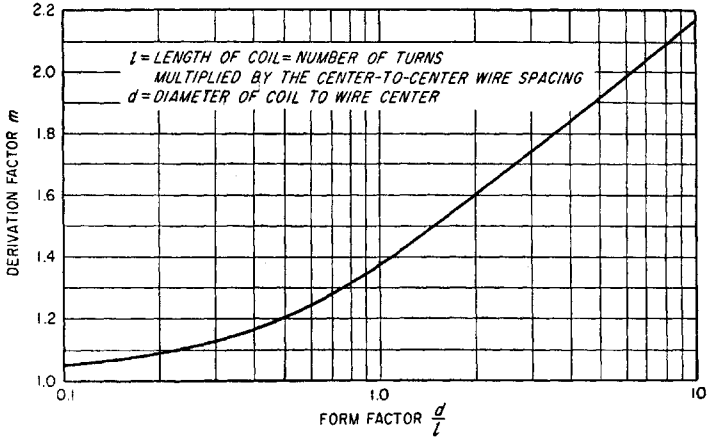
FIG. 1.3. Form factor d/l for inductors to be used in delay-line sections having a derivation factor m greater than unity.
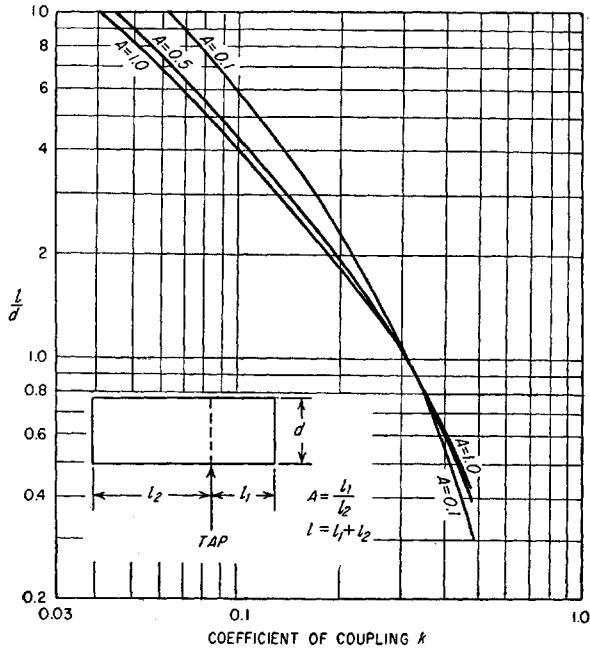
FIG. 1.4. Coefficient of coupling between sections of a tapped single-layer coil.
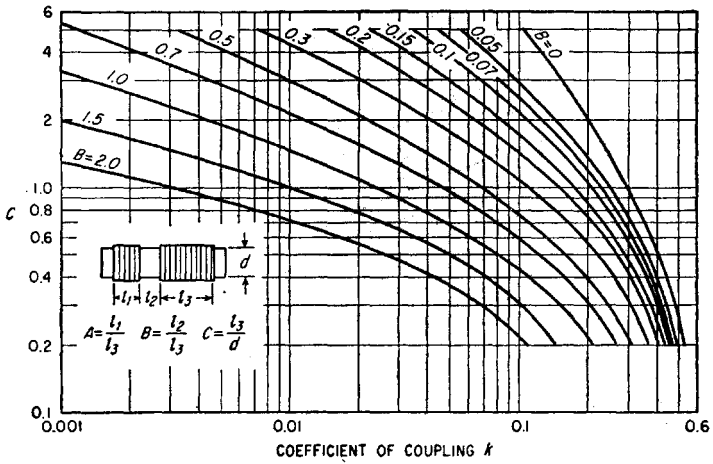
FIG. 1.5. Coefficient of coupling between two-layer wound coaxial coils having a ratio A of winding lengths equal to 0.1.
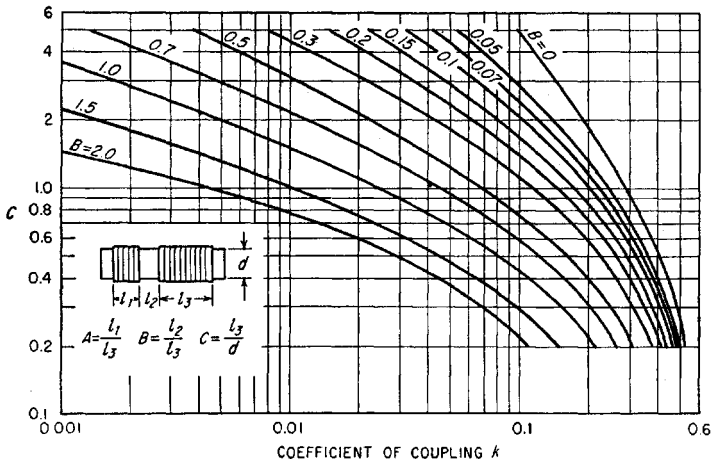
FIG. 1.6. Coefficient of coupling between two-layer wound coaxial coils having a ratio A of winding lengths equal to 0.15.
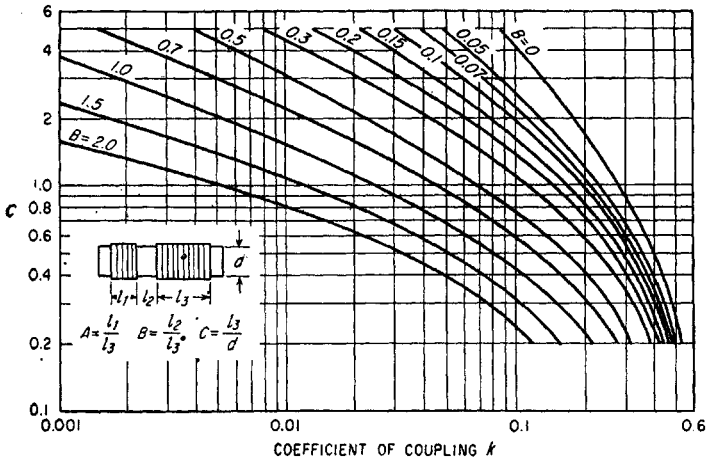
FIG. 1.7. Coefficient of coupling between two-layer wound coaxial coils having a ratio A of winding lengths equal to 0.2.
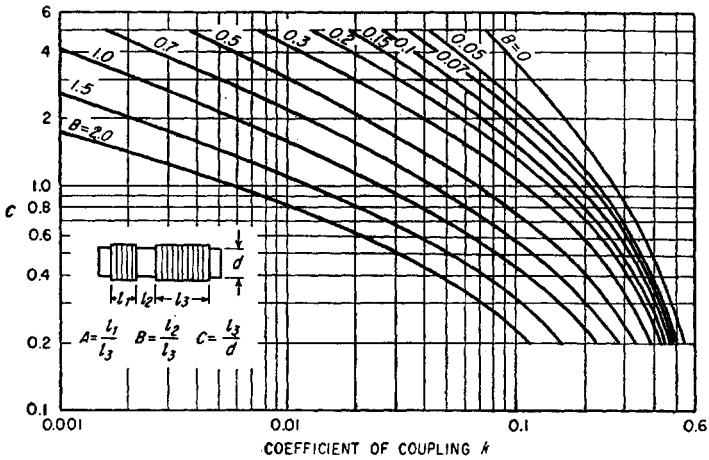
FIG. 1.8. Coefficient of coupling between two-layer wound coaxial coils having a ratio A of winding lengths equal to 0.3.
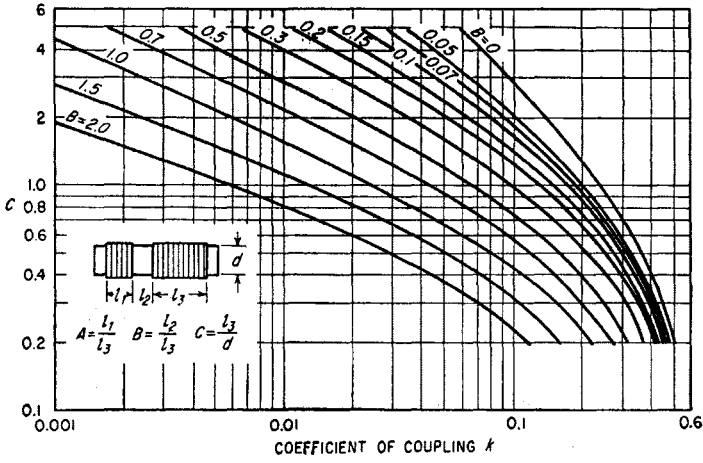
Fig. 1.9. Coefficient of coupling between two-layer wound coaxial coils having a ratio A of winding lengths equal to 0.5.
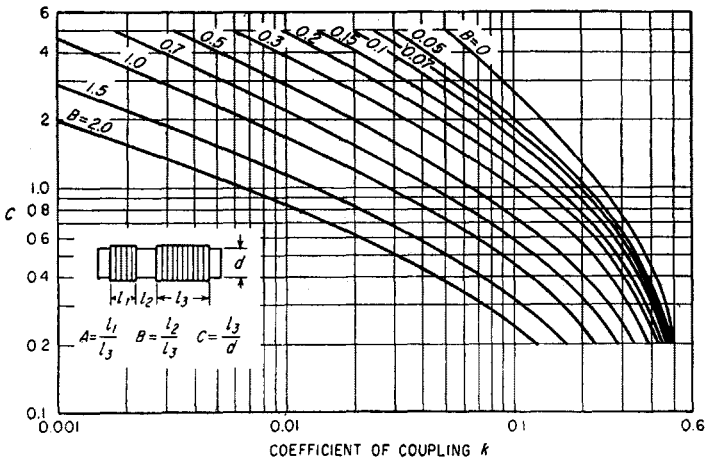
Fig. 1.10. Coefficient of coupling between two-layer wound coaxial coils having a ratio A of winding lengths equal to 0.7.
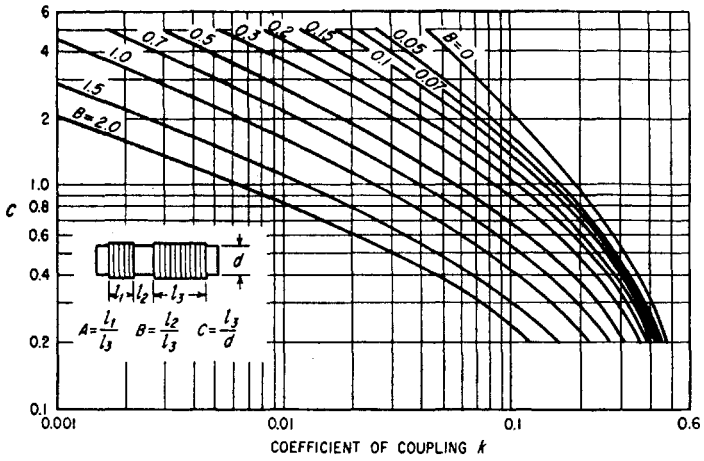
FIG. 1.11. Coefficient of coupling between two-layer wound coaxial coils having a ratio A of winding lengths equal to 1.0.
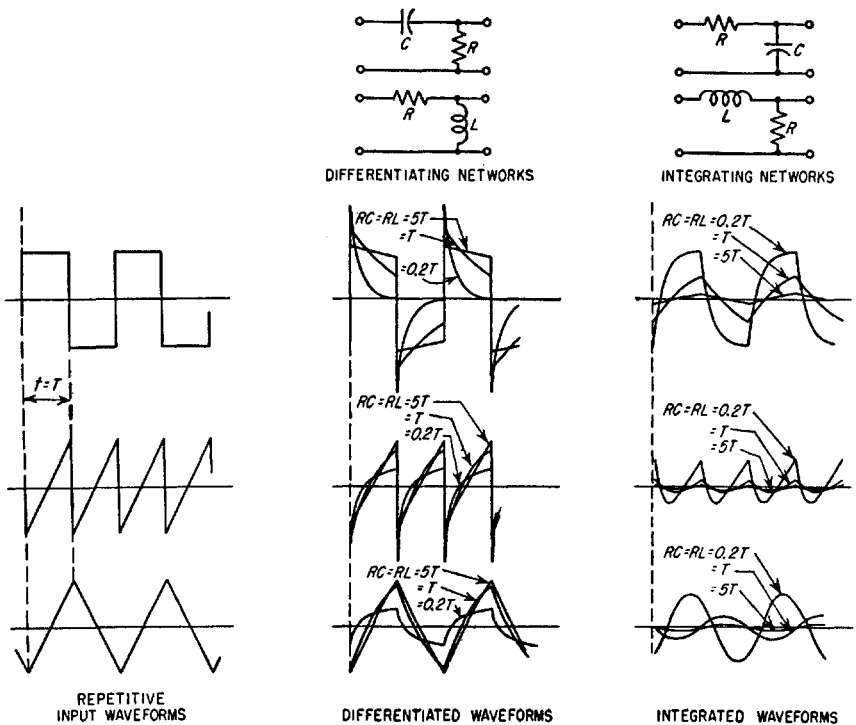
FIG. 1.12. Differentiated and integrated waveforms.

Vacuum Tubes and Transistors

2.1. Electron Motion. The motion of an electron in a vacuum is determined by the forces acting upon the electron and its mass. These forces are the resultant of electric and/or magnetic fields surrounding the electron and the electrostatic charge of the electron. The physical characteristics of the electron are:

$$\text{Mass of electron} = M_e = 9.11 \times 10^{-31} \text{ kg}$$
$$\text{Charge of electron} = e = -1.60 \times 10^{-19} \text{ coulomb}$$
$$\text{Apparent diameter} = D = 3.8 \times 10^{-15} \text{ m}$$

2.1a. Motion of Electrons in a Uniform Electric Field. An electron is accelerated in accordance with the relation

$$F = M_e a \tag{2.1}$$

where F = force acting upon electron, newtons (1 newton = 10^5 dynes)
  M_e = mass of electron, kg
  a = acceleration of electron, m/sec²
The force acting upon the electron in an electric field is

$$F = \mathcal{E}e \tag{2.2}$$

where F = force, newtons

  $\mathcal{E}$ = electric field = negative electric gradient, volts/m = $-\dfrac{\Delta V}{\Delta s}$

  e = charge of electron, coulombs
  ΔV = change in potential in distance Δs
The electric field $\mathcal{E}$ is positive in the direction from a positive potential to a negative potential. The acceleration of the electron in an electric field is given by

$$a = \frac{\mathcal{E}e}{M_e} \qquad \text{m/sec}^2 \tag{2.3}$$

In a uniform field, i.e., a region where $\mathcal{E}$ is constant, the velocity v of the electron at any time t after starting from rest at $t = 0$ is

$$v = \int_0^t a \, dt = \frac{\mathcal{E}et}{M_e} \qquad \text{m/sec}$$
$$= -1.76 \times 10^{11} \mathcal{E}t \qquad \text{m/sec} \tag{2.4}$$

The distance traveled by the electron at time t starting from rest at $t = 0$ is

$$s = \int_0^t v \, dt = \frac{\mathcal{E}e}{2M_e} t^2 \qquad \text{meters}$$
$$= -0.88 \times 10^{11} \mathcal{E}t^2 \qquad \text{meters} \tag{2.5}$$

When the electron has an initial velocity v_0 parallel to the electric field, Eqs. (2.4)

and (2.5) are modified to include the effect of initial velocity as follows:

$$v = v_0 + \frac{\mathcal{E}et}{M_e} \quad \text{m/sec}$$

$$= v_0 - 1.76 \times 10^{11}\mathcal{E}t \quad \text{m/sec} \tag{2.6}$$

or

$$v = \sqrt{\frac{-2Ve}{M_e} + v_0^2} \quad \text{m/sec} \tag{2.7}$$

$$s = v_0 t + \frac{\mathcal{E}et^2}{2M} = v_0 t - 0.88 \times 10^{11}\mathcal{E}t^2 \quad \text{meters} \tag{2.8}$$

where $V = -\mathcal{E}s$ = total potential through which electron is accelerated

The total kinetic energy KE of an electron accelerated through a change in potential of V volts regardless of the uniformity of the field is

$$KE = \tfrac{1}{2}M_e v^2 = -Ve + \tfrac{1}{2}M_e v_0^2 \quad \text{joules} \tag{2.9}$$

$$= -Ve + (KE)_0 \quad \text{joules} \tag{2.9a}$$

where v_0 = initial velocity of electron

$(KE)_0$ = initial kinetic energy

The energy of an electron is commonly expressed in electron-volts. An electron which has accelerated through a potential of one volt has increased its kinetic energy by one electron-volt. One electron-volt is equal to 1.60×10^{-19} joule. Thus, for an electron Eq. (2.9) can be written as

$$KE = V + (KE)_0 \quad \text{electron-volts} \tag{2.9b}$$

When the electron has an initial velocity v_0 which is not parallel to the electric field $\mathcal{E}$, the velocity can be resolved into a component of velocity v_p parallel to the field and a component of velocity v_n normal to the field. The parallel component of velocity v_p is changed by the field $\mathcal{E}$ and the initial value is substituted for v_0 in Eqs. (2.6), (2.7), and (2.8). The normal component of velocity v_n is unaffected by the field $\mathcal{E}$. Electron motion in this type of field is further discussed in Sec. 2.7d.

The above equations for electron motion are accurate for electron velocities sufficiently less than the velocity of light so that relativity effects can be neglected. These effects must be considered for electrons accelerated through more than about 30,000 volts.[1]

Application of the fundamental laws of electron motion is illustrated by the following example.

Example 2.1

An electron is emitted from a cathode with 10 electron-volts of energy and a velocity normal to the cathode surface. Another electrode parallel to the cathode and spaced 10 cm away has a potential of +1,000 volts with respect to the cathode. Find the kinetic energy of the electron as it strikes the anode. Determine the time required for the electron to reach the anode, assuming the field between the cathode and anode is uniform.

Solution

1. Determine the kinetic energy of the electron as it strikes the anode.
From Eq. (2.9b)

$$KE = V + (KE)_0$$
$$= 1,000 + 10 = 1,010 \text{ electron-volts}$$
$$= 1.616 \times 10^{-16} \text{ joule}$$

---

[1] For a discussion of electron velocity and relativity effects, see Karl A. Spangenberg, "Vacuum Tubes," pp. 103–107, McGraw-Hill Book Company, Inc., New York, 1948.

2. Determine the initial velocity v_0.
From Eq. (2.9)

$$v_0 = \sqrt{\frac{2(\text{KE})_0}{M_e}}$$

$$= \sqrt{\frac{2 \times 10 \times 1.6 \times 10^{-19}}{9.11 \times 10^{-31}}}$$

$$= 1.87 \times 10^6 \text{ m/sec}$$

3. Determine the final velocity v.
From Eq. (2.9)

$$v = \sqrt{\frac{2 \times 1,010 \times 1.6 \times 10^{-19}}{9.11 \times 10^{-31}}}$$

$$= 1.88 \times 10^7 \text{ m/sec}$$

4. Determine the time required for the electron to reach the anode.
From Eq. (2.6)

$$t = \frac{M_e(v - v_0)}{\mathcal{E}e}$$

$$= \frac{9.11 \times 10^{-31} (1.88 \times 10^7 - 1.87 \times 10^6)}{(-10^4)(-1.6 \times 10^{-19})}$$

$$= 9.64 \times 10^{-9} \text{ sec}$$

2.1b. Motion of Electrons in a Uniform Magnetic Field.   The force exerted upon an electron in motion in a uniform magnetic field is given by

$$F = |Bev \sin \theta| \qquad (2.10)$$

where F = force, newtons

B = magnetic flux density, webers/m² (1 weber/m² = 10^4 gauss)

e = charge of electron, coulombs

v = velocity of electron, m/sec

θ = angle between direction of magnetic flux and direction of electron velocity

The magnetic field exerts a force upon the electron only if the electron has a component of velocity normal to the direction of the magnetic field.   The force F upon the electron has a direction which is normal to the directions of both the magnetic field and the electron velocity.   This is illustrated in Fig. 2.1.   Since the force exerted upon the electron by the magnetic field is normal to the velocity of the electron, the magnitude of the electron velocity is unchanged by the magnetic field.   However, the force F causes the direction of the electron velocity to change, the force F always remaining normal to the magnetic field and to the electron velocity.
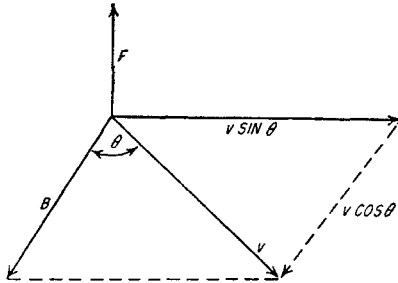
FIG. 2.1. Force exerted upon an electron of velocity v in a uniform magnetic field.

For the case where the electron velocity is entirely normal to a uniform magnetic field, the force exerted upon the electron by the magnetic field causes the electron to move in a circular path as shown in Fig. 2.2a.   The radius R of the electron path is

$$R = \left| \frac{M_e v}{eB} \right| \qquad \text{meters} \qquad (2.11)$$

where R = radius, m

    M_e = mass of electron, kg

    v = velocity of electron normal to magnetic field, m/sec

    e = charge of electron, coulombs

    B = magnetic flux density, webers/m²

At this radius of curvature the centripetal force Bev is equal to the centrifugal force $M_e v^2/R$.

The period of one revolution of the electron in its circular path is

$$T = 2\pi \left| \frac{M_e}{eB} \right| \quad \text{sec} \tag{2.12}$$

If an electron also has a component of velocity parallel to the magnetic field, the electron will traverse a spiral path of constant radius R and pitch P where

$$R = \left| \frac{M_e v \sin \theta}{eB} \right| \quad \text{meters} \tag{2.13}$$

$$P = 2\pi \left| \frac{M_e v \cos \theta}{eB} \right| \quad \text{meters} \tag{2.14}$$

where θ = angle between velocity vector v and magnetic field vector B

This is illustrated in Fig. 2.2b.

(a) THE PATH OF AN ELECTRON HAVING AN INITIAL VELOCITY v NORMAL TO THE MAGNETIC FIELD. DIRECTION OF B OUT OF PAGE

(b) THE PATH OF AN ELECTRON HAVING A COMPONENT OF INITIAL VELOCITY PARALLEL TO THE MAGNETIC FIELD. DIRECTION OF B OUT OF PAGE
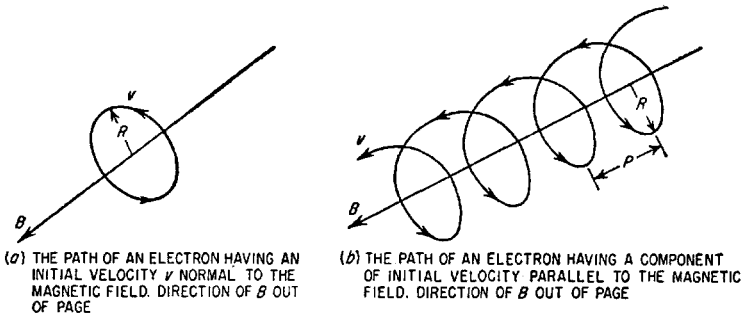
FIG. 2.2. Paths of electrons in a uniform magnetic field.

The use of the equations for the motion of an electron in a uniform magnetic field is illustrated by the following example.

Example 2.2

An electron is injected into a uniform magnetic field of 0.1 weber/m² with an initial velocity of 10^4 m/sec. The initial direction of the electron is 30° from the direction of the magnetic field. Find the radius and the pitch of the spiral electron path.

Solution

1. Determine the radius of the electron path from Eq. (2.13).

$$\begin{aligned} R &= \left| \frac{M_e v \sin \theta}{eB} \right| \\ &= \frac{9.11 \times 10^{-31} \times 10^4 \times 0.5}{1.60 \times 10^{-19} \times 0.1} \\ &= 2.85 \times 10^{-7} \text{ meter} \end{aligned}$$

2. Determine the pitch of the electron path from Eq. (2.14).

$$P = \left| \frac{2\pi M_e v \cos \theta}{eB} \right|$$
$$= \frac{2\pi \times 9.11 \times 10^{-31} \times 10^4 \times 0.866}{1.60 \times 10^{-19} \times 0.1}$$
$$= 3.1 \times 10^{-6} \text{ meter}$$

2.1c. Motion of Electrons in Combined Uniform Electric and Magnetic Fields. The motion of an electron in combined electric and magnetic fields can be determined by the solution of Eq. (2.1) for the combined electrostatic and magnetic forces. Several special cases yield comparatively simple results which are of interest.

CASE I. ε PARALLEL TO B, v_0 PARALLEL TO ε AND B. For this case the electron motion is unaffected by the presence of the magnetic field. The electron velocity is given by Eq. (2.6) for an initial velocity of v_0 and by Eq. (2.4) if v_0 is zero.

CASE II. ε NORMAL TO B, $v_0 = 0$. For this condition the electron will travel a cycloidal path as described in Fig. 2.3.

CASE III. ε NORMAL TO B, v_0 NORMAL TO ε AND B. For the direction of the ε, B, and v_0 vectors as indicated in Fig. 2.4, the magnetic field will exert a downward force
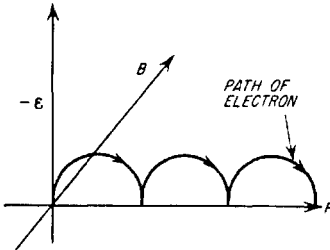
FIG. 2.3. Path of an electron having zero initial velocity in perpendicular magnetic and electric fields.
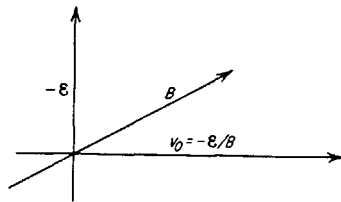
FIG. 2.4. Direction and magnitude of initial electron velocity necessary for the electric and magnetic field forces to cancel.

on the electron while the electric field will exert an upward force on the electron. If these forces are equal, the path of the electron and its velocity will remain unchanged. This condition is satisfied when the electron has an initial velocity given by

$$v_0 = -\varepsilon/B \qquad \text{m/sec} \qquad (2.15)$$

where $\varepsilon = -$ potential gradient, volts/m

B = magnetic-flux density, webers/m²

2.2. Electron Emission. In a conductor, certain of the valence electrons of each atom are displaced from their normal position by the molecular arrangement of the atoms. The forces restraining these displaced electrons are relatively small, and, as a consequence, these electrons can move freely throughout the conductor. These electrons are known as free electrons, and current flow in a conductor consists of the movement of these free electrons within the conductor. The thermal energy of the conductor causes the free electrons as well as the atoms and molecules of the material to be in constant motion in the conductor. As the absolute temperature of the conductor is increased, the velocity, and hence the kinetic energy, of the electrons is increased.

2.2a. Work Function. When a free electron in a conductor, as a result of thermal agitation, acquires a velocity of such a direction and magnitude as to cause it to pass through the surface of the conductor, the conductor is left with a net positive charge on its surface. The attractive force between the positive conductor and the negative

electron will return the electron to the conductor unless it possesses sufficient initial velocity to overcome this attractive force. The kinetic energy which the electron must possess in order to escape from the surface of a conductor is called the *work function W*. The work function is normally given in electron volts. Table 2.1 contains the values of the work function W of several conductors.

At room temperatures the kinetic energy of the free electrons in a conductor due to thermal agitation is insufficient to allow any appreciable number of electrons to escape from the surface of the conductor. However, electrons will be emitted in larger numbers from the surface of the conductor if the energy level of the free electrons is raised by additional energy from another source such as incident light, X rays, high-velocity particles striking the surface, or by increased temperature of the conductor.

TABLE 2.1. APPROXIMATE WORK FUNCTION OF SEVERAL CONDUCTORS

Metal	W electron volts	A amp/cm² deg²
Barium.....................	2.52	60
Carbon....................	4.7	60.2
Calcium...................	3.2	60.2
Copper....................	4.1	65
Molybdenum...............	4.4	60.2
Nickel.....................	5.03	26.8
Oxide-coated nickel..........	1.0–2.0	0.01
Platinum..................	6.0	32
Tantalum..................	4.06	37.2
Thorium...................	3.35	37
Tungsten..................	4.52	60.2
Thorium on tungsten........	2.63	60.2

2.2b. Thermionic Emission. Thermionic emission is the emission of electrons from the surface of a conductor because of the increased energy level of the free electrons in the conductor caused by increasing the temperature of the conductor. The rate of emission of electrons from the surface of any material is related to the absolute temperature and the work function of the material by Eq. (2.16).

$$J = 1,000 A T^2 \epsilon^{eW/kT} \tag{2.16}$$

where J = electron current, ma/cm² of emitter surface
A = constant dependent upon type of emitter (see Table 2.1)
W = work function of emitter, electron-volts
T = absolute temperature, °K
k = Boltzman's universal gas constant = 1.380×10^{-23} joules/°K
e = electron charge = 1.60×10^{-19} coulomb
ϵ = 2.718

Equation (2.16) is valid provided that the emitter is in a vacuum and that another electrode is present in the vacuum which has a potential sufficiently positive with respect to the emitting surface to collect all of the emitted electrons.

A measure of the efficiency of a thermionic emitter is given by

$$\eta_e = \frac{J}{P} \tag{2.17}$$

where η_e = emission efficiency, ma/watt

      J = emission current, ma/cm^2

      P = heating power applied to emitter, watts/cm^2

2.2c. Types of Emitters. Thermionic emission of electrons is quite small in any material at temperatures below about 1000°K. A satisfactory emitter must provide high emission efficiency together with long life. The requirements of high operating temperature, high emission efficiency and/or life limit the number of practical emitters to a few substances. They are tungsten, thoriated tungsten, and oxide-coated nickel or nickel alloy.

Tungsten. Although the work function of tungsten is quite high compared to other metallic emitters (see Table 2.1), tungsten has a higher melting point than any other metal (3643°K) and can be operated at sufficiently high temperatures to obtain a very large emission per square centimeter of emitter before evaporation of the metal becomes a serious factor in limiting the life of the emitter. Tungsten emitters are normally operated at temperatures between 2400 and 2600°K. Tungsten is commonly used as an emitter only in high-power tubes where the anode voltage is above 3,500 volts. In tubes of this type, the emitter is subject to bombardment by positive ions resulting from the ionization of any residual gas molecules in the tube. These positive ions have extremely high energy because of the high anode voltage. This bombardment has no appreciable effect on the life of a pure tungsten emitter, but greatly reduces the life of other emitters.

Thoriated Tungsten. A monatomic layer of thorium deposited on the surface of tungsten produces a thermionic emitter which has a work function considerably less than that of either pure thorium or pure tungsten alone (see Table 2.1). This occurs because thorium is electropositive with respect to tungsten and an intense electric field is present at the boundary between the two metals. Therefore a thoriated-tungsten emitter will emit several thousand times as many electrons per unit surface area as pure tungsten operating at the same temperature. In order to form the monatomic layer of thorium on the surface of the tungsten, the emitter containing thorium oxide is carbonized and then heated to 2600 to 2800°K for 1 or 2 min to cause the carbon to reduce some of the thorium oxide to thorium. It is then operated at about 2100°K for 15 to 30 min to allow some of the thorium to diffuse through the tungsten to the surface. The thorium which is evaporated from the surface of the emitter is continually replaced by diffusion of thorium from the interior of the tungsten to the surface.

In order to minimize evaporation of thorium from the surface of the emitter and to reduce the effects of positive-ion bombardment, a surface layer of tungsten carbide is placed over the thin thorium layer. This is accomplished by heating the emitter in an atmosphere of hydrocarbon vapor. This outer carbonized surface allows the emitter to be operated at considerably higher temperature than could otherwise be done without excessive thorium evaporation. The normal operating temperature for such an emitter is about 1900°K.

Oxide-coated Emitters. In the oxide-coated emitter, a core of nickel or nickel alloy is covered with a combination of barium and strontium carbonates. When this emitter is activated, it has a work function of only 0.5 to 1.5, which provides very high emission currents at temperatures of 1000 to 1150°K. The activation process consists of heating the emitter to about 1500°K under vacuum, which reduces the carbonates to oxides and liberates carbon dioxide which must be removed. The carbon dioxide can be removed by vacuum pumping, but more often a metal "getter" is vaporized within the tube to combine chemically with the carbon dioxide. The emitter is then heated to about 1150°K, and a positive potential of about 100 volts is applied to the anode through a high resistance. Under this condition, particles of pure barium and strontium are produced which diffuse to the surface, and the

electron emission builds up to a steady value. Although the exact mechanism of emission is not completely understood, it is generally believed that the emission current is supplied directly by the free metal which has diffused to the surface or by semiconductor action made possible by the presence of the free metal at the cathode surface. Oxide-coated emitters are normally operated at 1000 to 1100°K.

Because oxide-coated emitters provide a higher emission current per watt of heating power, they are used whenever possible. However, positive-ion bombardment severely damages the emitter surface, and for this reason oxide-coated emitters are only used in tubes having anode voltages less than about 1,000 volts except in special cases.

Oxide-coated cathodes are capable of emitting very large quantities of electrons for periods of several microseconds. For this reason, they are used exclusively in high-power pulse magnetrons where the average current may be only 10 ma while the peak current may be as high as 10 to 20 amp. Table 2.2 gives a comparison of the three types of thermionic emitters.

(a) TYPICAL FILAMENT CATHODES   (b) TYPICAL INDIRECTLY HEATED CATHODES
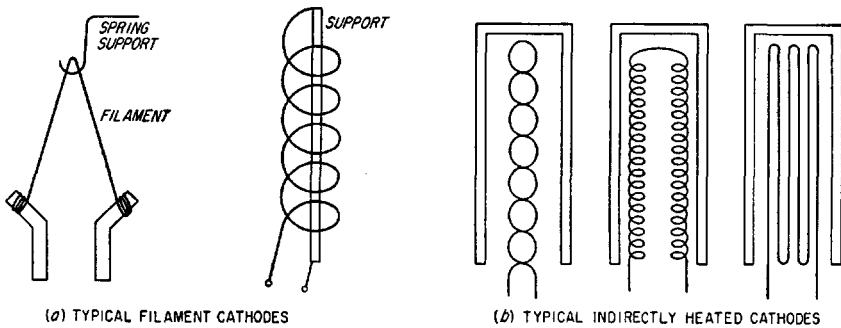
FIG. 2.5. Examples of cathode construction.

The foregoing discussion on emission is valid only if all the emitted electrons are collected by another electrode of sufficiently high positive voltage to prevent the formation of any space charge within the tube (see Secs. 2.3a and 2.3b).

2.2d. Vacuum-tube Cathode. The emitter in a vacuum tube is known as the cathode. The cathodes used in vacuum tubes are of two types: filament cathodes and indirectly heated cathodes. Typical configurations of filament cathodes are shown in Fig. 2.5a. The filament is heated by passing a current through it. Tungsten and thoriated-tungsten cathodes are always of the filament type, as are some oxide-coated cathodes. Most oxide-coated cathodes, however, are of the indirectly heated type. Several types of indirectly heated cathode construction are shown in Fig. 2.5b. The cathode

TABLE 2.2. COMPARISON OF THE PROPERTIES OF COMMONLY USED EMITTERS

Emitters	Normal operating temperature, °K	Emission efficiency, ma/watt	Emission current, ma/cm²
Tungsten filament...................	2400–2600	2–10	100–1000
Thoriated tungsten filament..........	1900	50–100	700–3000
Oxide-coated filaments...............	1000–1100	200–1000	400–3000
Oxide-coated indirectly heated cathodes	1000–1100	10–200	1000–3000

consists of a metal cylinder, usually nickel, which is coated with oxides as described in Sec. 2.2c. Inside the cylinder is placed a heating filament which is electrically insulated from the cathode cylinder.

The filament cathode has the disadvantage that if it is heated with alternating current, the resultant electric field and alternating voltage drop across the filament produce variations in the tube plate current. Also, in a circuit using several tubes, the tubes must all be operated at the same cathode potential or separate filament sources must be employed. In the indirectly heated cathode tubes, the cathode has a unipotential surface and the fields produced by the heater are shielded by the metal cylinder. Most indirectly heated cathodes have adequate electrical insulation between the cathode cylinder and the heater to allow 100 volts or more difference in potential between the cathode cylinder and the heater. This is sufficient in most applications to allow all tubes in a circuit to be operated from a common heater source.

The filament-type cathode has the advantages of fast warmup time (for small receiving tubes approximately 1 to 2 sec compared to about 10 sec for indirectly
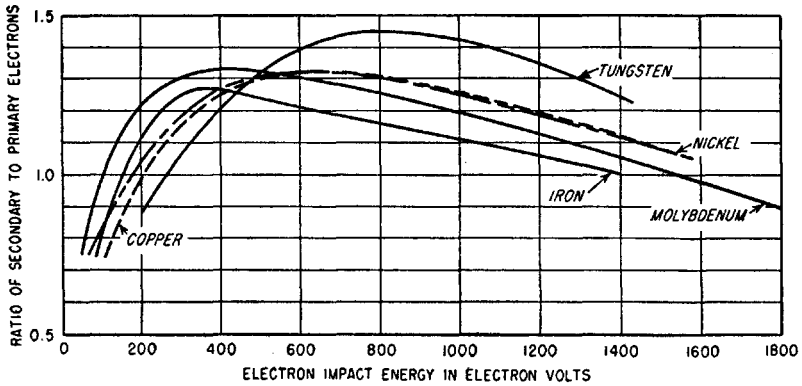
FIG. 2.6. Secondary-emission characteristics of various metals. (*Electronics.*)

heated cathodes) and higher emission efficiency because of the lower heat loss in the filament cathode due to its small mass.

2.2e. Secondary Emission. Secondary emission is the release of electrons from the surface of a material which is bombarded by electrons. Both metals and insulators exhibit secondary emission. The ratio of the average number of secondary electrons to the number of primary electrons striking the surface of a material is dependent upon the velocity of the primary electrons and the type of material. This ratio is given in Fig. 2.6 for a number of common materials as a function of the energy of the primary electrons expressed in electron volts. The average velocities of the secondary electrons will always be less than the average velocity of the primary electrons. A typical velocity distribution of secondary electrons is shown in Fig. 2.7. The small percentage of secondary electrons which have nearly the same velocity as the primary electrons are not really secondary electrons but primary electrons which have been elastically reflected from atoms near the surface of the plate. The ratio of secondary to primary electrons increases appreciably as the angle of incidence of the primary electrons is increased from the normal.

Certain compounds, notably alkali halides on a base of the alkali metal and alkali oxides on various metal bases, are capable of emitting a high ratio of secondary to primary electrons. The secondary-emission characteristics of several of the alkali oxides are shown in Fig. 2.8.

A — SECONDARY ELECTRONS HAVING ENERGY LEVELS OF 20 ELECTRON VOLTS OR LESS. APPROXIMATELY 90 PERCENT OF THE SECONDARY ELECTRONS FALL IN THIS REGION

B — SECONDARY ELECTRONS HAVING ENERGY LEVELS BETWEEN 20 ELECTRON VOLTS AND 98 PERCENT OF THE PRIMARY ELECTRON ENERGY LEVEL. APPROXIMATELY 7 PERCENT OF THE SECONDARY ELECTRONS FALL IN THIS REGION

C — SECONDARY ELECTRONS HAVING ENERGY LEVELS 98 TO 100 PERCENT OF THE PRIMARY ELECTRON ENERGY LEVEL. APPROXIMATELY 3 PERCENT OF THE EMITTED ELECTRONS FALL IN THIS REGION

RELATIVE NUMBER OF SECONDARY ELECTRONS

RATIO OF SECONDARY TO PRIMARY ELECTRON ENERGIES

FIG. 2.7. Typical velocity distribution of secondary electrons when the primary electron energy exceeds 50 electron volts. (*K. R. Spangenberg, "Vacuum Tubes," McGraw-Hill Book Company, Inc., New York,* 1948.)

RATIO OF SECONDARY TO PRIMARY ELECTRONS

$Cs-Cs_2O-Ag$

$Rb-Rb_2O-Ag$

$K-K_2O-Ag$

$Na-Na_2O-Ag$

PRIMARY BEAM VOLTAGE
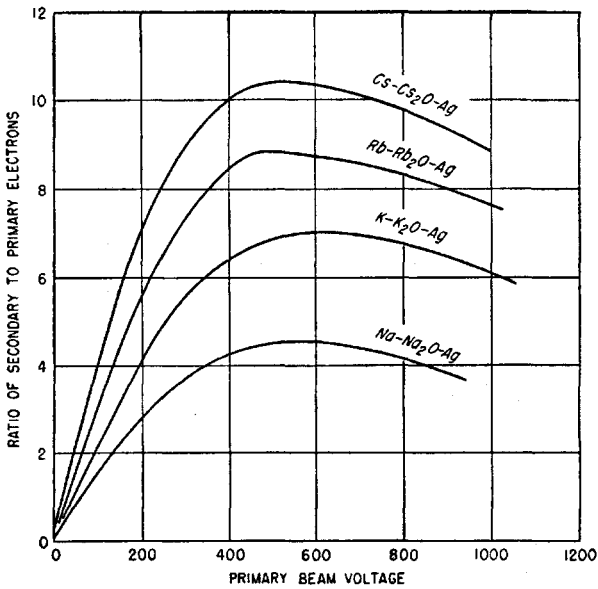
FIG. 2.8. Secondary-emission characteristics of several alkali oxides.

2.3. Diodes. A diode consists of an evacuated envelope containing an emitter called the cathode and one other electrode called the anode or plate which is separated a short distance from the emitter.

2.3a. Space Charge. When the plate of a diode is connected to a voltage source which is positive with respect to the cathode, the electrons emitted by the cathode will travel to the plate because of the force exerted on the electrons by the electric field existing between the cathode and plate (see Sec. 2.1a). However, because of their negative charge, the electrons in the space between the cathode and plate exert an opposing force on the electrons on the surface of the cathode. For a given cathode-plate potential, the space current flowing from cathode to plate will increase until the electric-field force at the cathode surface due to the plate potential is slightly exceeded by the opposing force created by the electrons already in transit between the cathode and plate. In this equilibrium condition the space current is constant because, on the time average, only enough electrons leave the cathode to compensate for those that strike the plate. The charge of all the electrons in transit between cathode and plate is known as the space charge, and under the above condition the diode current is said to be space-charge-limited.
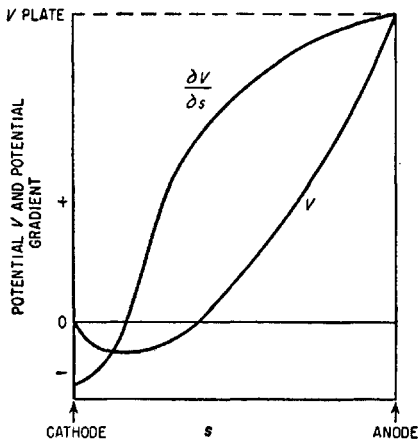
FIG. 2.9. Potential and potential gradient in a parallel-plane diode whose current is space-charge-limited.

The resultant potential gradient at the cathode surface is slightly negative when the diode current is space-charge-limited because of the initial velocities of the emitted electrons. Since the average velocity of emitted electrons is greater than zero, electrons will be emitted even when the resultant field at the cathode surface is zero, thus increasing the space current and producing a negative potential gradient at the cathode surface. Equilibrium is reached when the resultant force of the negative potential gradient is sufficient to return to the cathode all emitted electrons which have an initial velocity below the value which just allows the space charge creating the force to sustain itself, i.e., the number of emitted electrons having an initial velocity sufficient to overcome the negative force adjacent to the cathode is just enough to maintain the space current at a constant value. The potential distribution and potential gradient of a parallel-plane diode are shown in Fig. 2.9.

2.3b. Child's Law. When the current in a diode is space-charge-limited, the current is independent of the temperature of the emitter provided that the temperature is high enough to supply more than the necessary space current. The excess electrons emitted because of any higher emitter temperature are all returned to the cathode by the negative potential gradient near the cathode surface. However, if the plate voltage is increased, the space current will also increase until a new equilibrium condition is reached. The relation between the plate voltage and space current in a diode was first established by Child and is known as Child's law.[1] If the following assumptions are made, the relation is given by Eq. (2.18).

1. The thermionic emission from the cathode is great enough so that the space current is limited by space charge only.

2. The cathode and plate are parallel plane surfaces whose areas are large compared to their spacing.

[1] C. B. Child, *Phys. Rev.*, vol. 32, p. 492, May, 1911.

3. The cathode and anode are equipotential surfaces.

4. The electrons emitted from the cathode have zero emission velocity.

5. The electrons emitted from the cathode have no collisions with gas molecules between electrodes.

Under these conditions,

$$J_b = 2.33 \times 10^{-6} \frac{E_b^{3/2}}{d^2} \qquad (2.18)$$

where J_b = space current density, amp/cm^2

E_b = plate voltage, volts

d = spacing between cathode and plate, cm

2.3c. Temperature-limited and Space-charge-limited Diode Current. From the previous sections it is evident that two conditions of operation can exist in a diode: one, when the diode plate voltage is sufficiently high and the cathode temperature sufficiently low so that the current between cathode and anode is determined primarily by the cathode temperature; and the other, when the cathode temperature
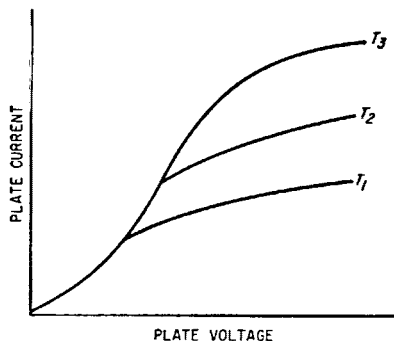
Fig. 2.10. Diode plate current as a function of plate voltage for various cathode temperatures.
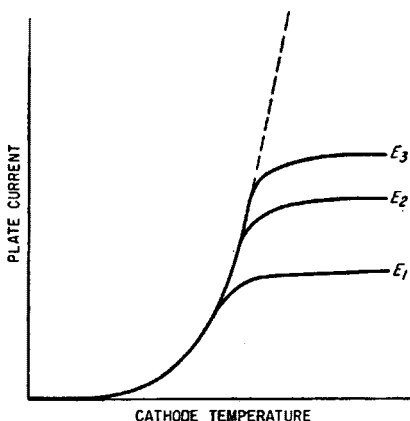
Fig. 2.11. Diode plate current as a function of cathode temperature for various plate voltages.

is sufficiently high and the plate voltage sufficiently low so that the current between cathode and anode is determined primarily by the plate voltage. The first condition is known as *temperature saturation* and is shown in Fig. 2.10. When the emitting surface is subject to an appreciable positive-voltage gradient, the space current continues to rise with plate voltage beyond the point where temperature saturation occurs because the work function of the emitter is effectively reduced by the electrostatic field present at the cathode surface. This is known as the Schottky effect. The decrease in work function as the field is increased can be taken into account by modifying Eq. (2.16) by the appropriate factor. Thus,

$$J_E = J_b \epsilon^{\frac{4.4\sqrt{-\varepsilon}}{T}} \qquad (2.19)$$

where J_E = emission current density taking potential gradient at cathode into account

J_b = emission current density given by Eq. (2.16)

ε = −potential gradient at emitter surface, volts/cm

T = absolute temperature, °K

The second condition is known as *voltage saturation* and is shown in Fig. 2.11. The diode current is limited by space-charge effects as described in Secs. 2.3a and 2.3b.

In normal use, a diode is operated under conditions of voltage saturation, where Child's law is applicable.

2.3d. Plate Heating. Each electron upon striking the plate of a diode transfers all of its kinetic energy to thermal energy unless it causes the release of a secondary electron. The plate of a diode or other vacuum tube must be capable of dissipating the heat released by the impact of all the electrons forming the tube plate current. Neglecting the small energy transferred to secondary electrons, the total energy transformed to heat at the plate in a given time interval assuming a constant plate voltage is given by

$$\text{KE} = n \frac{M_s v^2}{2} = -nVe \tag{2.20}$$

where V = plate-cathode potential of the diode

n = total number of electrons received in specified time interval

Power is the rate of expenditure of energy. Therefore the power transferred to the plate in the form of heat is given by

$$P = VI \qquad \text{watts} \tag{2.21}$$

where V = plate-cathode voltage, volts

I = plate current, amp = rate of change of charge = $e \dfrac{dn}{dt}$

When the plate voltage and plate current vary as a function of time, the average power dissipated at the plate in the period T is given by

$$P_{\text{av}} = \frac{1}{T} \int_0^T ei \, dt \tag{2.22}$$

where e = instantaneous plate voltage

i = instantaneous plate current

The maximum power which can be dissipated by the plate of a tube is determined by the rate at which heat can be removed from the plate and the maximum allowable plate temperature. The maximum plate temperature is limited by three factors: amount of gas released from the plate material at high temperatures, allowable maximum temperature of the glass envelope, and the melting point of the plate material. Heat is removed from the plate by conduction through the plate supporting structure and by radiation from the plate.

2.3e. Edison Effect. If the plate terminal of a diode is connected to the cathode terminal through an ammeter without any potential applied to the plate, but with the cathode at its operating temperature, a small current will flow through the ammeter. This phenomenon was first observed by Edison and is often referred to as *Edison effect.* The current exists because the initial velocity of the electrons emitted from the cathode surface causes a small percentage of the electrons to arrive at the plate. In order completely to cut off the current between cathode and plate, a negative voltage of approximately 1 or 2 volts must be applied to the plate.

2.3f. Diode Characteristics. The characteristics of a diode are completely described by a plot of plate current as a function of plate voltage. This plot is called the E_b-I_b characteristic. A typical receiving-type diode E_b-I_b characteristic curve is shown in Fig. 2.12.

The *static plate resistance* R_p of the diode is defined as

$$R_p = \frac{E_b}{I_b} \qquad \text{ohms} \tag{2.23}$$

where E_b = plate voltage, volts

I_b = plate current, amp

The *dynamic plate resistance* r_p of the diode is defined as the inverse of the slope of the E_b-I_b characteristic curve at any point. Thus,

$$r_p = \frac{\partial E_b}{\partial I_b} \simeq \frac{\Delta E_b}{\Delta I_b}\bigg|_{T=\text{const}} \quad \text{ohms}$$

(2.24)

where T = cathode temperature

In Fig. 2.12, the dynamic plate resistance of the diode at a current of 5 ma is given by

$$r_p \simeq \frac{(30-23)}{(6-4)\times 10^{-3}} = 3,500 \text{ ohms}$$

If a diode is placed in a circuit as shown in Fig. 2.13, the following equation relating the loop voltage and current may be written

$$E_{bb} = I_b R_b + E_b \qquad (2.25)$$

Fig. 2.12. E_b-I_b characteristics of a typical diode.

If an equation is written relating the diode current and voltage as shown graphically by the E_b-I_b characteristic curve, this equation and Eq. (2.25) could be solved simultaneously to determine E_b and I_b. A graphical solution of the two equations is much
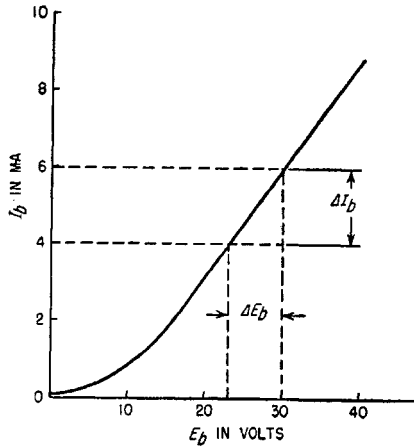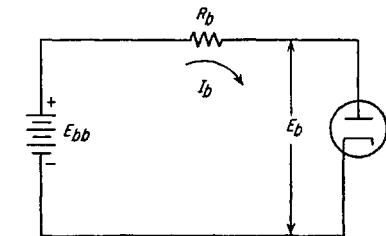
FIG. 2.13. Diode in series with battery and resistor.
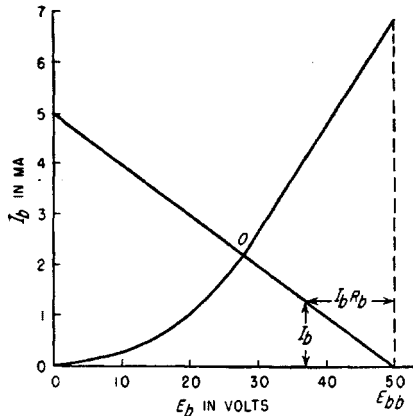
FIG. 2.14. Construction of the d-c load line for a diode.

more easily made, however. The E_b-I_b curve of the diode in Fig. 2.13 is shown in Fig. 2.14. This is a plot of the E_b-I_b equation for the diode. The sum of the voltage drop E_b across the diode and the voltage drop $I_b R_b$ across R_b must equal the supply voltage E_{bb}. Therefore, if the voltage drop across R_b is plotted as a function of the current through R_b on the diode E_b-I_b characteristic as shown in Fig. 2.14, the value of current for which Eq. (2.25) is valid is readily determined as the current at point O. The plot of the voltage drop across the load resistor R_b as a function of diode current I_b is called the *d-c load line*. The point O, which is the solution of the two simultaneous equations, is called the quiescent operating point. The slope of the load

line is $-1/R_b$.   The construction of d-c load lines is illustrated by the following example.

Example 2.3

A 5U4-G diode rectifier is connected to a 100-volt supply through a 500-ohm load resistor as shown in Fig. 2.15a.   Determine the quiescent operating point, the static and dynamic

(a) CIRCUIT OF A
5U4-G DIODE &
RESISTOR LOAD

(b) DIODE CHARACTERISTIC CURVE AND LOAD LINE
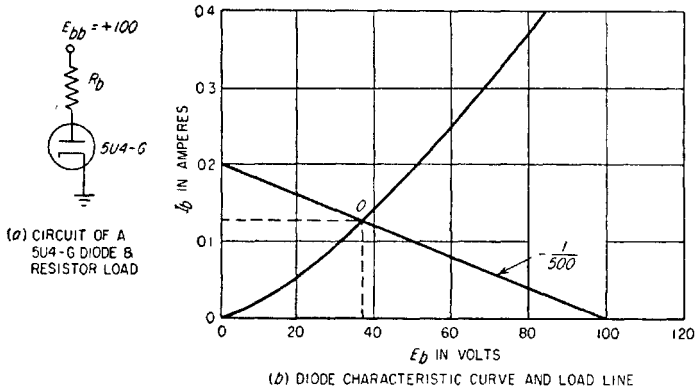
FIG. 2.15. 5U4-G diode rectifier with resistor load.

plate resistance of the diode at the quiescent operating point, and the plate dissipation of the diode.   The E_b-I_b curve of the 5U4-G is given in Fig. 2.15b.

Solution

1. Construct the load line on the E_b-I_b curve.
Since the diode load is 500 ohms, the load line has a slope of $-\frac{1}{500}$ and starts at the supply voltage of $+100$ volts.   The load line is constructed in Fig. 2.15b.
2. Determine the quiescent operating point.
The quiescent operating point is the intersection of the E_b-I_b curve and the load line. Point O in Fig. 2.15b is the operating point.
3. Determine the static diode resistance.
The static resistance, as given by Eq. (2.23), is

$$R_p = \frac{37.0}{0.125}$$
$$= 296 \text{ ohms}$$

4. Determine the dynamic plate resistance, r_p.
From Eq. (2.24)

$$r_p = \frac{\Delta E_b}{\Delta I_b}\bigg|_{T=\text{const}}$$
$$= \frac{40 - 32}{0.140 - 0.100} = 200 \text{ ohms}$$

FIG. 2.16. Equivalent circuit of diode.

5. Determine the plate dissipation of the diode.
From Eq. (2.21)

$$P = 37 \times 0.125$$
$$= 4.63 \text{ watts}$$
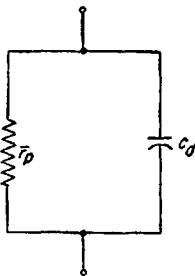
2.3g. Diode Equivalent Circuit.   The a-c equivalent circuit of a diode is shown in Fig. 2.16.   It consists of a resistance $\bar{r}_p$ in parallel with a capacitor c_d.   The capacitance c_d is the capacity between the cathode and plate of the diode.   The resistance $\bar{r}_p$ is the average value of the diode dynamic plate resistance averaged over one cycle of the applied signal.   Thus,

$$\bar{r}_p = \frac{1}{2\pi} \int_0^{2\pi} r_p \, d\theta \qquad (2.26)$$

where r_p = instantaneous value of dynamic plate resistance as a function of θ

θ = angle of fundamental component of applied signal, radians

When a small a-c voltage is superimposed upon a larger d-c voltage across the diode, the diode resistance will be very nearly constant throughout the cycle of the a-c voltage and equal to the value of r_p at the d-c operating point.

At frequencies of about 100 Mc or greater, depending upon the particular diode, the transit time of the electrons traveling from cathode to plate will become an appreciable portion of one cycle of the applied voltage. As a result of this delay, r_p will change, becoming infinite when the transit time is equal to the period of one cycle of the applied voltage, and the effective capacitance c_d will increase. At higher frequencies r_p will oscillate between negative and positive values.[1] Diode tubes are used as power rectifiers, mixers, and detectors in receivers and in other specialized applications requiring a nonlinear element.

2.4. Triode Tubes. A triode is a vacuum tube in which a control electrode has been placed between the cathode and the plate. This electrode is called the grid, or
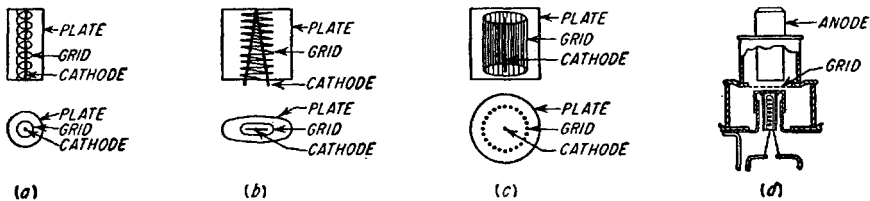
FIG. 2.17. Typical triode-tube configurations.

control grid. The potential applied to this grid controls the magnitude of the cathode-plate current. The grid achieves its control of plate current by controlling the electrostatic field near the cathode. Several configurations of triode tubes are shown in Fig. 2.17.

2.4a. The Action of the Control Grid. In Sec. 2.3a the action of space charge in controlling the plate current of a diode was explained. The current adjusts itself to the level where the negative potential gradient near the cathode surface due to space charge slightly exceeds the positive potential gradient due to the positive plate voltage. The number of emitted electrons which have initial velocities great enough to overcome the force of the slightly retarding field at the cathode surface is just sufficient to maintain the space charge at the equilibrium value.

When a control grid is inserted between the cathode and plate with close spacing between the cathode and the grid, a relatively small grid-cathode potential can produce a field at the cathode surface which will overcome the field due to the plate-cathode potential. When the grid potential is made negative with respect to the cathode, fewer of the electrons which are emitted by the cathode have sufficient initial velocity to overcome the increased negative potential gradient. As a result, the space charge increases, causing the space current to decrease until a new equilibrium condition is reached. If the grid potential is made sufficiently negative, the space current will be reduced to zero. The value of grid voltage which just reduces the tube current to zero is called the *cutoff voltage* E_{co}. Above the cutoff voltage the

[1] For greater detail on the effects of electron transit time in diodes see K. R. Spangenberg, "Vacuum Tubes," McGraw-Hill Book Company, Inc., New York, 1948.

grid is able to vary the plate current between wide limits with small changes in grid potential. The grid is normally operated at negative potentials with respect to the cathode in order that the grid may not intercept any appreciable number of electrons.

2.4b. Amplification Factor. Because of the close spacing between the grid and cathode, a change in grid voltage has a much greater effect upon plate current than an equal change in plate voltage. The amplification factor μ of a vacuum tube is defined as the ratio of the change in plate current with a change in grid voltage to the change in plate current for an equal change in plate voltage, i.e.,

$$\mu = \frac{\partial I_b / \partial E_c}{\partial I_b / \partial E_b} = - \frac{\partial E_b}{\partial E_c} \tag{2.27}$$

From Eq. (2.18), the total cathode current can be expressed as

$$I_c + I_b = k(E_b + \mu E_c)^{3/2} \tag{2.28}$$

where I_c = grid current
    I_b = plate current
    k = constant dependent upon tube configuration
    E_b = plate voltage
    E_c = grid voltage

When the grid potential is negative with respect to the cathode, the grid current will be zero and Eq. (2.28) reduces to

$$I_b = k(E_b + \mu E_c)^{3/2} \tag{2.29}$$

When the term on the right side of Eq. (2.29) becomes zero, the plate current is reduced to zero and remains zero for negative values of $(E_b + \mu E_c)$. From Eq. (2.29) it would appear that the value of bias necessary to reduce the plate current to zero is

$$E_{co} = - \frac{E_b}{\mu} \tag{2.30}$$

where E_{co} is defined as the projected cutoff bias in volts. However, because of the nonuniformity of the electrostatic field in the tube due to edge effects, and for other practical considerations, the value of cutoff bias required for a particular plate voltage is not sharply defined, and Eq. (2.30) gives a value which is less than the actual bias required.

2.4c. Tube Characteristic Curves. The relationships between plate current, plate voltage, grid current, and grid voltage in a triode tube are best described by a series of curves. The most important set of curves for the triode is the plot of plate current as a function of plate voltage for various values of grid voltage. This is known as the E_b-I_b characteristics, or the plate characteristics of the tube, and the curves for a typical triode are shown in Fig. 2.18. The value of grid current I_c is also indicated on this series of characteristic curves for various values of positive grid voltage. The reversal in the curvature of the lines of constant grid voltage for positive values of grid voltage is due to the grid current which flows when the grid is positive. At zero plate voltage and at positive grid voltage, the total cathode current (except for Edison effect) flows to the grid. The majority of the electrons accelerated by the grid initially pass through the space between the grid wires, approach the plate, and then return to the grid because of its positive potential. As the plate voltage is increased from zero, the plate current increases rapidly because an increasing proportion of the electrons which were initially accelerated by the positive grid voltage are now collected by the plate. As the plate voltage becomes much higher than the

grid voltage, only those electrons initially intercepted by the grid are collected by it, and all of the electrons passing through the grid spaces are collected by the plate. Since the total cathode current is also increasing with increasing plate voltage because of the changing field at the cathode surface, the plate current increases very rapidly at low plate voltages and then approaches the three-halves power relationship expressed in Eq. (2.28). When the grid voltage exceeds the plate voltage, any secondary electrons emitted from the plate are drawn to the grid, thus lowering the net plate current and increasing the grid current.

From the information contained in the E_b-I_b curves, other sets of curves which are sometimes useful can be constructed. One set, the E_c-I_b curves, shows the relation between plate current and grid voltage for various values of plate voltage. A plot of this type is shown in Fig. 2.19. Another plot of the same information which is used in class C amplifier design (see Sec. 4.4) is the E_b-E_c, or constant plate current plot. This is illustrated in Fig. 4.18.
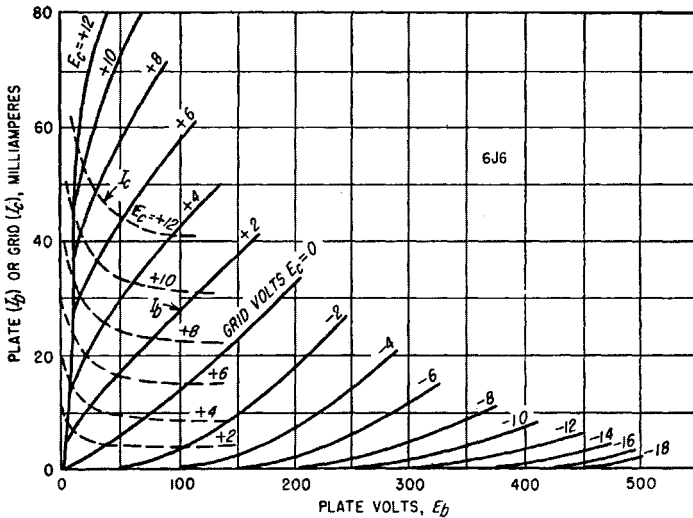
FIG. 2.18. E_b-I_b curves for a typical triode.

The curves in Figs. 2.18 and 2.19 are obtained by making static measurements on the tube as shown in Fig. 2.20. In making measurements of this type, care must be exercised to ensure that the average power dissipation ratings of the grid and plate are not exceeded. Manufacturers' published tube characteristic curves are generally obtained by averaging measurements of a large number of tubes.

2.4d. Tube Coefficients. The characteristics of a particular triode are completely described by the characteristic curves for that tube. After the point of operation of the tube on the characteristic curves has been established, however, it is possible to describe quantitatively the characteristics of the tube in a region near the point of operation by three tube coefficients which are assumed constant in the region of the point. These tube coefficients are the amplification factor μ, the grid-plate transconductance g_m, and the plate resistance r_p.

Amplification Factor μ. The amplification factor μ of a triode was discussed in Sec. 2.4b. It is a ratio describing the relative control of grid voltage and plate voltage upon plate current and is given by Eq. (2.27). The amplification factor of a tube can be determined graphically from the E_b-I_b curves of that tube at any operating point by assuming a small variation ΔE_c in grid voltage and finding the resultant varia-
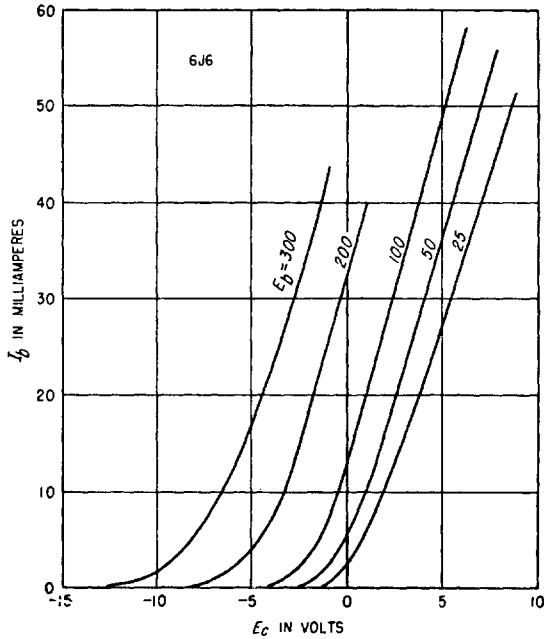
FIG. 2.19. E_b-I_b characteristic curves for a typical triode.
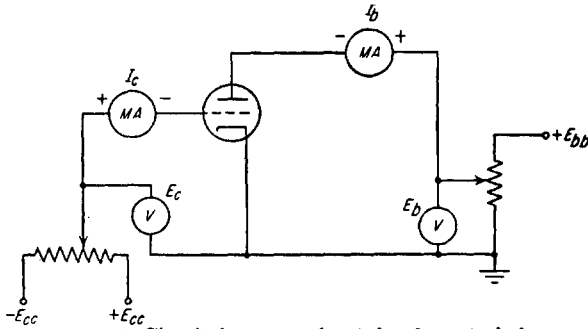
FIG. 2.20. Circuit for measuring tube characteristics.

tion ΔE_b in plate voltage necessary to maintain the plate current constant. The amplification factor is then given by

$$\mu = -\frac{\Delta E_b}{\Delta E_c}\bigg|_{I_b = \text{const}} \tag{2.31}$$

This is illustrated in Fig. 2.21, where μ in the region selected is

$$\mu = -\frac{(175 - 145)}{-6 - (-4)}$$
$$= 15$$

The amplification factor of a tube is dependent primarily upon the grid structure and the ratio of grid-cathode to plate-cathode spacing. Large grid wires and close spacing of the grid wires increase the amplification factor because of increased shielding

between plate and cathode.   The amplification factor of triode tubes ranges from approximately 2 to 300.   In the negative grid region, the amplification factor of a tube would be a constant independent of the grid and plate voltages if the grid and plate voltages had the same ratio of control of the electrostatic field at all parts of the
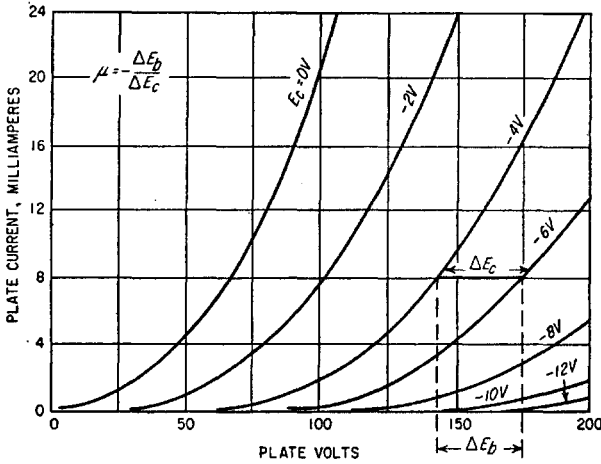
Fig. 2.21. Graphical determination of amplification factor.
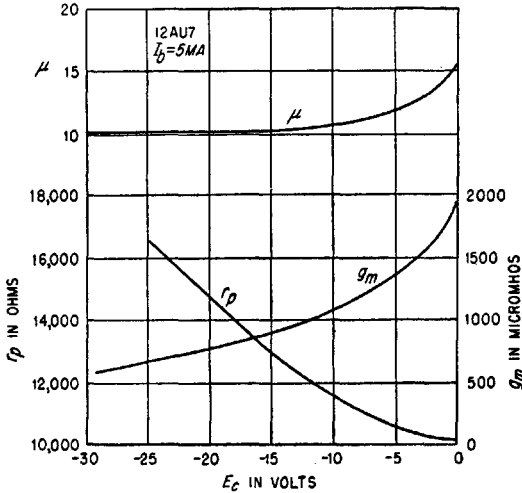
Fig. 2.22. Variations in tube coefficients of a typical low-μ triode as a function of grid bias.

cathode surface.   In practice, however, support wires and edge effects cause different parts of the tube to have different amplification factors, and as a result, the amplification factor is not constant for all grid and plate voltages.   The amplification factor of a typical triode is plotted in Fig. 2.22 as a function of grid voltage for a constant plate current.

Dynamic Plate Resistance r_p.   The dynamic plate resistance r_p of a triode is the ratio of an incremental change in plate voltage to the resultant change in plate current

for a constant grid-cathode voltage. Thus,

$$r_p = \frac{\Delta E_b}{\Delta I_b}\bigg|_{E_c = \text{const}} \qquad (2.32)$$

The dynamic plate resistance is determined from the E_b-I_b characteristic curves as shown in Fig. 2.23. It is the inverse, or reciprocal, of the slope of the curves of constant grid bias at any operating point. The plate resistance in the region selected in Fig. 2.23 is given by

$$r_p = \frac{119 - 78}{(11.2 - 3.2)10^{-3}}$$
$$= 5.13K$$

The static plate resistance R_p of the tube is the ratio of the plate voltage to plate current at the operating point.
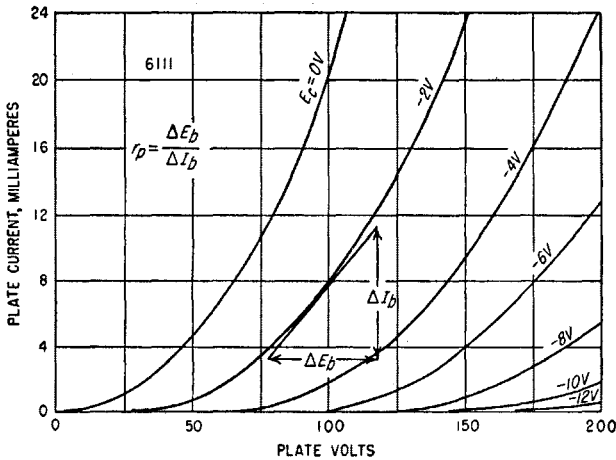
FIG. 2.23. Graphical determination of dynamic plate resistance.

In an ideal tube where the relation between plate current and grid and plate voltages can be expressed by Eq. (2.29), the dynamic plate resistance is given by

$$r_p = \frac{2}{3k^{2/3}I_b^{1/3}} \qquad (2.33)$$

where I_b = a-c plate current
k = constant dependent upon tube configuration
From Eq. (2.33) it is seen that the dynamic plate resistance of an ideal tube is independent of the actual plate and grid voltages and varies inversely as the cube root of the plate current. In practical tubes, however, r_p is not constant as a function of grid and plate voltages, as shown in Fig. 2.22. Dynamic plate resistance varies inversely with cathode surface area and directly with cathode-plate spacing, assuming that as the plate-cathode spacing is changed the grid-cathode spacing is also modified so as to keep the amplification factor of the tube constant. The dynamic plate resistance r_p is plotted as a function of plate current in Fig. 2.24 for a constant plate voltage.

Grid-plate Transconductance. The grid-plate transconductance g_m of a tube is a measure of the effectiveness of the grid in controlling plate current when the plate

voltage is held constant.   It is defined as

$$g_m = \frac{\Delta I_b}{\Delta E_c}\bigg|_{E_b\,=\,\mathrm{const}} \qquad\qquad (2.34)$$

The transconductance can be graphically determined at any operating point from
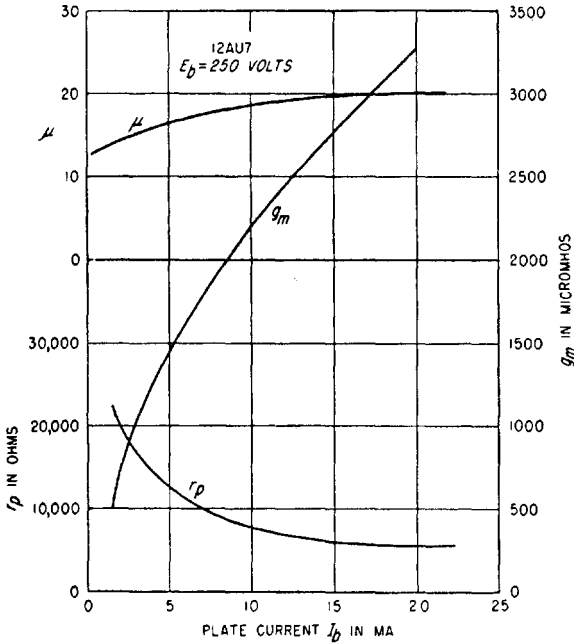
Fig. 2.24. Variations in tube coefficients of a typical low-μ triode as a function of plate current.

the E_b-I_b characteristic curves as shown in Fig. 2.25.   The g_m in the region selected in Fig. 2.25 is

$$g_m = \frac{(8.8 - 7.5)10^{-3}}{(-2.0) - (-2.3)}$$
$$= 4.33 \times 10^{-3}\ \mathrm{mho}$$

The transconductance of a tube can be related to the μ and r_p of the tube by substituting Eqs. (2.31) and (2.32) into Eq. (2.34).   Thus,

$$g_m = \frac{\mu}{r_p} \qquad\qquad (2.35)$$

Since in an ideal triode the amplification factor is independent of the tube voltages and currents, Eqs. (2.33) and (2.35) show that the g_m of an ideal triode varies as the cube root of the plate current.   This is approximately true in practical tubes as is indicated in Fig. 2.24 where g_m is plotted as a function of plate current for a fixed plate voltage.

   Since in most tube applications it is desirable to have a high amplification factor and a low dynamic plate resistance, the g_m of a tube can be used as a measure of the performance capabilities of the tube.

2.4e. Load Lines. When an external impedance is connected in series with a vacuum tube and a voltage source, the voltage across the load impedance and the current through the load impedance can be determined for any value of grid-cathode voltage by the simultaneous solution of the voltage-current relationships in the
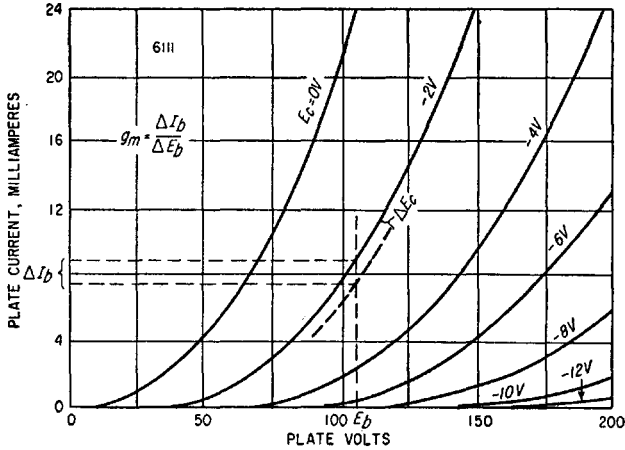
FIG. 2.25. Graphical determination of transconductance in a triode.
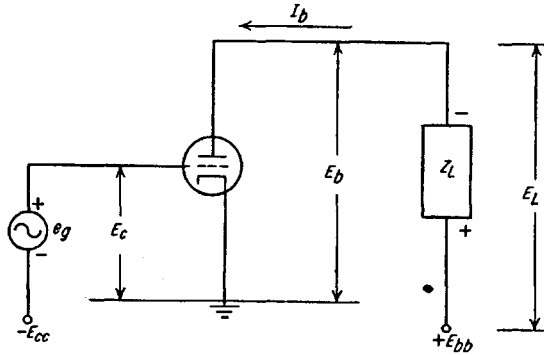
FIG. 2.26. Triode tube and load impedance.

vacuum tube and the load impedance. In an ideal triode, this could be done analytically by the simultaneous solutions of the following equations:

$$I_b = k(E_b + \mu E_c)^{3/2} \qquad (2.36)$$
$$E_L = Z_L I_b \qquad (2.37)$$
$$E_L + E_b = E_{bb} \qquad (2.38)$$

In practical tubes, however, Eq. (2.36) is only approximate, and the solutions are much more easily obtained graphically by utilizing the E_b-I_b tube characteristic curves, which are a graphic representation of Eq. (2.36). The method of construction of load lines is discussed in detail for single tube amplifiers in Secs. 3.3 and 3.4; load lines for push-pull amplifiers are discussed in Secs. 4.2a and 4.2b; and the construction of diode detector load lines is discussed in Sec. 7.7a.

2.4f. Equivalent Circuits for Vacuum Tubes. If the voltage and current variations in a vacuum tube are small enough so that g_m, r_p, and μ can be considered constant

over the region of operation, the vacuum tube can be represented by an equivalent circuit containing linear elements. At any operating point on the E_b-I_b curves, a small change ΔI_b in the tube plate current can be expressed as

$$\Delta I_b = \frac{\partial I_b}{\partial E_c}\,\Delta E_c + \frac{\partial I_b}{\partial E_b}\,\Delta E_b \qquad (2.39)$$

From Eqs. (2.32) and (2.34), this is equivalent to

$$\Delta I_b = g_m\,\Delta E_c + \frac{\Delta E_b}{r_p} \qquad (2.40)$$

If the tube is connected in a circuit as shown in Fig. 2.26, the load impedance Z_L will have a voltage E_L developed across it equal to $I_b Z_L$. If the grid-cathode voltage is varied by an amount ΔE_c of such a polarity as to cause the plate current to increase by an amount ΔI_b, the

FIG. 2.27. Constant-voltage form of equivalent circuit for Fig. 2.26.

voltage drop across the load will increase by an amount ΔE_L. The consequent change in plate voltage ΔE_b must be equal to $-\Delta E_L$ since the sum of the voltages across the tube and the load equals the supply voltage. Thus, Eq. (2.40) can be rewritten as

$$\Delta I_b = g_m\,\Delta E_c - \frac{\Delta E_L}{r_p} \qquad (2.41)$$

or

$$-i_p = g_m e_g + \frac{e_p}{r_p}$$

where $e_g = \Delta E_c$
$i_p = -\Delta I_b$
$e_p = \Delta E_b = -\Delta E_L$

The direction of i_p (see Fig. 2.27) is the assumed direction of signal current flow in the plate circuit. Eq. (2.41) can be rearranged as

$$-\mu e_g = i_p(r_p + Z_L) \qquad (2.42)$$

The d-c grid and plate voltages and currents do not enter directly into the small signal operation of a vacuum tube as indicated in Eq. (2.42). The d-c conditions are implicit in the values of μ, r_p, and g_m existing at the operating point. From Eq. (2.42) the equivalent plate circuit of Fig. 2.27 can be constructed. The tube is replaced by an equivalent voltage generator whose internal impedance is r_p and whose open circuit terminal voltage is $-\mu e_g$. The 180° phase shift between the grid signal voltage and the plate signal voltage is taken into account by the polarity of the equivalent generator and the assumed direction of cur-

FIG. 2.28. Constant-current form of equivalent circuit for Fig. 2.26.

rent flow. It should be remembered that $-\mu e_g$ is only an equivalent voltage generator and does not actually exist. The only a-c voltages actually existing in the plate circuit are e_p and ΔE_L. They are equal in magnitude and opposite in polarity. The a-c current i_p in the plate circuit is caused by the control action of the grid, and the generator $-\mu e_g$ is an equivalent way of expressing this control action.

The circuit of Fig. 2.27 is known as the equivalent constant-voltage form of the

plate circuit because the equivalent circuit contains a voltage generator $-\mu e_g$ having zero internal impedance (see Sec. 23.3). A similar equivalent constant-current form of the plate circuit can be constructed by dividing both sides of Eq. (2.42) by r_p and noting that i_p in Eq. (2.42) is the current through Z_L. Thus,

$$-g_m e_g = i_p \left(1 + \frac{Z_L}{r_p}\right) \tag{2.43}$$

From Eq. (2.43) the equivalent circuit of Fig. 2.28 can be constructed.

2.4g. Vector Diagrams for Vacuum-tube Circuits. By expressing each sinusoidal signal voltage in the constant-voltage form of the equivalent plate circuit of a tube as a vector, a vector diagram can be constructed which shows the relative phase and magnitude of the various voltages in the circuit for any load impedance. Examples for various load impedances are shown in Fig. 2.29. Similar current vector diagrams can be constructed from the constant-current form of the equivalent plate circuit.

2.4h. Tube Capacitances. The close spacing of the electrodes in a tube and the close proximity of the leads extending from the electrodes to the base of the tube
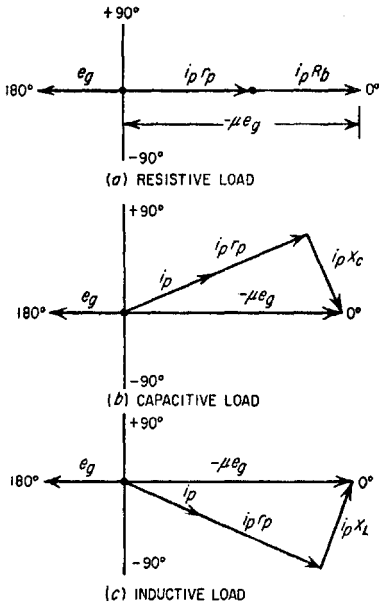
Fig. 2.29. Voltage vector diagrams of a vacuum-tube plate circuit having various load impedances.
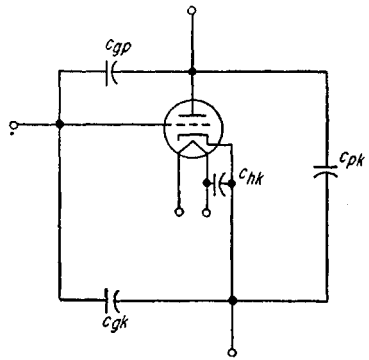
Fig. 2.30. Interelectrode capacitances existing in triode tubes.

cause capacitances to exist between the various electrodes in the tube. The interelectrode capacitances that exist in a triode are shown in Fig. 2.30. They are the capacitance between grid and plate c_{gp}, the capacitance between grid and cathode c_{gk}, the capacitance between cathode and the plate c_{pk}, and the capacitance between heater and cathode c_{hk} in a tube having an indirectly heated cathode. Because the grid-cathode spacing is much smaller than the grid-plate spacing, the grid-cathode capacitance c_{gk} is higher than the grid-plate capacitance c_{gp}. Since the grid shields the cathode from the plate, the cathode-plate capacitance c_{pk} is somewhat smaller than either c_{gk} or c_{pk}. Because of the proximity of the various leads from the electrodes to the base, and because of the various electrode supports within the tube, the differences in the values of c_{gk}, c_{gp}, and c_{pk} are not as great as would be expected from the physical spacing of the electrodes alone. In tubes designed for use at very high frequencies, the conventional socket arrangement is not used and instead the tubes have the electrode leads and the external connections located at well-separated parts of the tube. Typical values of inter-electrode capacitances for octal base

triodes, miniature base triodes, subminiature triodes, and very high frequency triodes are given in Table 2.3. The inter-electrode capacitances c_{gp}, c_{gk}, and c_{pk} of the tube itself are usually specified by the manufacturer of the tube. The value of c_{hk} is ordinarily not given, but for receiving-type triodes it will ordinarily be between 20 and 25 $\mu\mu f$.

TABLE 2.3. TYPICAL VALUES OF INTERELECTRODE CAPACITANCES FOR DIFFERENT TYPES OF TRIODE TUBE CONSTRUCTION

	Octal base	Miniature base	Subminiature	Uhf
Tube type	6SN7	12AU7	6111	416A
c_{gk}, $\mu\mu f$	3.2	1.8	1.9	7.5
c_{gp}, $\mu\mu f$	4.0	1.6	1.5	1.25
c_{pk}, $\mu\mu f$	3.4	1.3	0.3	0.0095

In circuit analysis, the tube inter-electrode capacitances must be added to the tube equivalent circuit together with socket capacitances, wiring capacitances, Miller-effect capacitance, transit-time capacitance, etc. At very high frequencies the self-inductance of the electrode leads must also be taken into account (see Sec. 7.4h).

2.4i. Methods of Grid Bias. In normal operation of vacuum tubes, the grid is operated at a d-c potential that is negative with respect to the cathode. There are

(a) FIXED-GRID BIAS

(b) CATHODE BIAS

(c) CONTACT-BIAS

(d) GRID-LEAK BIAS
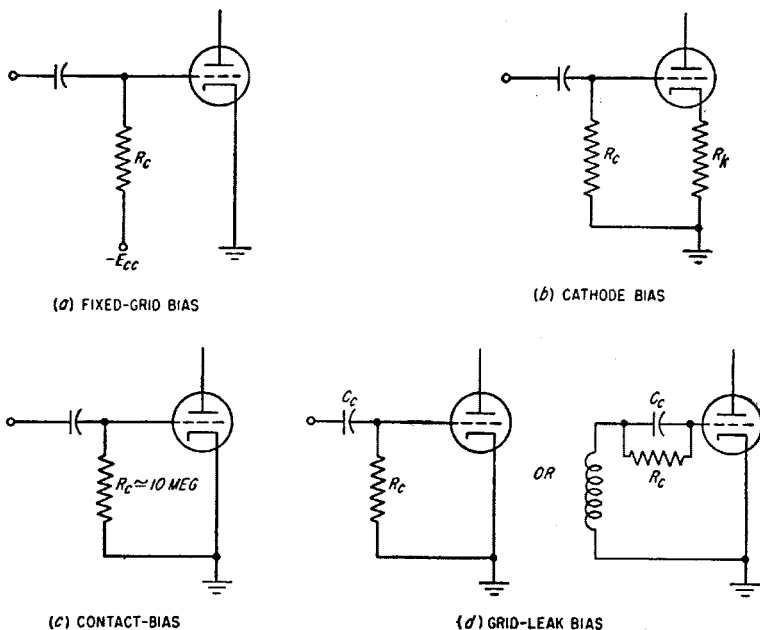
FIG. 2.31. Various methods of obtaining grid bias in vacuum-tube circuits.

four methods by which this negative grid-cathode voltage, known as grid bias, can be obtained.

One method is to return the grid to a voltage source which is negative with respect to the cathode. This is known as *fixed grid bias* and is illustrated in Fig. 2.31a.

A second method is to insert a small resistance between the cathode and the point to which the d-c grid return is made. This is known as *cathode bias* and is illustrated in Fig. 2.31b. The cathode resistor is made large enough so that the plate current through the tube develops the proper value of bias across the cathode resistor. The cathode resistor must be bypassed by a capacitor which has a negligible reactance at any signal frequency if there is to be no loss in amplification due to cathode degeneration (see Sec. 3.5b).

A third method of obtaining grid bias is to insert a high resistance, usually about 10 megohms, between grid and cathode. The bias voltage is developed by those electrons in the tube space charge which strike the grid and return to the cathode through the external resistance. This is known as *contact bias* and is shown in Fig. 2.31c. When the external voltages are first applied to the tube, the grid is at the same potential as the cathode, and some of the emitted electrons will have a high enough initial velocity and the proper direction to strike the grid. The flow of these electrons through the external grid resistor causes the grid to become slightly negative, reducing the number of electrons which have a high enough velocity to reach the grid and causing an equilibrium condition to be reached when the number which reach the grid produce a bias which is just sufficient to allow that number to reach the grid. The value of external grid resistance R_c used when contact bias is utilized is normally between 1 megohm and 10 megohms. Any gas in the tube will tend to offset the negative bias thus produced because some of the electrons in traveling to the plate will strike the gas molecules with sufficient velocity to knock an electron from the gas molecule, forming a positive ion. These positive ions will travel to the negative grid and develop a positive bias across the grid-cathode resistor. In receiving-type tubes the amount of residual gas within the tube is quite small, and the negative contact bias predominates (see Sec. 3.8a). The amount of bias that can be developed by contact bias is usually about one volt with grid resistors of several megohms or more. For this reason, contact bias is only useful in applications where the grid-cathode signal amplitude is quite small.

A fourth method of obtaining grid bias which is often used in oscillators and r-f amplifiers is shown in Fig. 2.31d. This is called *grid-leak bias*. The positive portion of the incoming signal initially raises the grid to a potential which is positive with respect to the cathode, causing grid current to flow which charges the coupling capacitor C_c. The discharge of C_c through R_c is made very nearly constant by making the time constant $C_c R_c$ long compared to the period between the positive peaks of the input waveform. When R_c is made large, sufficient bias will be developed so that grid current will flow only during the most positive peak of the incoming signal.

2.4j. Grid Current and Grid Power Limitations. When the grid of a triode is made positive with respect to the cathode, grid current will flow. The grid current is a function of the plate voltage and grid voltage as indicated by Eq. (2.28). When the tube is operated in the positive grid voltage region of the E_b-I_b curves, the grid-cathode resistance becomes quite low and the grid may consume considerable power from the source which is driving it. For this reason, the grid is seldom operated at positive potentials except in power amplifiers and special circuits.

The maximum current which the grid can handle is limited by the ability of the grid structure to dissipate the heat generated by the electrons which comprise the grid current striking the grid wires. In small receiving-type tubes, the grid structure is not designed to handle any appreciable amount of power and the average grid current must be kept quite low. In large transmitting tubes designed for class C oscillator and amplifier applications, the tube grid structures are quite rugged and capable of dissipating many watts of power.

2.4k. Plate Circuit Power Relations. As stated in Sec. 2.3d, the power that must be dissipated by the plate of a tube is determined by the energy per unit time released

by the electrons which strike the plate. The plate power at any instant is then given by the product of the instantaneous values of plate voltage and plate current. The plate voltage supply delivers an amount of power to the vacuum tube and its load impedance given by the product of the instantaneous values of plate supply voltage and plate supply current. The average plate supply power P_{bb} is then

$$P_{bb} = \frac{1}{T} \int_0^T e_{bb} i_b \, dt \qquad (2.44)$$

where e_{bb} = instantaneous value of plate supply voltage

i_b = instantaneous value of plate supply current

Ordinarily e_{bb} and i_b are constant, or nearly so, and the average plate supply power is the product of E_{bb} and I_b. The sum of the power P_t dissipated in the tube and the power P_l delivered to the load equals the plate supply power.

(a) TYPICAL PHYSICAL CONFIGURATION   (b) SCHEMATIC REPRESENTATION
OF TETRODE TUBE   OF A TETRODE TUBE
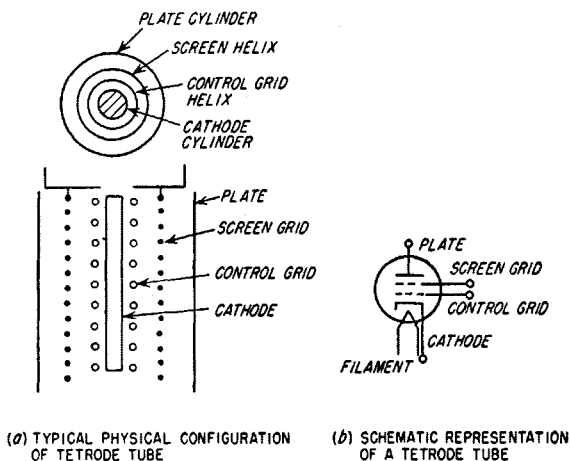
FIG. 2.32. The tetrode tube.

2.5. Tetrode Tubes. The tetrode is a four-element vacuum tube containing, in addition to the cathode, grid, and plate, a second grid known as the *screen grid*, located between the control grid and the plate. The physical construction of the screen grid is similar to that of the control grid, being a wire mesh or helix. A typical configuration of a tetrode is shown in Fig. 2.32a, and the schematic representation of a tetrode is illustrated in Fig. 2.32b.

The tetrode tube was originally developed to overcome two inherent disadvantages present in the triode tube caused by the relatively large grid-plate capacity. These effects are (1) Miller effect, which causes the input capacity of a triode to increase with the gain of the tube (see Sec. 3.8c) and (2) the necessity of neutralization of a tuned triode amplifier to prevent the feedback through the grid-plate capacitance from causing oscillation. Since the screen grid forms an electrostatic shield between the grid and plate, the capacity between these two electrodes is reduced considerably, thus minimizing Miller effect and eliminating the necessity of neutralization in tetrode tuned amplifiers at low and moderate frequencies. (See Sec. 7.4i for a discussion of the upper frequency limit of unneutralized tuned amplifiers.)

Since the screen grid forms an additional electrostatic shield between the cathode and the plate, plate voltage has very little effect upon the potential gradient at the cathode surface and therefore has little effect upon the tube current. The screen

is normally operated at a positive voltage of 0.25 to 1.0 times the plate voltage. The potential distribution within a typical tetrode is shown in Fig. 2.33. The screen intercepts a certain number of the electrons forming the tube current because of its positive potential. In normal circuit applications, the screen current will be between 0.1 and 0.3 times the plate current.

The grid-plate transconductance g_m of a triode tube is approximately proportional to the cube root of plate current. Since the introduction of a screen grid into a triode reduces the plate current by approximately 25 per cent, the g_m of a tetrode is reduced about 10 per cent from that of an equivalent triode. However, the electrostatic shielding effect of the screen grid greatly increases the plate resistance r_p because the plate voltage has much less effect upon plate current. Thus, the amplification factor μ is greatly increased over that of an equivalent triode.

FIG. 2.33. Potential distribution in a typical tetrode.

2.5a. *Tetrode Characteristic Curves.* The E_b-I_b characteristic curves for a typical tetrode are shown in Fig. 2.34. As the plate voltage is increased from zero there is a pronounced dip in the plate current for plate voltages slightly less than the screen
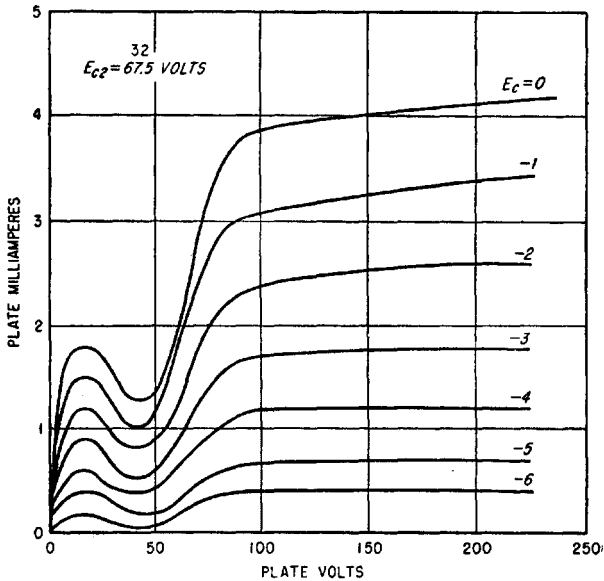
FIG. 2.34. E_b-I_b characteristic curves of a typical tetrode.

voltage. These dips in the characteristic curves are the result of secondary emission from the plate. At plate voltages above 25 volts the electron velocity is great enough to cause some secondary emission from the plate surface. As the plate potential

is increased, the rate of increase of secondary electron emission from the plate may exceed the rate of increase of primary plate electrons and, as long as the screen grid is at a higher potential than the plate, nearly all of the secondary electrons from the plate will be collected by the screen grid. Thus, there is a region of plate voltage where, for a constant control grid potential, the plate current decreases as the plate voltage is increased. The dynamic plate resistance r_p of the tube is negative in this region. The ratio of secondary to primary electrons can, under certain conditions, exceed unity (see Sec. 2.2e), causing a reversal in the direction of the net electron current to the plate unless special treatment is given the plate material to ensure a low ratio of secondary to primary electrons. As soon as the plate voltage exceeds the screen voltage, however, the number of secondary electrons which return to the plate increases rapidly, and for plate voltages only slightly higher than the screen voltage, nearly all secondary electrons are returned to the plate.

The screen grid also emits secondary electrons because of the velocity of the electrons which strike it. However, these secondary electrons contribute very little to the plate current because their number is proportionally smaller and because they are, in general, emitted on the control-grid side of the screen grid rather than on the plate side and are not directly affected by the plate potential. A plot of screen current I_{c2}
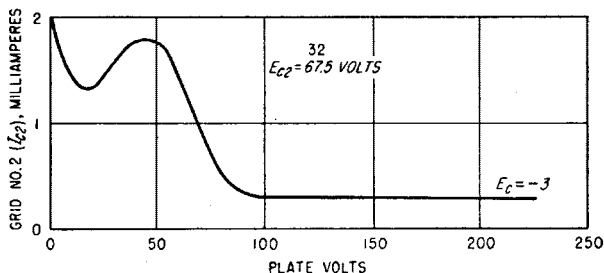
FIG. 2.35. E_b-I_{c2} characteristic curves for a typical tetrode.

as a function of plate voltage for a typical tetrode is shown in Fig. 2.35. The sum of the plate and screen currents is almost constant for a given screen-grid voltage independent of the value of plate voltage. The plate voltage determines the division of the total cathode current between screen grid and plate for a given screen-grid voltage.

2.5b. Tetrode Tube Coefficients. In addition to the amplification factor μ, transconductance g_m, and dynamic plate resistance r_p relating the electrode voltages and currents in the triode, the tetrode contains several other tube coefficients. In tetrodes and pentodes (see Sec. 2.6b) the definitions of the tube coefficients assume that all voltages and currents in the tube other than those specified are held constant. The new coefficients are:

1. Dynamic screen-grid resistance

$$r_{sg} = \frac{\partial E_{c2}}{\partial I_{c2}} \tag{2.45}$$

2. Control-grid–screen grid amplification factor

$$\mu_{c2} = -\frac{\partial E_{c2}}{\partial E_c} \tag{2.46}$$

3. Control-grid–screen grid transconductance

$$g_{mc2} = \frac{\partial I_{c2}}{\partial E_c} \tag{2.47}$$

4. Screen-grid–plate amplification factor

$$\mu' = -\frac{\partial E_b}{\partial E_{c2}} \tag{2 48}$$

5. Screen-grid–plate transconductance

$$g'_m = \frac{\partial I_b}{\partial E_{c2}} \tag{2.49}$$

The additional relationships existing among these tube coefficients in the tetrode are

$$\mu_{c2} = g_{mc2}r_{sg} \tag{2.50}$$
$$\mu' = g'_m r_p \tag{2.51}$$

The equivalent circuit of a tetrode having cathode, screen, and plate circuit impedances is somewhat more complex than that of the triode because of the effect of the control-grid voltage upon screen current and the effect of the screen voltage upon plate

(a) COMPLETE EQUIVALENT CIRCUIT OF TETRODE TUBE HAVING CATHODE, SCREEN AND PLATE CIRCUIT IMPEDANCES

(b) TETRODE EQUIVALENT PLATE CIRCUIT WITH ZERO EXTERNAL SCREEN IMPEDANCE

(c) TETRODE EQUIVALENT PLATE CIRCUIT WITH SCREEN BYPASSED TO CATHODE
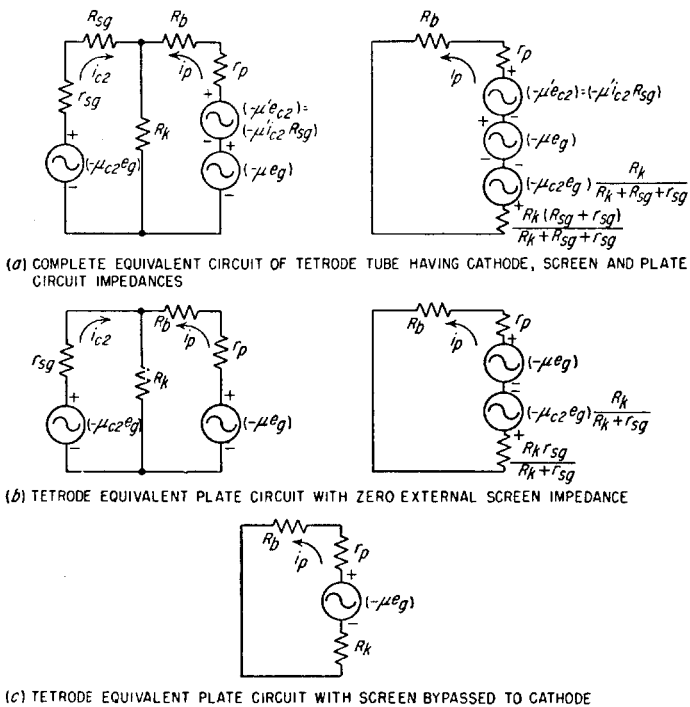
FIG. 2.36. Equivalent circuits of tetrode tubes.

current. The complete equivalent circuit is shown in Fig. 2.36. For plate voltages greater than the screen voltage, variations in plate voltage have virtually no effect upon screen current, so that the equivalent circuit shows that the screen current I_{c2} changes only as a function of the control grid voltages E_c. When the external screen circuit impedance R_{sg} is zero, the screen voltage does not vary with changes in screen current and the equivalent circuit reduces to that of Fig. 2.36b. If the screen grid is bypassed to the cathode, the equivalent plate circuit of a tetrode reduces to that shown in Fig. 2.36c, which is identical to that of a triode.

The variations in the μ, r_p, and g_m of a typical tetrode as a function of screen voltage for the condition of zero screen and cathode circuit impedances are illustrated in Fig. 2.37.

2.5c. The Beam-power Tetrode. The principal disadvantage of the tetrode tube is the very pronounced dip in the E_b-I_b characteristics caused by the secondary emission of electrons from the plate at low plate voltages. Although this disadvantage can
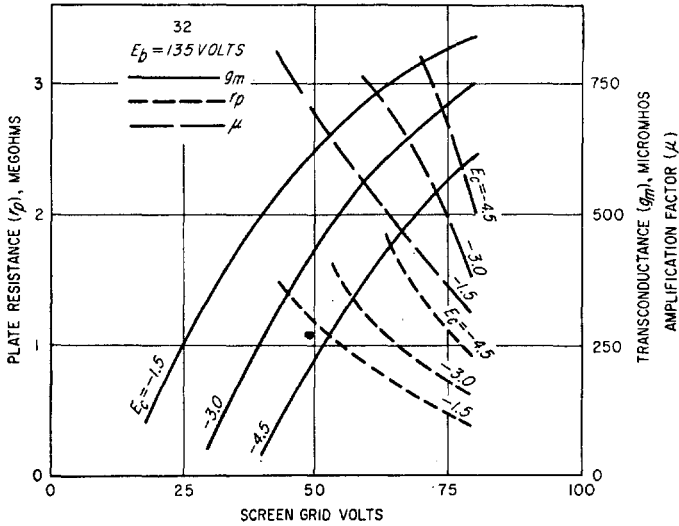
FIG. 2.37. Variations in the parameters of a typical tetrode as a function of screen voltage.

be overcome by the addition of another grid between the screen and plate (see Sec. 2.6), it can also be eliminated by special design of the tetrode tube.

The special tube design forms the electron paths from cathode to plate into high-current density beams and utilizes the effects of space charge between the screen and plate to eliminate the secondary emission effect. Tubes of this type are known as beam-power amplifiers, beam-power tetrodes, and beam-power pentodes. The internal structure of such tubes is illustrated in Fig. 2.38. The cathode surface is flat, and the screen and grid helices are wound with the same pitch. The screen wires are aligned with those of the grid so that the grid wires shield the screen from the cathode. Beam-forming plates located on each side of the cathode are internally connected to the cathode and cause the tube current to form into beams having small width and relatively high current density. The screen-plate distance is increased

FIG. 2.38. Internal structure of a typical beam-power tetrode.

over that of a conventional tetrode so that a larger percentage of the electrons comprising the tube current are in the region between the screen and the plate at any instant.

The electrons which are between the screen and plate comprise a negative space charge and create electrostatic fields at the screen plane and at the plate surface. The density of this space charge is inversely proportional to the plate voltage and directly proportional to the tube current. The effect of the large space charge thus
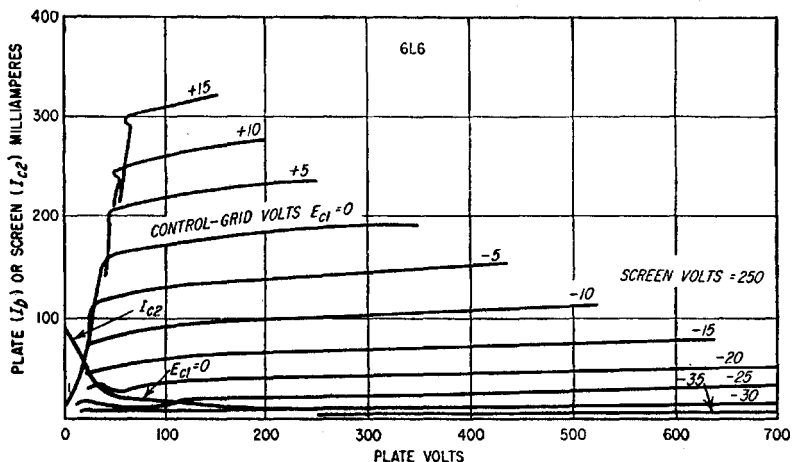
FIG. 2.39. E_b-I_b characteristic curves of a typical beam-power tetrode.
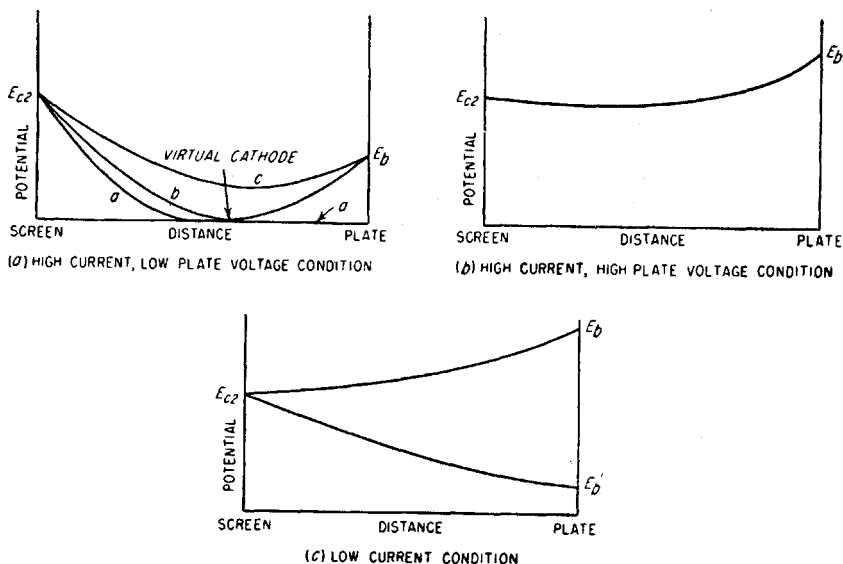
FIG. 2.40. Potential distribution between screen and plate of a typical beam-power tetrode

created is twofold; it causes the secondary electrons emitted by the plate to be returned to the plate, and it causes electrons injected into the screen-plate region to be decelerated, thus increasing the space charge. The result of these effects produces the E_b-I_b characteristic curves illustrated in Fig. 2.39. In the region of low plate voltage and a constant, high injected current, the space charge between the screen and plate is large enough to reduce the potential to zero at a plane between the screen and plate, forming a virtual cathode at this plane as shown by line a in Fig. 2.40a. Under these

conditions the potential gradient at the screen plane is such that only those injected electrons having sufficient velocity to overcome the force of the negative field will reach the plate. All other electrons will be returned toward the screen and will either be collected by the screen or pass through into the negative-space-charge region near the cathode. A relatively small increase in plate voltage produces a large increase in plate current. The plate current increases more rapidly than the three-halves power of plate voltage because the virtual cathode moves toward the plate as plate potential is increased. This condition is illustrated by the steep slope of the constant grid voltage curves at low plate voltages in Fig. 2.39.

When the plate voltage is increased to the value where the plate current equals the injected current, the increase in plate current stops abruptly. All of the injected current is going to the plate, and a further increase in plate voltage will not increase the plate current since the shielding action of the screen prevents the plate potential from changing the cathode current. At this plate potential, the virtual cathode has just disappeared, as shown by line b in Fig. 2.40a. The space charge in the region between the virtual cathode and the plate increases with plate current. When the plate current can no longer increase with plate voltage, the increasing electron velocity due to the increasing plate voltage decreases the space charge in this region. The decreased space charge causes less deceleration of the electrons between screen and plate, which in turn reduces the space charge. This regenerative condition causes the potential gradient to shift from line b to line c in Fig. 2.40a. Under this condition, the space charge is still sufficient to return all secondary electrons emitted by the plate back to the plate. Further increases in plate voltage now have almost no effect upon plate current. This is the condition existing in the region of the E_b-I_b curves of Fig. 2.39 where the plate current is nearly independent of plate voltage. The sharp knee to these curves is the result of the rapid transition between the two states described above. For higher plate voltages, the potential distribution between screen and plate will vary between that of line c in Fig. 2.40a and the curve of Fig. 2.40b as a function of the ratio of plate potential to the screen potential.

At very low plate currents, corresponding to the condition of high negative grid bias, the space charge between the screen and plate may be insufficient to return all secondary electrons to the plate, and a dip in the plate current for low values of plate voltage may occur as shown in Fig. 2.39 for plate currents less than about 35 ma. The potential distribution for this condition contains no minimum, and the tube operates as a conventional tetrode. This is illustrated in Fig. 2.40c.

Because the screen is well shielded from the cathode, the number of electrons intercepted by the screen is small, resulting in a low ratio of screen to plate current. The linearity of the E_b-I_b curves allows large plate swings with low distortion, and this together with the low ratio of screen to plate current allows the efficiency of the beam-power tetrode to be high. The beam-power tetrode has largely superseded the conventional tetrode and the triode for small- and medium-power applications.

2.6. Pentode Tubes. The disadvantages of the conventional tetrode can be overcome by the addition of another grid to the tube as well as by the beam-tetrode type of tube construction. The additional grid is placed between the screen and plate and is called a suppressor. Such tubes contain five electrodes and are, therefore, called pentodes. The suppressor is normally either connected to the cathode or grounded. The action of the suppressor is to form a region of reduced potential between the screen and plate, just as the large space charge does in the beam-power tetrode. The effect of this region of reduced potential is to return secondary emission electrons from the plate back to the plate, thus eliminating the dip in the E_b-I_b curves which is so pronounced in the conventional tetrode. The suppressor also acts as an additional electrostatic shield between the cathode and the plate, reducing the effect of plate voltage upon cathode current even more than in the tetrode. The result

is a higher dynamic plate resistance r_p and a lower grid-plate capacitance c_{gp} than in a tetrode.   However, the g_m of the pentode is nearly independent of the presence of the suppressor.

The internal structure of a typical pentode is shown in Fig. 2.41a, and the schematic representation of a pentode is given in Fig. 2.41b.

(a) STRUCTURE OF TYPICAL
PENTODE TUBE

(b) SCHEMATIC REPRESENTATION OF
A PENTODE TUBE
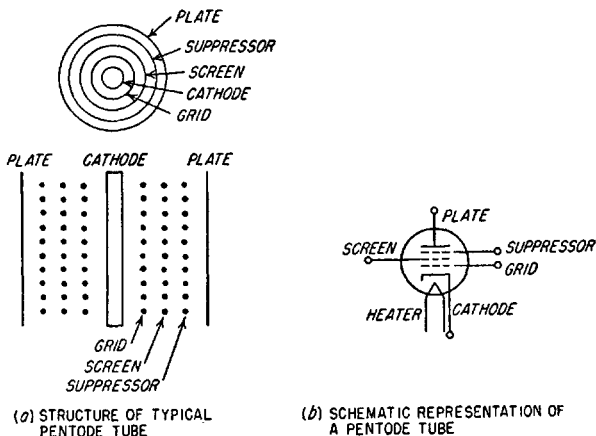
FIG. 2.41. Pentode-tube construction.
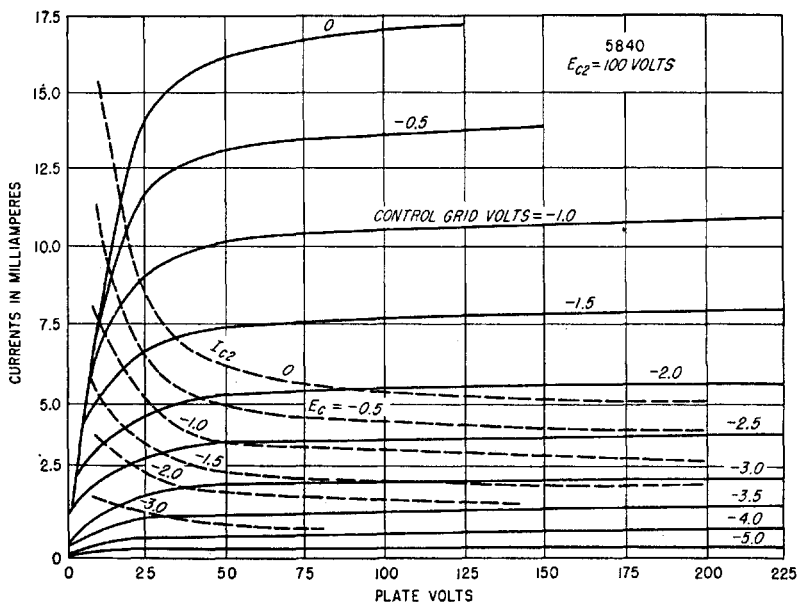
FIG. 2.42. E_b-I_b characteristic curves of a typical pentode.

2.6a. Pentode Characteristic Curves.   The E_b-I_b characteristic curves for a typical pentode are shown in Fig. 2.42.   In the region where the plate current is nearly independent of plate voltage, the potential distribution between the screen and the plate is as shown in Fig. 2.43a.   The region of reduced potential at the suppressor plane returns all secondary electrons emitted from the plate back to the plate, and

although the injected electrons entering the screen-suppressor region are decelerated somewhat by the retarding field, all electrons pass through the suppressor plane and reach the plate. As the plate voltage is reduced, the negative potential gradient between screen and suppressor becomes greater, and the average potential at the suppressor plane approaches the actual suppressor voltage. The increased negative potential gradient causes a reduction in the velocity of electrons entering the screen-suppressor region, increasing the space charge which further increases the negative
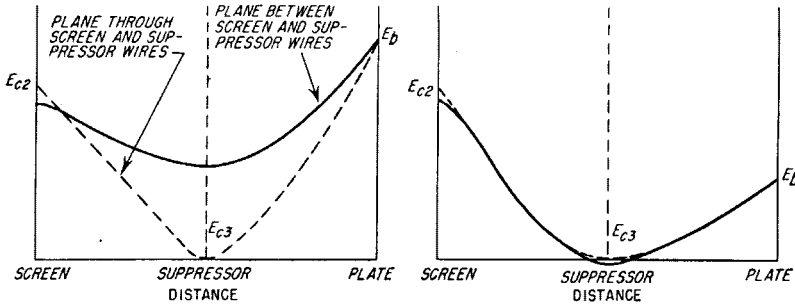
FIG. 2.43. Potential distribution between screen and plate of a typical pentode.

potential gradient between screen and suppressor. When the plate potential is reduced somewhat below that of the screen, this effect causes the formation of a virtual cathode at the suppressor plane, resulting in a potential at the suppressor plane which is slightly negative with respect to the cathode and suppressor voltage. As a result, only those electrons entering the screen-suppressor region with sufficient velocity to overcome the negative potential gradient are able to reach the plate, all others being returned to the screen or to the space charge near the cathode. Under this condition, small changes in the plate potential greatly affect the number of electrons reaching the plate by causing relatively large changes in the negative potential gradient between suppressor and screen. The potential distribution for this condition is illustrated in Fig. 2.43b. As seen in Fig. 2.42, where the screen current as a function of plate voltage is indicated in dotted lines, the total cathode current in the pentode is almost unaffected by plate potential, the plate potential merely determining the division of cathode current between screen and plate.

In addition to the E_b-I_b curves, the E_c-I_b curves and the E_c-g_m curves of the pentode tube are useful. These curves are shown in Fig. 2.44 for a typical pentode. Although the E_c-I_b curves are determined for a constant plate voltage, the very high dynamic plate resistance of the pentode makes the E_c-I_b transfer characteristic almost independent of plate voltage, thus making the E_c-I_b curves very nearly correct for all load impedances as long as operation is kept to the right of the knee of the E_b-I_b curves.
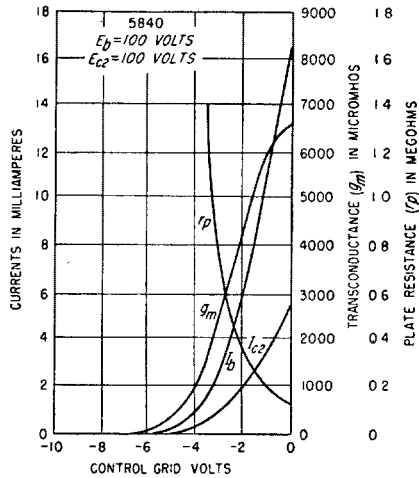
FIG. 2.44. E_c-I_b and E_c-g_m characteristic curves for a typical pentode.

2.6b. Pentode Tube Coefficients. In addition to the tube coefficients of the tetrode and the triode, the pentode tube contains coefficients relating the suppressor voltage to the plate and screen voltages and currents. Ordinarily, the pentode is operated with the suppressor either connected directly to the cathode or bypassed to it, so that the coefficients relating the suppressor voltage variations to the other electrode voltages and currents are eliminated. Under these conditions, the equivalent circuit of the pentode reduces to that of the tetrode as shown in Fig. 2.36.

In some applications the suppressor of a pentode is used as an auxiliary control electrode in addition to the control grid. In such instances, the screen is normally bypassed to the cathode and the suppressor is operated at potentials negative with respect to the cathode, so that the suppressor current is negligible. For these conditions, the equivalent circuit is as shown in Fig. 2.45.

When the suppressor voltage is such that the potential distribution between screen and plate is as shown in Fig. 2.43a, small variations in suppressor voltage produce no change in plate current since all the electrons passing through the screen have sufficient velocity to overcome the negative potential gradient existing between the screen and the suppressor. The suppressor to plate transconductance is zero for this condition. However, if the suppressor voltage is reduced until the average potential at the suppressor plane is slightly negative, only those electrons having sufficient velocity to overcome the negative potential gradient will pass through the suppressor plane and reach the plate. Small variations in suppressor voltage will now result in large variations in plate current. The action of the suppressor and the plate and the space charge is completely analogous to the action of the grid, plate, and space charge in a triode (see Sec. 2.4). The suppressor coefficients are given by

FIG. 2.45. Equivalent plate circuit of pentode having screen bypassed to cathode and suppressor acting as auxiliary control electrode.

$$g_{mc3} = \frac{\partial I_b}{\partial E_{c3}} \tag{2.52}$$

$$\mu_{c3} = -\frac{\partial E_b}{\partial E_{c3}} \tag{2.53}$$

$$g_m'' = \frac{\partial I_{c2}}{\partial E_{c3}} \tag{2.54}$$

$$\mu'' = \frac{\partial E_{c2}}{\partial E_{c3}} \tag{2.55}$$

where g_{mc3} = suppressor to plate transconductance
   μ_{c3} = suppressor to plate amplification factor
   g_m'' = suppressor to screen transconductance
   μ'' = suppressor to screen amplification factor

Since the suppressor voltage controls the division of current between screen and plate, g_{mc3} and g_m'' are related approximately by

$$g_{mc3} = -g_m'' \tag{2.56}$$

Because some of the electrons which return toward the screen after being stopped by the negative potential at the suppressor plane pass between the screen wires and add to the space charge near the cathode, the total space current is reduced, and the screen current does not increase by an amount exactly equal to the decrease in plate current.

2.6c. Variable-μ Pentodes.   In the conventional pentode, the control-grid wire spacing is uniform, and, neglecting edge effects, all portions of the control grid reach cutoff at the same grid potential.

The result is a relatively sharp cutoff of the plate current as a function of grid voltage as shown in Fig. 2.44.   However, if the spacing between the grid wires is varied, different parts of the grid plane reach cutoff at different values of grid bias so that there will be no sharply defined cutoff.   This is illustrated in Fig. 2.46.   Tubes of this type are known as variable-μ pentodes or remote-cutoff pentodes.   They find application in r-f and i-f amplifiers where it is desired to minimize the distortion of large signals with the application of AGC to the amplifier stages (see Sec. 7.4c).

A similar remote-cutoff characteristic can be obtained in a pentode by operating the screen through a large dropping resistor from the plate supply with the screen bypassed to the cathode.   As the grid bias is made more negative, the screen current is reduced, decreasing the drop across the screen resistor and increasing the screen voltage.   The increased screen voltage increases the value of negative grid bias
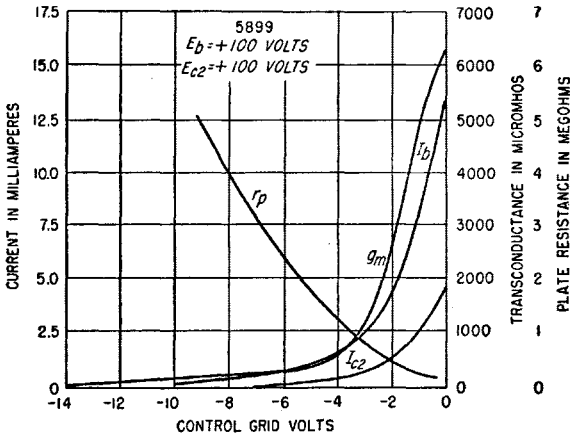
FIG. 2.46. Characteristics of a typical remote-cutoff pentode.

necessary to cut off the plate current.   The cutoff voltage is extended to a value determined by μ_{c2} and the screen supply voltage.

2.7. Cathode-ray Tubes.   A cathode-ray tube is a special type of vacuum tube in which a portion of the glass envelope is coated with a material which fluoresces when bombarded by a beam of electrons.   The tube also contains a source of electrons, a means of focusing the electrons into a beam, and a means of deflecting the electron beam so that it strikes different portions of the fluorescent screen.   Because the electron beam is virtually inertialess, the cathode-ray tube provides a means of visually displaying electrical signals which occur too rapidly to be displayed on any mechanical recorder or meter.

A typical cathode-ray tube using electrostatic deflection is shown in Fig. 2.47. It consists of a cathode, an intensity-control grid, focusing electrodes, accelerating electrode, deflection electrodes, fluorescent screen, and collector electrode.   Focusing and deflection of the electron beam can also be accomplished by the use of magnetic fields.

2.7a. The Electron Gun.   The purpose of the electron gun in a cathode-ray tube is to produce at the screen a focused beam of electrons of the proper density.   The electron gun consists of a cathode, control grid, accelerating anode, and focusing anodes or a focusing magnetic field.

The cathode of a cathode-ray tube is of the indirectly heated type and is usually made in the shape of a cylinder having a cap over one end. The heater for the cathode is placed inside the cylinder, and the emissive oxide coating is placed only on the outside surface of the end cap. The control grid is another cylinder slightly larger than the cathode cylinder having an end cap which has an aperture at its center as shown in Fig. 2.47. Opposite the aperture in the control grid is the accelerating electrode, or second grid, which also contains one or more apertures. The accelerating electrode is made positive with respect to the cathode by approximately 1,000 to 2,000 volts, and the control grid is made slightly negative with respect to the cathode.
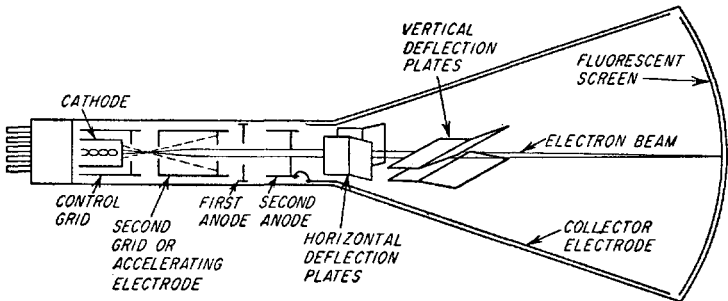
FIG. 2.47. Typical electrostatic-deflection cathode-ray tube.

The cathode, control grid, and accelerating electrode act very much like the cathode, grid, and screen of a tetrode. The potentials of these electrodes are such as to cause the formation of a space charge and virtual cathode between the cathode and control grid. The negative potential of the control grid causes all of the electrons emitted by the cathode which do not pass through the aperture in the control grid to return to the virtual cathode formed by the space charge between the cathode and the control grid. The potential of the control grid also controls the density of the electron beam passing through the control-grid aperture, thus controlling the intensity of the spot produced on the fluorescent screen by the electron beam.

2.7b. Electrostatic Focusing of Electron Beams. The focusing of the electron beam upon the screen of a cathode-ray tube is accomplished in two steps. The first step is to cause the electrons emitted from the cathode to pass through the aperture of the control grid in such a manner as to form a beam of small diameter at a plane near the control-grid aperture. The second step is to form on the screen the image of the electron beam that existed at the plane of smallest beam diameter.

FIG. 2.48. Focusing action of lens formed by cathode, control grid, and accelerating electrode.

The initial step in focusing is accomplished by the relative positions, shapes, and potentials of the cathode, control grid, and accelerating electrode. This is illustrated in Fig. 2.48. The apertures in the control grid and the accelerating electrode cause the equipotential lines of the electrostatic field between the three electrodes to have the shapes shown. The electrostatic force is always normal to the equipotential lines, thus causing the emitted electrons to be directed toward the axis of the apertures as shown in Fig. 2.48. After crossing the axis, the electrons diverge. All of

the electrons do not cross the axis at the same point.   However, in the region beyond the control grid there is a plane where the cross section of the electron beam is a minimum.   This is called the "crossover," and the image of this crossover is what is focused on the cathode-ray-tube screen.   The diameter of the electron beam at the crossover point is considerably smaller than the area of the cathode or the aperture in the control grid.

The second focusing step is accomplished by the accelerating electrode and the first and second anodes.   This focusing action is illustrated in Fig. 2.49.   The electrostatic field between the accelerating electrode and the first anode is such as to
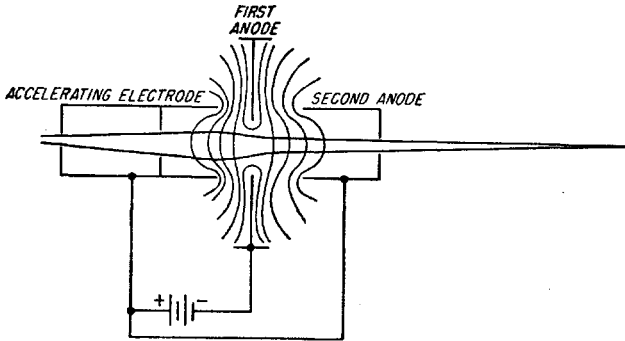
FIG. 2.49. Focusing action of accelerating electrode and first and second anode in an electrostatic-focus cathode-ray tube.

cause the electrons which are diverging after passing the crossover point to become convergent and again acquire a component of velocity toward the axis of the tube. Although the field between the first anode and the second anode exerts a force upon the electrons in such a direction as to again cause divergence, the increased spacing of the second anode relative to the first anode reduces the magnitude of the electric field existing in this region compared to the field in the region between the accelerating electrode and the first anode.   As a result, the divergence of the lens is less than the convergence of the first lens, and the image of the crossover occurs at a point past the second anode.   When this second crossover occurs at the tube screen the beam is in focus, since the spot produced on the screen will have its minimum diameter.   The size of the spot on the tube screen is determined by the diameter of the first crossover and the

FIG. 2.50. Spherical aberration in electrostatic lens.

magnification of the second lens.   Spot size can be reduced by reducing the size of the electrode apertures to remove the more divergent electrons.

Focusing is normally accomplished by varying the potential of the first anode, maintaining the potentials of the accelerating electrode and second anode constant. The adjustment of beam intensity will require refocusing, since variations in the control-grid voltage will cause some defocusing of the beam due to its effect upon the location of the first crossover.   Variation of the first anode voltage to accomplish focusing does not interact with the beam intensity because of the electrostatic shielding provided by the accelerating electrode.

Electrostatic lenses suffer from many defects, including those present in optical lenses, such as spherical aberration, astigmatism, magnification depending upon radial distance, etc.   Spherical aberration is illustrated in Fig. 2.50.   Electrons entering

the lens parallel to the axis have a focal length that is a function of the distance of the initial path from the axis. In electrostatic lenses, the focal length is always less for electrons farther from the axis, thus making it desirable to utilize electrodes having small apertures to remove the more divergent electrons.

2.7c. Magnetic Focusing of Electron Beams. The laws concerning the action of electrons in magnetic fields were discussed in Sec. 2.1*b*. Consider electrons emitted from a point source as shown in Fig. 2.51. The electrons are attracted toward the anode by the electrostatic force between cathode and anode. If a uniform magnetic field is placed parallel to the axis between cathode and anode, the initially divergent electrons will be refocused to a spot on the anode in the following manner. The magnetic field exerts a force on the electrons only when they have a velocity component normal to the direction of the magnetic flux, and the resulting force is normal to both the electron velocity and the lines of magnetic flux. Thus, the axial component of velocity of each electron is unaffected by the presence of the magnetic field, and since all electrons are subject to the same force in the axial direction, all electrons will require the same intervals of time to reach the anode, neglecting the component of initial electron velocity along the axis. However, the axial magnetic field causes each electron having an initial radial velocity component to travel a spiral
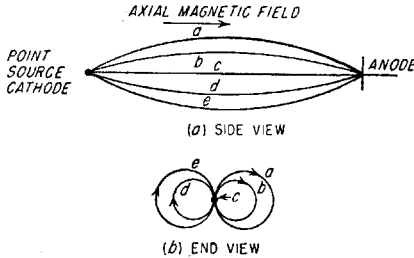
FIG. 2.51. Focusing of an electron beam by means of an axial magnetic field.
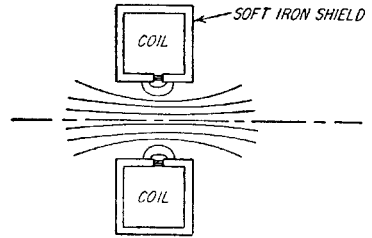
FIG. 2.52. Magnetic focusing with short axial magnetic field.

path between cathode and anode. This is illustrated in Fig. 2.51. The radius of the spiral is a function of the strength of the magnetic field and the radial electron velocity component. However, the time interval required to complete one spiral is dependent only upon the strength of the magnetic field and the ratio of mass to charge of the particle. Thus, for an electron,

$$T = 2\pi |M_e/eB| \sec \qquad (2.57)$$

where M_e = mass of electron = 9.11×10^{-31} kg
e = charge of electron = -1.6×10^{-19} coulomb
B = magnetic flux density, webers/m^2

Since the time T required to complete one spiral is the same for all electrons, and since all of the electrons were emitted from a point on the axis, all of the electrons will return to a point on the axis at time T and all multiples of T regardless of the radii of the individual spirals. If the magnetic-field strength is adjusted relative to the voltage between cathode and anode so that the time required to complete one spiral just equals the travel time between cathode and anode, all electrons will be refocused to a spot at the anode surface.

In practical cathode-ray tubes, the axial magnetic field cannot be made to extend completely between the cathode and the screen. However, the electrons from an electron gun which have been initially focused so as to form the first crossover and are then divergent can be refocused on the cathode-ray-tube screen by passing them through an axial magnetic field of shorter length as shown in Fig. 2.52. If the length

and strength of the magnetic field are adjusted properly, the divergent electrons will be caused to spiral just sufficiently when in the magnetic field to form an image of the first crossover at the cathode-ray-tube screen. The strength of the magnetic field is quite critical, since if the electrons spiral too much or too little while in the magnetic field they will not reach the spot on the cathode-ray-tube screen corresponding to the image of their location at the first crossover. The magnetic field required for magnetic focusing in a cathode-ray tube can be achieved by winding a coil around the neck of the cathode-ray tube as shown in **Fig. 2.52.** Magnetic lenses are subject to several types of distortion in addition to those which occur in electrostatic lenses.[1]

In magnetic focus tubes, variations in spot intensity caused by changes in the control-grid potential will cause defocusing just as in electrostatic focus tubes since the first crossover position will be changed. However, variations in focusing produced by changing the flux in the magnetic field will have no interaction with the beam current.

2.7d. Electrostatic Deflection Systems. The electron beam of a cathode-ray tube can be deflected from its normal position along the axis of the tube by either electrostatic or magnetic means. Electrostatic deflection is accomplished by locating two pairs of deflection plates as shown in **Fig. 2.47.** The pairs of plates are located symmetrically with respect to the axis of the tube, and one pair of plates is normal to the other. One pair of plates is called the horizontal deflection plates, and the other pair is called the vertical deflection plates.

The deflection plates are normally returned through a high resistance to approximately the same potential as the second anode. The best focus at the screen is normally obtained by varying slightly the potential difference between the horizontal and vertical deflection plates and between both pairs of plates and the second anode. The adjustment of either of these potential differences is termed "astigmatism control" and is usually used in electrostatic deflection cathode-ray tubes of 5-in. diameter or greater.

The electron beam can be deflected by the application of signal voltages to the deflection plates. The basic relationships among the parameters which determine the deflection of the electron beam when projected into an electrostatic field are discussed in Sec. 2.1. If the deflection plates are assumed to be flat parallel plates as shown in Fig. 2.53 and the electrons enter the region between the

FIG. 2.53. Electron-beam deflection in passing through an electrostatic field.

plates with a constant axial velocity and zero radial velocity, the deflection of the spot on the cathode-ray-tube screen is given by

$$d_s = \frac{lLV_d}{2dV_a} \tag{2.58}$$

where d_s = deflection on screen, cm
  l = length of plates, cm
  L = distance between center of plates and screen, cm
  d = distance between plates, cm
  V_d = potential difference between deflection plates, volts
  V_a = accelerating potential = voltage between cathode and anode preceding deflection plates, volts

[1] For greater detail, see Soller, Starr, and Valley, "Cathode Ray Tube Displays," chap. 3, Radiation Laboratory Series, McGraw-Hill Book Company, Inc., New York, 1948.

2-44

Equation (2.58) assumes no distortion of the electrostatic field between deflection plates due to edge effects and no change in axial-beam velocity after deflection. Electrostatic deflection in cathode-ray tubes is rated by either of the two following terms:

$$\text{Deflection sensitivity} = \frac{d_s}{V_d} = \frac{1}{\text{deflection factor}}$$

Common electrostatic deflection cathode-ray tubes have deflection factors in the order of 20 to 100 volts/cm.

The residual gas molecules present within a cathode-ray tube will be ionized by collisions with the electrons traveling from cathode to anode. Negative ions formed by these collisions are accelerated toward the fluorescent screen just as the electrons are. An electrostatic focus system will focus the ions in the same manner as it does the electrons even though they are much heavier than electrons. This occurs because, although the ions are heavier and consequently do not accelerate as rapidly as the electrons under the forces of an electrostatic field, they remain under the influence of the field for a longer period of time and are deflected just as much as the electrons. Accordingly, the negative ions are also deflected by an electrostatic deflection system by the same amount as the electrons. This can be seen from Eq. (2.58); the deflection of the electron is independent of its charge and mass, and therefore any negatively charged particle will be deflected by the same amount under the same conditions. The focusing of the heavy ion particles in one spot will cause a discoloration of the screen if the bombardment persists in one spot for a period of time. However, since the heavier ions are deflected by the electrostatic deflection plates, the ions will be distributed about the face of the tube as the electron beam is deflected. Under these conditions, the heavy ions will not produce any noticeable effect upon the screen.

The deflection voltage may be applied to the deflection plates either by connecting one plate to ground and applying the deflection signal to the other plate or by applying the deflection signal to both plates in push-pull. The latter method is preferable because the average potential of the deflection plates does not change in push-pull operation. In single-ended operation the average potential of the deflection plates changes as a function of the deflection voltage, causing a difference in potential between the second anode and the deflection plates. This potential variation will cause defocusing of the electron beam due to the electrostatic field established between the second anode and the deflection plates. Under this condition, best focusing can be achieved at only one point on the cathode-ray-tube screen.

The brightness of the spot on any particular cathode-ray-tube screen material is a function of both the number of electrons striking the screen per unit of time and the velocity of the impinging electrons. However, the deflection sensitivity is inversely proportional to the axial electron velocity. A high rate of deflection of the beam across the tube screen will require a high beam voltage in order to produce sufficient intensification of the trace to allow visual observation. However, a higher beam voltage will reduce the deflection sensitivity of the tube. This difficulty can be partially overcome by employing postdeflection acceleration. In tubes of this type, the potential of the second anode is maintained low enough to permit satisfactory deflection sensitivity, and an additional intensifier electrode consisting of a conducting ring around the inside of the tube is located near the screen. This intensifier electrode is operated at potentials up to approximately 2.3 times that of the second anode. In this manner, the velocity of the electrons and, therefore, the spot intensity are increased considerably without reducing the deflection sensitivity of the tube. Higher velocities may be obtained by adding additional intensifier electrodes to gradually increase the postacceleration voltage to as high as ten times that of the second anode.

This is necessary to minimize deflection distortion due to the fields existing between the intensifier electrodes, the second anode, and the deflection plates.

Since the deflection sensitivity of an electrostatic deflection cathode-ray tube is directly proportional to the distance between the deflection plates and the screen, one pair of deflection plates will normally have a slightly higher deflection sensitivity than the other pair. The pair having the highest deflection sensitivity is normally used for the vertical deflection in an oscilloscope while the other pair is used for horizontal deflection. The signals to be observed on the cathode-ray tube are applied to the vertical plates while the time-base sweep voltage is applied to the horizontal plates. It is possible to adjust the spacing and the length of the deflection plates to provide essentially equal sensitivity in both planes.

The principal advantages of electrostatic deflection are (1) high impedance deflection elements, requiring low driving power, (2) ease with which high writing speeds can be achieved, and (3) the linearity of the spot deflection with respect to the deflection voltage.

2.7e. Magnetic Deflection Systems. The deflection of the electron beam in a cathode-ray tube can also be accomplished by a magnetic field perpendicular to the axis of the tube. Since the force exerted upon the electrons is perpendicular to the direction of electron motion and to the direction of the magnetic field, deflection is produced in a plane perpendicular to the magnetic field. In the region of the magnetic field, the electrons are deflected in the arc of a circle of radius r where

Fig. 2.54. Magnetic deflection of an electron beam in a cathode-ray tube.

$$r = \left| \frac{M_e v}{Be} \right| \quad \text{meters} \tag{2.59}$$

where M_e = mass of electron = 9.11×10^{-31} kg
  v = velocity of electrons, m/sec
  B = magnetic flux density, webers/m^2
  e = charge of electron = -1.6×10^{-19} coulomb

At the termination of the magnetic field the electrons have an axial and a radial component of velocity. The deflection is proportional to the radial velocity and the time required for the electron to reach the screen. For an idealized magnetic deflection system having a uniform transverse magnetic field of fixed length and having no edge effects (see Fig. 2.54), the deflection of any charged particle is given by

$$d_s = 100lLB \sqrt{\frac{-q}{2MV_a}} \tag{2.60}$$

where d_s = deflection distance, cm
  l = axial length of magnetic field, cm
  L = distance from center of magnetic field to screen, cm
  B = magnetic flux density, webers/cm^2
  q = charge of the particle, coulombs
  M = mass of the particle, kg
  V_a = accelerating potential = voltage between anode and cathode

From Eq. (2.60) it is seen that the deflection of a charged particle by a magnetic field
is proportional to the ratio of charge to mass of the particle.   This is in contrast to
the deflection of a charged particle by an electrostatic field [see Eq. (2.58)] where the
deflection is independent of the particle charge or mass.   When the charge and mass
of the electron are substituted into Eq. (2.60) the deflection is given by

$$d = 2.96 \times 10^7 \frac{lLB}{\sqrt{V_a}} \qquad \text{cm} \qquad\qquad (2.61)$$

There are numerous configurations of deflection coils which may be used to achieve
deflection in a cathode-ray tube.   Several shapes are illustrated in Fig. 2.55.   Figure

(a) FOUR POLE IRON CORE
    DEFLECTION COILS

(b) IRON CORE PARALLEL FIELD
    DEFLECTION COILS

(c) LUMPED WINDING AIR CORE
    DEFLECTION COILS

(d) DISTRIBUTED WINDING AIR
    CORE DEFLECTION COILS

(e) SEMIDISTRIBUTED WINDING
    WITH BENT-UP ENDS
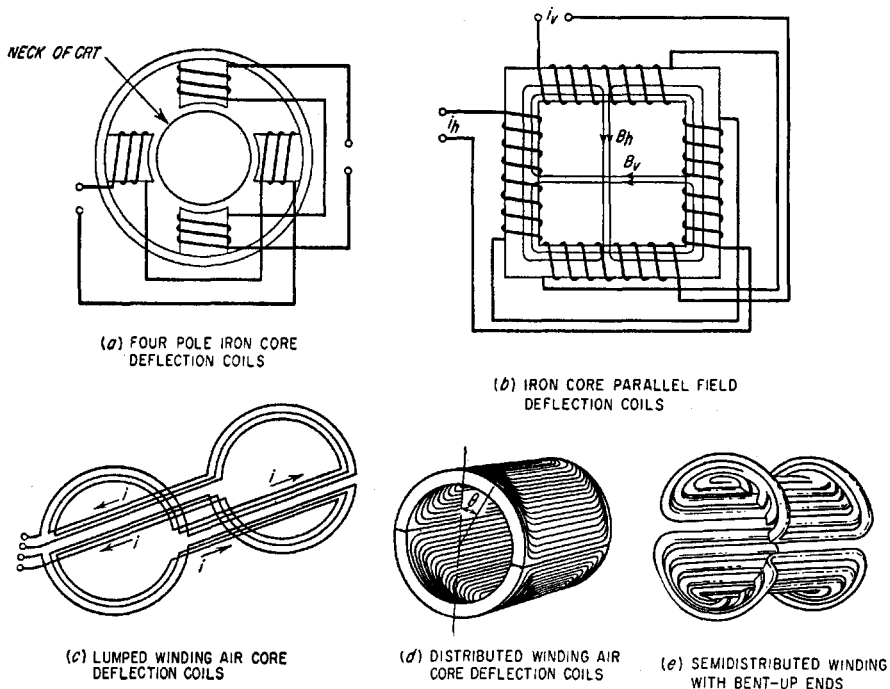
FIG. 2.55. Typical deflection coils for magnetic-deflection cathode-ray tubes.

2.55a shows a deflection system consisting of four iron pole pieces supported by a
circular yoke.   The deflection coils are wound on the pole pieces as shown.   The
yoke and pole pieces may be made of iron laminations, powdered Polyiron, or iron
ferrite.   In Fig. 2.55b the coils are wound directly on the square yoke, producing the
magnetic fields as indicated.   Because of the symmetrical construction, the two
horizontal windings induce equal and opposite voltages in each of the vertical windings
and vice versa.   Therefore there is very little interaction between the vertical and
horizontal deflection coils.   The deflection coils may be wound without an iron core
or yoke as shown in Fig. 2.55c.   The deflection coils are wound as flat loops and then
shaped to fit the contour of the cathode-ray-tube neck.   In order to provide a more
nearly uniform magnetic field across the neck of the cathode-ray tube, the lumped
loops of Fig. 2.55c may be modified as shown in Fig. 2.55d to provide a multilayer
distributed winding.   A uniform field is produced if the number of turns per degree
around the neck of the cathode-ray tube is made proportional to the sine of the angle

θ of Fig. 2.55d. A compromise between the simplicity of the lumped loops of Fig. 2.55c and the field linearity achievable in the distributed winding of Fig. 2.55d is illustrated in Fig. 2.55e. When horizontal and vertical deflection are desired, two pairs of coils are placed on the neck of the cathode-ray tube, with one pair rotated 90° with respect to the other and lying on top of the other.

There are numerous advantages to the use of magnetic deflection in cathode-ray-tube applications where its use is possible. Tubes employing magnetic deflection also usually employ magnetic focus. The combination of these two controls requires less axial space than electrostatic focus and deflection for the same diameter screen, thus allowing the neck length and, consequently, over-all tube length to be shorter. A comparison of Eqs. (2.58) and (2.60) shows that magnetic deflection is reduced by the square root of an increase in accelerating potential, while electrostatic deflection is reduced directly with any increase in accelerating potential. As the screen size is increased, the accelerating potential must be increased in order to provide the same spot intensity, since for the same angular deflection rate the electron beam ha
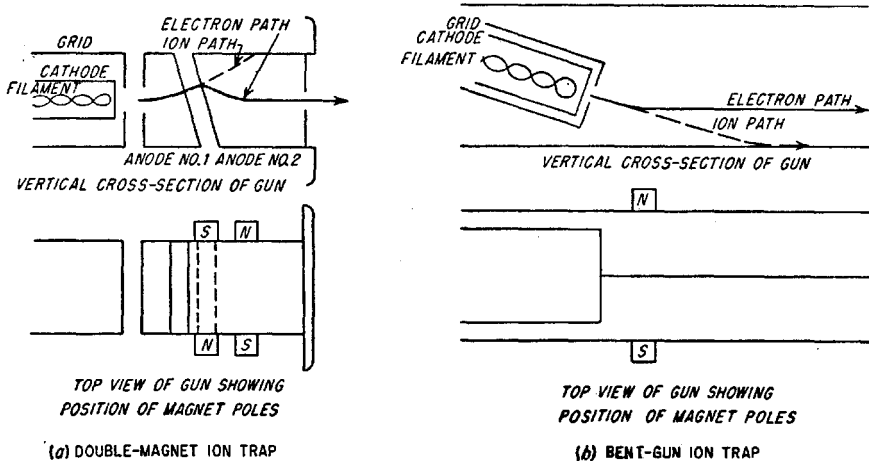
FIG. 2.56. Ion traps for electromagnetic-deflection cathode-ray tubes.

less time to spend on a given area of the screen. Magnetic deflection possesses an advantage over electrostatic deflection for large tubes because in electrostatic deflection tubes, wide angle deflection can be obtained only at the expense of deflection sensitivity. This occurs because closely spaced deflection plates will be in the beam path for wide deflection. All presently used cathode-ray tubes for television use which are larger than 10 in. in diameter utilize magnetic deflection.

A problem encountered in magnetic-deflection tubes is the prevention of an ion spot on the cathode-ray-tube screen. Comparison of Eqs. (2.58) and (2.60) shows that the electrostatic deflection of any negatively charged particle is the same regardless of its mass, while electromagnetic deflection is proportional to the square root of the ratio of charge to mass of the particle. Negative ions, which are always present within a cathode-ray tube in small quantities because of the capture of an electron by some gas molecules, have a mass which is at least 1,800 times greater than an electron but usually having the same charge. As a result, these ions are virtually unaffected by magnetic deflection, and a heavy concentration of ions will bombard the screen at the point of intersection of the electron gun axis and the screen. This ion bombardment will cause a discoloration of the screen phosphor and a reduction in the light emissivity if allowed to continue for extended periods of time. To prevent

this discoloration, the electron guns of tubes employing electromagnetic deflection contain ion traps to prevent the ions from reaching the screen. There are a number of different types of ion traps, two of which are illustrated in Fig. 2.56. All ion traps separate the ions from the electron beam by utilizing the fact that electrons are deflected readily by both electrostatic and magnetic fields while the heavier ions are deflected appreciably only by electrostatic fields.

2.7f. Cathode-ray Screens. The position of the electron beam is indicated by the emission of light from a phosphor coating on the inside surface of the tube face when bombarded by the electrons. This phenomenon is known as fluorescence. There are many different types of phosphors, having light outputs which vary in color, persistence, and efficiency. The general-purpose cathode-ray tubes used in most commercial oscilloscopes use willemite as the screen phosphor. Willemite consists of zinc orthosilicate containing traces of manganese to increase the light output. Other screen coatings utilize oxides and sulfides of zinc, cadmium, and magnesium.
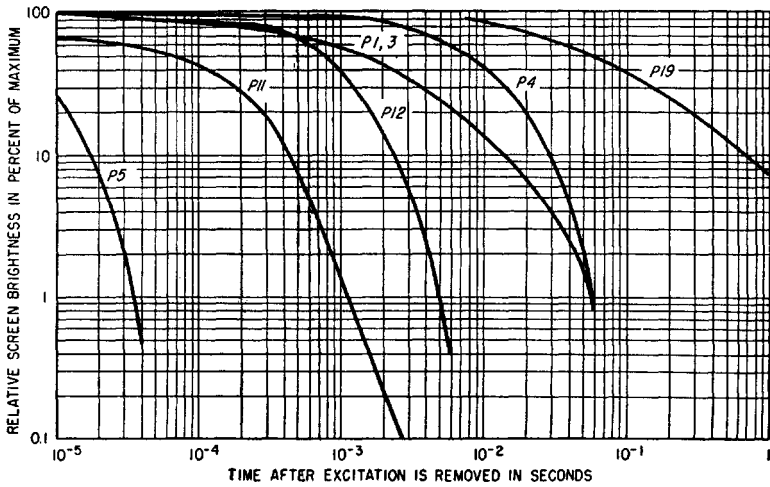
FIG. 2.57. Persistence of various cathode-ray-tube phosphors.

Certain impurities when present in extremely small proportions greatly increase the light output and affect the light color. These materials are called *activators*.

The amount of light emitted from a cathode-ray-tube-screen phosphor is given approximately by

$$CP = KI(V - V_o)\tag{2.62}$$

where CP = candlepower
   K = constant dependent upon screen material, CP/watt $\simeq 2$
   I = beam current, amperes
   V = potential of last accelerating electrode, volts
   V_o = threshold potential necessary to just excite fluorescence, volts

The light output from the screen material does not immediately reduce to zero when the electron beam is removed, but decays in an exponential manner. This phenomenon is known as phosphorescence. The persistence of the phosphorescence may vary from a few microseconds to several seconds depending upon the screen material. Commercial cathode-ray-tube screens are rated according to the spectrum of the emitted light and persistence. Persistence curves for several screen phosphors are shown in Fig. 2.57.

In order to prevent the cathode-ray-tube screen from building up a negative charge

sufficient to prevent additional electrons from reaching the screen, it is necessary for the screen to emit one or more secondary electrons for each primary electron striking the screen. As in other materials, the ratio of secondary to primary electrons is a function of the energy of the primary electrons, and the screen will function properly only over a certain range of primary electron velocities. The permissible range of electron velocities is quite wide, however, corresponding to accelerating potentials of about 200 volts as a minimum to 5,000 to 10,000 volts as a maximum, depending upon the screen material. In this region, where more than one secondary electron is emitted per primary electron, the screen potential will increase positively with respect to the collector anode potential until enough of the secondary electrons are returned to the screen to make the ratio of net secondary to primary electrons exactly unity. The secondary electrons which are not returned to the screen are attracted to the collector anode which usually consists of an aquadag coating on the inside wall of the tube near the screen.

Approximately one-half of the light emitted from the screen of a cathode-ray tube is emitted from the inner surface of the screen and serves no useful purpose. In addition, the light emitted on the inner surface of the screen tends to be reflected by the inner walls of the cathode-ray tube and then falls on the screen. This causes normally dark areas to be dark gray and reduces contrast.

Both of these difficulties and the limitations on maximum screen potential due to secondary electron emission from the screen are alleviated in certain types of cathode-ray tubes by the use of a very thin layer of aluminum on the back side of the fluorescent screen. The high-velocity electrons have sufficient energy to penetrate the aluminum layer and activate the screen. The light emitted from the inner surface of the screen is reflected by the aluminum layer, increasing the light output at the outer surface and greatly reducing reflections within the tube. The electron velocity must be quite high to allow the electron beam to penetrate the aluminum layer. A minimum accelerating potential of about 5,000 volts is required for penetration. The conducting surface formed by the aluminum layer prevents the screen from building up any charge and maintains the entire screen surface at the potential of the collector anode. This eliminates any restriction on the maximum screen potential due to secondary emission.

The aluminum layer also presents an effective barrier to the slower-moving negative ions and prevents their reaching the screen, thus eliminating the necessity of an ion trap.

2.8. Gas Tubes. Although the effects of even small quantities of gas are detrimental in conventional vacuum tubes, the characteristics exhibited by tubes filled with gas at a low pressure are very useful in special circuit applications.

2.8a. Cold-cathode Gas Diodes. A cold-cathode gas diode consists of two electrodes in a low-pressure gas atmosphere. The smaller electrode is the anode and the larger one the cathode. The gas pressure is usually between 0.001 and 0.1 mm Hg. The current-voltage characteristics of such a tube are illustrated in Fig. 2.58. As the

Fig. 2.58. Voltage-current characteristics of a cold-cathode gas diode.

anode potential is made progressively more positive with respect to the cathode, the tube current increases very slowly from an initial value of about 1 μamp until point a of Fig. 2.58 is reached. This initial current is known as the *dark current* because

under these conditions there is no visible glow in the gas. When the ignition or *break-down potential* corresponding to point *a* is reached, the ionized gas within the tube conducts heavily and, assuming a resistance in series with the tube to control the tube current, the potential from anode to cathode drops abruptly to a value determined by the type of gas in the tube and the cathode material. From point *b* to point *c* the tube potential remains very nearly constant as the tube current is increased. This is called the glow discharge region, because a portion of the tube becomes luminous, and it represents the normal operating range of cold-cathode gas diodes. It is a relatively high-voltage (75 to 150 volts) low-current (1 to 40 ma) discharge. The maximum tube current in this region is determined by the cathode area. At point *c* the tube current increases only with a substantial increase in applied voltage until point *d* is reached. This is known as the abnormal glow region. At some excessive current, such as that indicated by point *d*, the cathode surface becomes hot enough, because of the ion bombardment, to emit electrons, and the abnormal glow discharge changes to an arc discharge. The arc discharge is a low-voltage high-current discharge, and the tube voltage drops abruptly to point *e*. Beyond point *e* a larger tube current results in a slow decrease in tube voltage. When operating
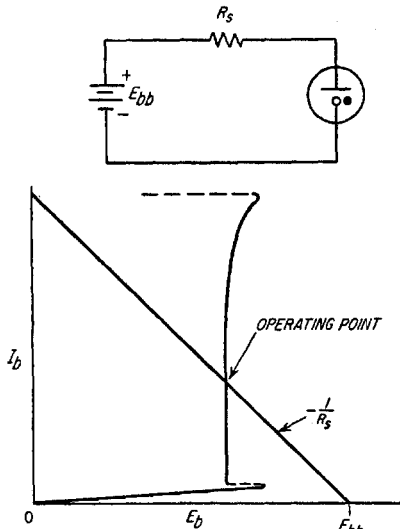
FIG. 2.59. Determination of gas diode operating point.

FIG. 2.60. Potential distribution for glow discharge between parallel-plane electrodes.

in the range from *b* to *c*, if the voltage is reduced below the value at *b*, the current abruptly drops to a value corresponding to the tube dark current. The voltage at point *b* is known as the extinction voltage.

When no voltage is applied to the diode, there are always a number of ionized gas molecules within the tube because of the radiant energy level present. As voltage is applied, the positive ions travel to the cathode and constitute the dark current. The ignition potential is reached when the electric field across the tube is sufficient to cause the cathode surface to emit secondary electrons as a result of the bombardment of the cathode by the positive ions. These electrons are accelerated toward the anode by the potential gradient in the tube, and in traveling to the anode these electrons strike other gas molecules, forming more positive ions and electrons. This process is regenerative, increasing the tube current rapidly until either the voltage drop across an external series resistance stabilizes the tube current or the region of tube operation shifts to the region *c-d* where the current will not increase unless the tube voltage is increased.[1] The operating current of a gas diode may be determined from the tube characteristics and the external circuit parameters as shown in Fig. 2.59.

[1] For greater detail concerning the mechanism of breakdown in gas diodes, see D. V. Geppert, "Basic Electron Tubes," pp. 231–246, McGraw-Hill Book Company, Inc., New York, 1951.

The potential distribution within the diode is as shown in Fig. 2.60. Almost the entire tube voltage drop occurs within a short distance of the cathode. The kinetic energy of the positive ions after accelerating through this voltage drop is sufficient to cause secondary emission of electrons from the cathode surface. In the region of nearly constant potential the numbers of positive ions and electrons are approximately equal. The near equalization of the space charge in this region results in a very low potential gradient and, consequently, low velocity of the positive ions. Positive ions are created in this region by the collision of electrons which have been accelerated by the high potential gradient near the cathode with the gas molecules. This region is luminous and is known as the *plasma*. The high potential near the cathode is caused by a high concentration of positive ions at the edge of the plasma. The region near the cathode surface which is being bombarded by the positive ions is also luminous. This is referred to as cathode glow. As the tube current is increased from point *b* in Fig. 2.58, the area covered by the cathode glow increases until at point *c* it covers the entire cathode surface. The current density in the area covered by the cathode glow remains constant in this region of operation. The current density is dependent upon the gas, cathode material, and gas pressure. Beyond point *c* further increase in tube current requires an increase in the current density at the cathode surface and requires increased anode potential.

The principal application of cold-cathode gas diodes is as voltage regulators and voltage reference tubes. Since the voltage across the tube is nearly independent of tube current in the region of normal operation, the tube can be used to stabilize the voltage output of a varying voltage source for both supply-voltage variations and load-current variations. Design information on this application of cold-cathode gas diodes is given in Sec. 15.7. A list of commonly used cold-cathode gas diode tubes is given in Table 2.4.

2.8b. Cold-cathode Triodes. The breakdown or ignition of a gas tube can be controlled by a grid placed between the cathode and anode. This is possible because the ignition potential of a gas tube is an inverse function of the spacing between the two electrodes. Closely spaced electrodes require a high potential before breakdown can occur because the short path length for the few free electrons present in the gas before ignition takes place makes the probability of these electrons colliding with gas molecules and ionizing them very small.

TABLE 2.4. CHARACTERISTICS OF SEVERAL COLD-CATHODE
GAS DIODE VOLTAGE REGULATOR TUBES

Tube type	D-C operating voltage, volts	Operating current, ma		Breakdown potential, volts	Voltage regulation within operating-current range
		Min.	Max.		
OA2	150	5	30	185	2 volts
OA3/VR75	75	5	40	105	5 volts
OB2	108	5	30	133	2 volts
OC3/VR105	105	5	40	133	2 volts
OD3/VR150	150	5	40	185	4 volts
CK5783WA	86	1.5	3.5	125	0.1 volt
CK5787WA	100	1	25	141	{1–5 ma 1 volt {5–25 ma 4 volts
5644	95	5	25	130	5 volts
5651	87	1.5	3.5	115	3 volts

In order for ignition to take place, the probability of an electron ionizing a gas molecule must equal or exceed the ratio of positive ions striking the cathode to secondary emission electrons from the cathode. If this condition is not met, breakdown will not occur. Therefore, by locating a grid close to the anode, grid-anode breakdown cannot occur until the grid-anode potential is made much higher than the grid-cathode potential required for breakdown. The grid electrostatically shields the cathode from the anode and prevents anode-cathode breakdown when the anode-cathode potential exceeds the breakdown voltage. Breakdown can occur only when the grid-cathode potential is made large enough to allow the gas ions to initiate secondary electron emission from the cathode, i.e., when the grid-cathode breakdown potential is reached. When ignition does take place between cathode and grid, the major portion of the electron current shifts from the grid to the anode if the anode is at a sufficiently high potential with respect to the grid. Resistances in the grid and anode or cathode leads are necessary to limit the current flow to these electrodes.

After ignition takes place the grid no longer has any control. If the grid is made negative with respect to the cathode, positive ions are attracted to the grid in sufficient numbers to just neutralize the negative field caused by the negative grid voltage. As a result, the gas discharge is completely shielded from the grid. After breakdown occurs, the grid potential only affects the magnitude and polarity of the grid current. Deionization of the tube can occur only when the grid and plate voltages are reduced below the extinction value for the tube. The grid can regain control only after the tube has been deionized.

The operating characteristics of a typical cold-cathode gas triode are shown in Fig. 2.61.

2.8c. Hot-cathode Gas Diodes. In cold-cathode gas tubes the ignition potential is determined by the velocity required by the positive ions to cause secondary electron emission from the cathode. If the cathode is heated to a temperature which will cause sufficient thermionic emission of electrons, the gas in the tube will ionize when the anode-cathode potential becomes high enough to give the electrons sufficient energy to ionize the gas molecules by collision. This potential is the *ionizing potential* of the gas. It is about 16 volts for argon gas.

When the ionization potential is reached in a hot-cathode gas tube, the tube current increases immediately to the full emission current of the cathode unless limited by external circuit resistance. Any further increase in tube current can occur only by raising the tube voltage sufficiently to allow positive-ion bombardment of the cathode to increase the erection emission of the cathode. However, an oxide-coated cathode will be disintegrated if the energy level of the positive ions is sufficiently high. The voltage drop corresponding to this critical energy level is called the *disintegration voltage*. It is between 20 and 25 volts for the inert gases and mercury vapor. Therefore it is essential that the external circuit resistance be high enough to limit the tube voltage to a value below the disintegration voltage of the cathode material.

The low-voltage high-current discharge that takes place in a hot-cathode gas tube is termed an *arc discharge*. The characteristics of a typical diode are given in Fig. 2.62a. The potential gradient in a hot-cathode gas diode is shown in Fig. 2.62b. In normal operation the plate current is less than the emission current capabilities of the cathode at its operating temperature, and the negative space charge resulting from this condition causes a virtual cathode to be formed near the cathode surface. A short distance from the virtual cathode the potential within the tube rises to a value nearly equal to the total plate-cathode potential. The emitted electrons which travel to the plate are accelerated sufficiently by this potential to ionize the gas molecules upon collision. In the plasma region between this virtual cathode and the anode, electrons and positive ions are present in nearly equal numbers, and nearly complete space-charge neutralization occurs, thus maintaining a low potential drop
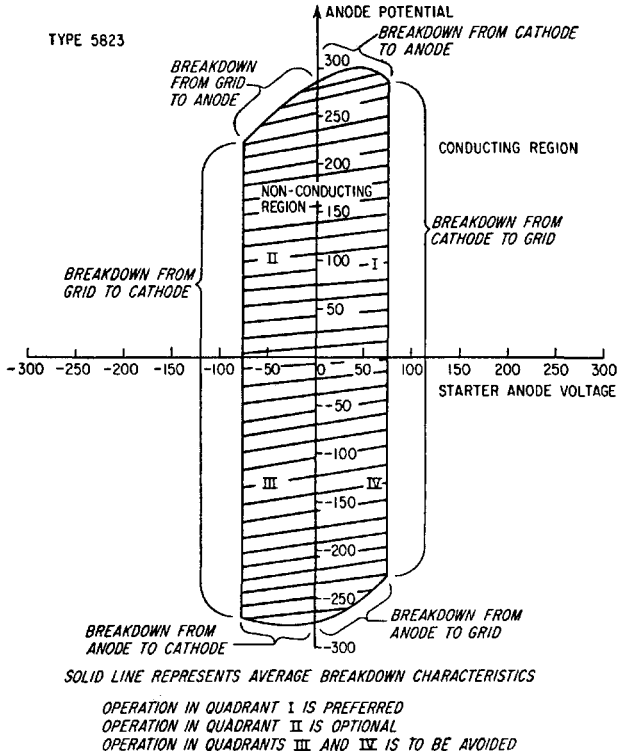
FIG. 2.61. Ignition characteristics of a typical cold-cathode gas triode.

(a) VOLTAGE-CURRENT CHARACTERISTICS

(b) POTENTIAL DISTRIBUTION WHEN CATHODE EMISSION EXCEEDS ANODE CURRENT
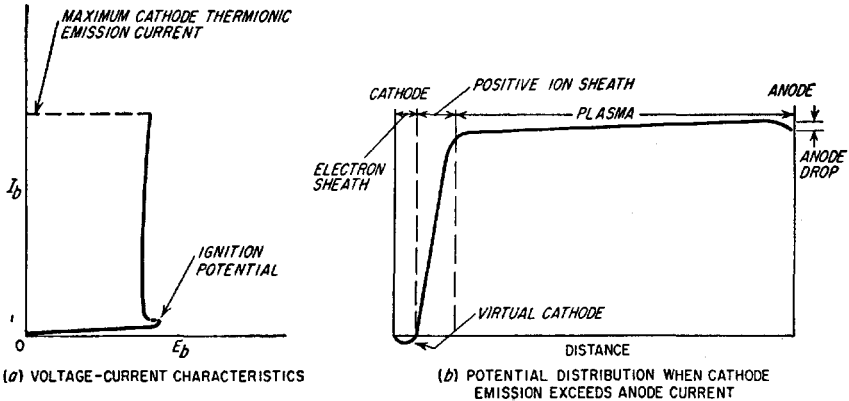
FIG. 2.62. Characteristics of a typical hot-cathode gas diode.

across this region. The low voltage drop of hot-cathodes gas tubes permits the use of anodes of much smaller size than in vacuum tubes for the same current ratings because of the reduced anode dissipation.

In hot-cathode gas tubes it is possible to obtain considerably higher emission efficiencies than in vacuum tubes. In a high-vacuum diode, the electrons leaving the

cathode are accelerated toward the anode by the electrostatic field between the two electrodes, and if a portion of the cathode is shielded from the anode, the resultant negative space charge in that region prevents electrons in that area from reaching the anode. In a gas diode, however, the plasma extends from the anode to within a very short distance of the cathode, and electrons emitted from the cathode will migrate through the region of the plasma to the anode regardless of the cathode shape. Because of the mobility of the positive ions and electrons in the plasma region, the cathode structure can be quite complex in shape and can, therefore, be designed for high emission efficiency by reducing the heat loss. Heat loss is minimized by designing the cathode structure to have a high ratio of emitting surface to volume and by using heat shields to minimize radiation. Two common cathode designs are illustrated in Fig. 2.63. High emission efficiency results in increased heating time, since the achievement of a high heat capacity with lower heating energy increases the time necessary to reach thermal equilibrium. In low-power, hot-cathode tubes the heating time for indirectly heated cathodes is usually between 1 and 5 min. Voltage must not be applied to a hot-cathode gas tube until the cathode has reached the
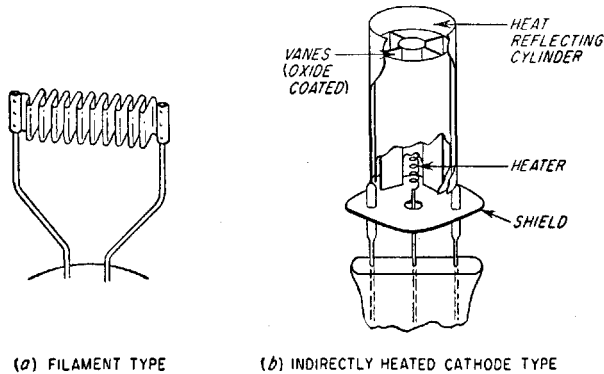
(a) FILAMENT TYPE   (b) INDIRECTLY HEATED CATHODE TYPE

FIG. 2.63. Typical cathode structures for hot-cathode gas tubes.

proper temperature to supply the necessary current since the voltage source will ordinarily be high enough to allow the tube potential to exceed the cathode disintegration potential.

Any of the inert gases or mercury vapor may be used in hot-cathode gas tubes. Mercury vapor has the disadvantage of being very sensitive to changes in gas pressure due to temperature. If the mercury-vapor pressure is too low, there may be insufficient gas ionization, resulting in a high voltage drop across the tube and damage to the cathode. If the mercury-vapor pressure is too high, the inverse voltage at which gas ignition occurs becomes quite low. The inert gases have the disadvantages of a higher ionization potential, resulting in a higher tube drop, and a lower peak inverse voltage rating due to the higher gas pressures necessary to minimize the effects of gas absorption in the tube. The peak inverse voltage rating on a hot-cathode gas tube is established by the reverse polarity voltage necessary to cause secondary electron emission from the anode by positive ion bombardment, resulting in a glow discharge.

Hot-cathode gas diodes are principally used as rectifiers. For high-current power supplies, hot-cathode gas diodes have the advantages over high-vacuum diodes of lower voltage drop, higher efficiency, lower filament power, and lower cost. They have the disadvantages of lower peak inverse voltage rating, production of r-f transients, and the requirement of a relatively long warmup period.

2.8d. Hot-cathode Gas Triodes (Thyratrons). A control grid can be added to a hot-cathode gas diode just as in the cold-cathode gas tubes. Hot-cathode gas triodes are known as *thyratrons*.

The mechanism of grid control in the hot-cathode triode is similar to that of the cold-cathode tubes. The grid is placed between the cathode and anode in such a manner as to electrostatically shield the cathode from the anode. With complete shielding by the grid, the electrons emitted by the cathode are not accelerated to the anode, and a negative space charge is formed near the cathode. When the grid potential is raised to a value which will allow electrons to pass into the grid-anode region, the electrons are accelerated by the grid to anode electric field and ionize the gas molecules by collision. The positive ions thus created will travel to the grid if it is negative, covering it with a sheath of ions which neutralizes the electrostatic field
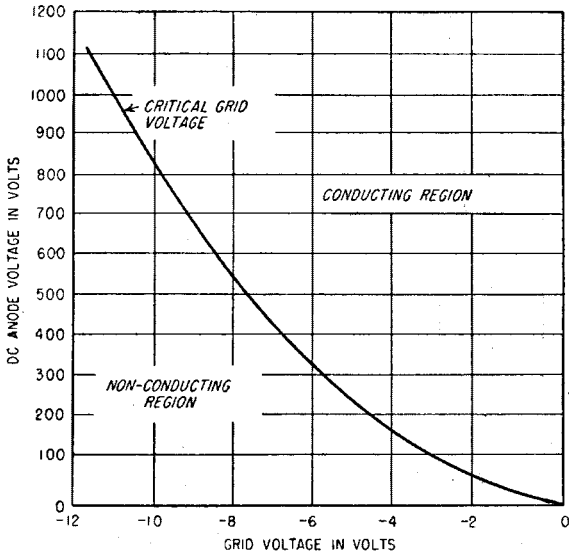
FIG. 2.64. Grid control characteristics of a typical thyratron.

of the grid and allows the anode field to accelerate more electrons. The process is regenerative until the current is limited by cathode emission or by external circuit resistance. The grid loses all control once it reaches the *critical grid potential* which allows enough electrons to enter the grid-anode region to initiate complete ionization. Grid control can be regained only by reducing the grid and anode potentials below the ionization potential of the gas. The critical grid voltage depends upon the electrode structure, anode voltage, and gas pressure. The performance of a thyratron is described by its grid-control characteristic. This is a plot of anode voltage versus critical grid voltage above which the tube ionizes. The grid-control characteristic of a typical tube is shown in Fig. 2.64. The major portion of the grid-control characteristic curve is nearly linear, and its slope $\Delta E_b/\Delta E_c$ is called the thyratron *control ratio*.

The grid structure of a thyratron is quite different from that of a high-vacuum tube. The grid shields the cathode from the anode very completely and usually contains one or more holes of much larger size than a high-vacuum-tube grid mesh. Typical electrode arrangements are illustrated in Fig. 2.65. By using several grid baffles it is possible to require positive grid voltage for ionization to occur.

Grid current may flow in a thyratron just prior to and during conduction. In any application this current must be taken into account in the design of the external grid circuit. The principal causes of grid current are electrons attracted to the grid from the cathode when the grid is positive or intercepted by the grid when it is just slightly negative with respect to the cathode (Edison effect); positive ions attracted to the grid when the grid is negative; electrons emitted from the grid because of contamination of the grid.

One limitation in the use of thyratrons is the time required for gas deionization after the anode and grid voltages have been reduced below the ionization potential. The deionization before the anode voltage is reapplied must be complete enough so that the remaining ions are insufficient to shield the grid and cause it to lose control. The necessary deionization time increases with anode voltage, increased spacing between electrodes and between electrodes and walls, increased current previous to extinction, and increased gas pressure. Deionization time is reduced by negative grid and anode voltages after conduction and low resistances in grid and anode circuits. The deionization time is also a function of the type of gas used in the tube. Mercury vapor, argon, and hydrogen are commonly used. Because of its high positive ion

(*a*) SINGLE GRID BAFFLE   (*b*) MULTIPLE GRID BAFFLE
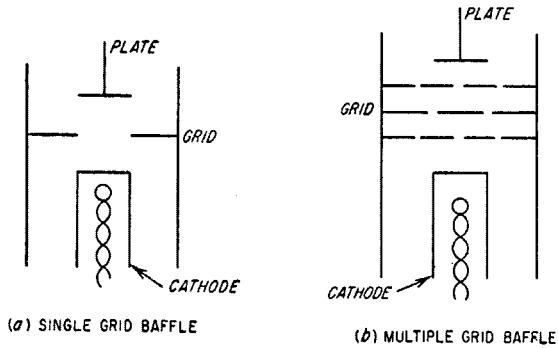
FIG. 2.65. Typical electrode arrangements in thyratrons.

mobility due to low ion mass, hydrogen ionizes and deionizes most rapidly. Representative values of deionization times are shown in Table 2.5.

Thyratrons are used for rectification, power control, power inversion (direct current to alternating current), and as switches. Hydrogen thyratrons are widely used in pulse modulators for radar transmitters. The characteristics of some commonly used thyratrons are given in Table 2.5.

2.8e. Voltage and Current Ratings of Hot-cathode Gas Diodes and Triodes. The various voltage and current ratings applied to hot-cathode tubes are listed as follows:

1. *Peak forward voltage.* This is the maximum voltage that can be applied to the anode of a thyratron without danger of a glow discharge forming between grid and anode, causing loss of grid control.

2. *Peak inverse voltage.* This is the maximum negative voltage that can be applied to the anode without danger of a glow discharge between the anode and the grid or cathode, allowing current flow in the reverse direction.

3. *Maximum peak cathode current.* This is the highest current that the tube can carry for short periods of time without damage to the cathode because of positive-ion bombardment.

4. *Maximum average anode current.* This is the highest current the tube can carry continuously without overheating of the anode.

5. *Maximum peak grid current.* This is the highest current that the grid can carry for short periods of time.

TABLE 2.5. CHARACTERISTICS OF SEVERAL THYRATRONS

Tube	Type	Gas	Anode voltage drop, volts	Maximum peak anode voltage, volts	Maximum peak cathode current, amp	Maximum average cathode current, amp	Ionization time, μsec	Deionization, μsec (approx.)
2D21	Tetrode	Inert gas	8	650	10	0.1	0.5	35
6D4	Triode	Inert gas	18	350	0.11	0.025		
2050	Tetrode	Inert gas	8	650	10	0.1	0.5	50
3C23	Triode	Inert gas and mercury vapor	15	1,250	120	1.0	10	1,000
3D22	Tetrode	Inert gas	10	650	30	0.8	0.5	150
105	Tetrode	Mercury vapor	16	10,000	160	8.0	10	1,000
3C45	Triode	Hydrogen	150	3,000	35	0.045	0.6	25
4C35	Triode	Hydrogen	1,500	8,000	90	0.1	0.6	
5C22	Triode	Hydrogen	2,500	16,000	325	0.2	1.0	
1907/5949	Triode	Hydrogen	5,000	25,000	500	0.5	1.0	
1257	Triode	Hydrogen	.....	38,000	2,000	2.5		
1258	Triode	Hydrogen	100	1,000	20	0.05	0.6	

6. *Maximum average grid current.* This is the highest current that the grid can carry continuously.

7. *Integration time.* The instantaneous anode current averaged over the integration time must not exceed the average current rating.

2.9. Transistors. A transistor is a semiconductor amplifier. By properly preparing and processing semiconductor materials such as alloyed germanium or silicon and attaching three or four electrodes in the correct manner, it is possible to obtain current amplification and/or considerable power gain. Such a device can thus perform functions analogous to vacuum tubes in various circuit applications.

Transistors are relatively small, require no filament power, have practically unlimited life when used properly, and can be made quite rugged mechanically.

2.9a. Properties of Semiconductors. Transistor operation can be better understood when some of the more important characteristics of semiconductor materials are known.[1]

Atomic Structure. In atomic theory, the atoms of each element consist of a nucleus having a specific mass and electric charge and one or more electrons revolving in orbits about the nucleus. The innermost orbit contains a maximum of two electrons, the next two orbits each contain a maximum of eight electrons, and the fourth and fifth orbits contain a maximum of ten and eight electrons, respectively. The electrons in the outer orbit are known as valence electrons, and only valence electrons enter into combination with other atoms. Those elements which have four valence

[1] For a comprehensive treatment of this subject see W. Shockley, "Electrons and Holes in Semiconductors," D. Van Nostrand Company, Inc., New York, 1950.

electrons and which require four additional electrons to complete their outer orbit are of specific interest because of their semiconductor properties. Two of these elements, silicon (14 electrons) and germanium (32 electrons), are currently being used in diodes and transistors.

In atomic theory, the electrical properties of the elements are explained by the concept of energy bands. The valence electrons are believed to have certain distinct energy levels or bands, and the conductivity of the element is determined by the energy required to elevate its valence electrons from their normal energy band to the conduction band, which has the highest energy level. Typical energy-band diagrams are shown in Fig. 2.66. In the case of an insulator, the energy gap separating the valence band and the conduction band is extremely large, and it is very difficult for a valence electron to reach the conduction band. In the case of a conductor, the valence band and the conduction band overlap, and the valence electrons are available for conduction. In a semiconductor, the energy gap separating the valence band and the conduction band is very small, and the thermal energy of the valence electrons at room temperature is sufficient to allow appreciable conduction. The energy gap for silicon is 1.11 electron-volts, and the energy gap for germanium is 0.72 electron-volt.

Crystalline Structure. The crystals of germanium and silicon are tetrahedral in shape, with each of the four valence electrons of each atom forming a *covalent bond*
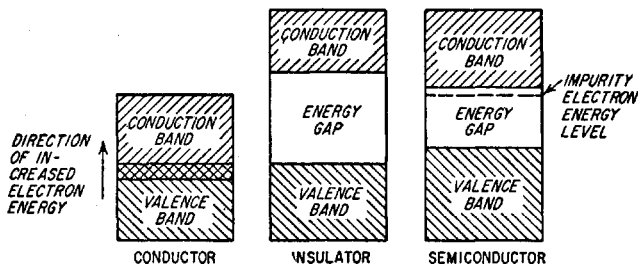
FIG. 2.66. Typical energy-band diagrams.

with one of the valence electrons of four adjacent atoms. In this form, crystalline germanium or silicon is in equilibrium, has no net positive or negative charge, and at low temperatures is an insulator.

If sufficient energy, such as thermal energy, is present, some of the valence electrons will be excited into the conduction band and will be freed from their covalent bonds. The removal of the electron from the atom leaves the atom with a net positive charge. This absence of an electron in the crystal is referred to as a *hole*, and the excitation of an electron into the conduction band results in the creation of a *hole-electron pair*. The application of an electric field will cause the electron and hole to drift in opposite directions in response to the field force. Motion of a hole is accomplished by the movement of an electron from an adjacent atom into the vacancy in the first atom, creating a hole in the second atom. The repetition of this process constitutes movement of the hole in response to an applied electric field. The density n_i of thermally generated hole-electron pairs increases exponentially with temperature. For germanium at room temperature (300°K), $n_i = 1.7 \times 10^{13}$ pairs/cm^3.

Impurity Conduction. The conductivity of germanium and silicon is greatly increased by certain impurities, and it is possible to carefully control the conductivity of germanium or silicon in this manner. Although the same principles apply to both silicon and germanium, only germanium is considered here. When a controlled amount of arsenic or antimony, both of which have five valence electrons, is added to crystalline germanium, its atoms unite in covalent bonds with the germanium atoms.

But since only four of the valence electrons can form covalent bonds with the surrounding germanium atoms, the fifth valence electron is left free. The energy required to make this electron available for conduction is approximately 0.01 electron-volt for germanium (0.05 electron-volt for silicon) at low concentration levels, and at room temperature the thermal energy of the electron is sufficient to elevate it to the conduction band. The absence of the extra valence electron from the arsenic or antimony atom leaves it with a net positive charge of 1 electron unit. This positive charge establishes an electric field about the atom, resulting in a distortion of the covalent bonds surrounding it in such a manner as to reduce the field. At room temperatures the thermal energy of the excess valence electrons is sufficiently high that the reduced attractive force of the arsenic atom is insufficient to retain the fifth valence electron. Impurity elements which add free electrons are referred to as *donors*, and germanium containing donors is known as *N-type germanium*, since current flow is accomplished by the motion of negatively charged electrons.

If controlled amounts of gallium or indium, which have only three valence electrons, are added to crystalline germanium, the valence electrons of these atoms will also join in the covalent bonds with the valence electrons of the surrounding germanium atoms. In this case, however, one of the four covalent bonds joining the gallium or indium atom with the four surrounding germanium atoms is incomplete. A valence electron from a germanium atom somewhere else in the crystal will fill in the incomplete bond, leaving a hole elsewhere in the crystal and giving the gallium or indium atom a net negative charge of e. The covalent bonds with adjacent electrons become distorted in such a manner as to reduce the resulting electric field, and at room temperature the field is of insufficient strength to attract the holes formed in the germanium. Impurity elements which result in a deficiency of electrons in the crystal are known as *acceptors*, and germanium containing these elements is known as *P-type germanium*, since current flow is effectively the result of the motion of the positively charged holes. The concentration of impurities in both N-type and P-type germanium is extremely low, about 10^{15} atoms per cubic centimeter.

The PN Junction. When P-type germanium and N-type germanium are joined together, a PN junction is formed. Such a junction exhibits the property of rectification; i.e., the current flow resulting from a voltage applied in one direction across the junction is different than the current flow which results from the same voltage applied in the opposite direction across the junction.

The P-type germanium contains holes having positive charges which are free to move about and acceptor atoms which have negative charges and are fixed in place. The N-type germanium contains electrons which are free to move about and donor atoms which have positive charges and are fixed in place. Since the number of holes in the P-type germanium equals the number of acceptors and the number of electrons in the N-type germanium equals the number of donors, neither crystal has any initial charge. However, when they are placed together, thermal energy will cause some electrons to diffuse into the P region and some holes to diffuse into the N region. This diffusion causes the P region to attain a slight negative charge and the N region to attain a slight positive charge. This is illustrated in Fig. 2.67. As a result, a potential gradient is developed across the junction which causes the electrons in the N region to tend to stay in the N region and move away from the junction and causes the holes in the P region to tend to stay in the P region and move away from the junction. The surplus charge in each region thus becomes concentrated near the junction. The thermal energies of the holes and electrons are sufficient to cause a certain percentage of the holes in the P region to overcome the potential gradient at the junction and move into the N region, and a corresponding percentage of electrons in the N-region to move into the P-region. These electrons and holes diffuse about in their new regions for some average time called the *lifetime* and then combine with an

opposite charge in that region. This net current flow is just balanced, however, by the rate at which new electrons are created in the P region and new holes are created in the N region by the thermal generation of hole-electron pairs in both regions and

P-REGION   N-REGION

(*a*) CHARGE DISTRIBUTION BEFORE JUNCTION IS MADE   (*b*) CHARGE DISTRIBUTION AFTER JUNCTION IS MADE

(*c*) ELECTROSTATIC POTENTIAL AT JUNCTION   (*d*) CHARGE DENSITY DISTRIBUTION AT JUNCTION
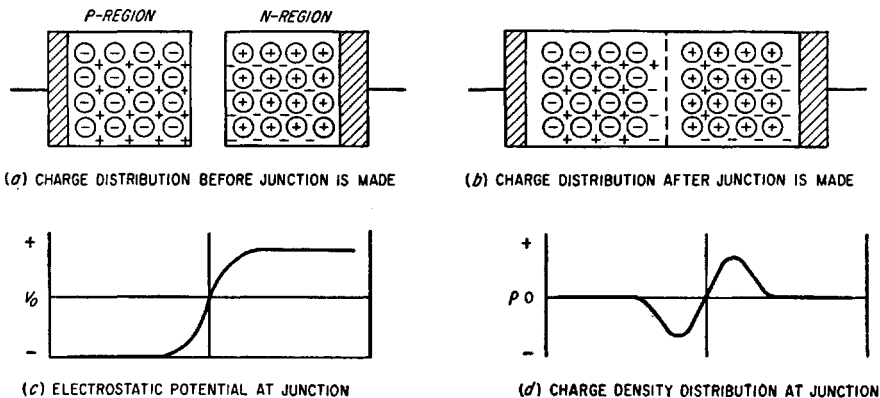
FIG. 2.67. The distribution of charges in a P-N junction under thermal equilibrium.

are swept back into their respective N-region and P-region as a result of the combined effects of thermal energy and the potential gradient across the junction. At any temperature, the generation and the recombination of hole-electron pairs are in equilibrium, resulting in a specific concentration of holes and electrons in the materials determined by thermal energy. As a result, there is no net current flow across the junction.

(*a*) ZERO APPLIED POTENTIAL

(*b*) REVERSE BIAS

(*c*) FORWARD BIAS
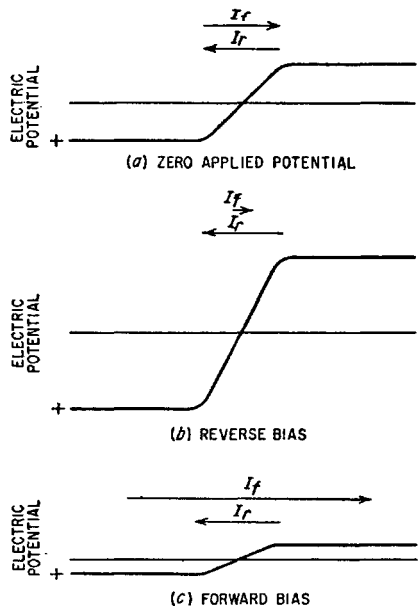
FIG. 2.68. Junction potential variation with forward and reverse bias.

When a potential is applied across the junction which aids the potential gradient already existing across the junction (the N-type germanium made positive with respect to the P-type germanium), the flow of forward current I_f due to electrons in the N region passing to the P region and holes in the P region passing into the N region is proportionately reduced, while the flow of reverse current I_r due to newly created electrons in the P region and newly created holes in the N region diffusing across the junction remains essentially constant. An applied potential of this polarity is referred to as a *reverse potential*. The net current flow approaches a constant value as the forward current I_f is reduced to zero while the reverse current I_r remains constant.

If a *forward potential* is applied across the junction (the P-type germanium made positive with respect to the N-type germanium), the number of electrons moving from the N region to the P region and the number of holes moving from the P region to the N region increase greatly, while the reverse current I_r remains nearly constant

as before. The net current $I_f - I_r$ increases exponentially as the forward potential is increased. The variations in the electric potential across the junction with forward and reverse biases are illustrated in Fig. 2.68. For a forward bias causing a very large forward current flow, the reverse potential across the junction is very nearly, eliminated.

Values of forward bias of more than a few tenths of a volt result in relatively large junction currents. However, because of the resistivity of the P and N materials and the voltage drop across these materials due to the junction current, a considerably larger voltage (in the order of 1 volt) must be applied across the terminals of the PN junction diode to obtain a junction bias of a few tenths of a volt. At high applied forward voltages, the junction potential is essentially zero and the only restriction to current flow is the resistance of the impure germanium.

(a) N-P-N JUNCTION TRANSISTOR CONFIGURATION   (b) P-N-P JUNCTION TRANSISTOR CONFIGURATION

(c) N-P-N JUNCTION TRANSISTOR POTENTIALS
    WITH NO APPLIED POTENTIALS

(d) N-P-N JUNCTION TRANSISTOR BIASING POTENTIALS

(e) N-P-N JUNCTION TRANSISTOR POTENTIALS
    WITH APPLIED BIAS POTENTIALS
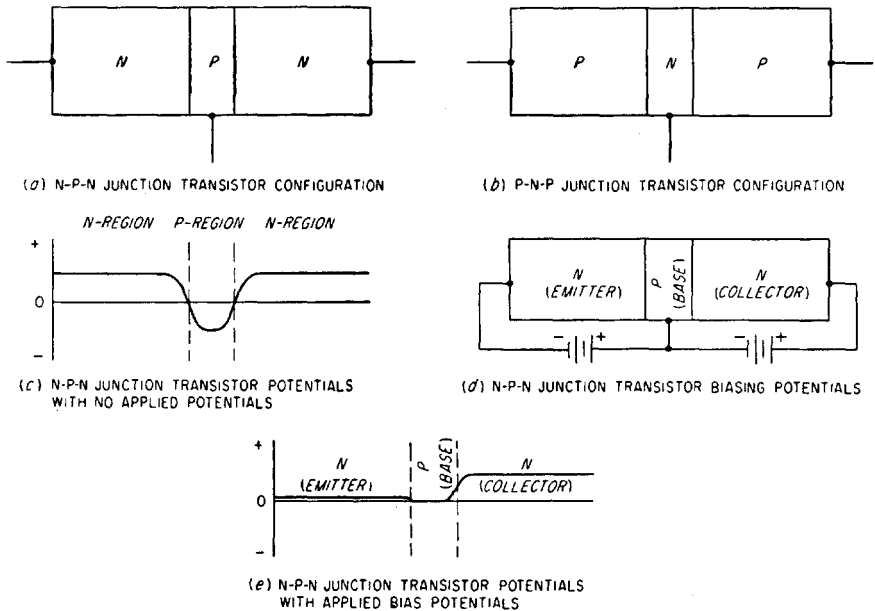
FIG. 2.69. Junction transistor operation.

2.9b. Junction Transistors. By joining another P or N layer to the PN junction, either an NPN or a PNP junction transistor is formed. The junction configurations for these two basic transistor types are shown in Fig. 2.69a and b. The center material is called the *base*, one of the outer layers is called the *emitter* and the other the *collector*.

Junction Transistor Operation. In the NPN junction transistor with no voltages applied, the electrostatic potentials at the two junctions will be as shown in Fig. 2.69c. If the NPN junction transistor is biased as shown in Fig. 2.69d, there will be a relatively large foward current flow across the emitter-base junction, and a relatively low reverse current flowing across the base-collector junction because of the reverse potential applied across this junction. However, if the width of the P-type germanium base is small compared to the average path length of an electron in the base before recombining with a hole, most of the electrons crossing the emitter-base junction will diffuse through the base to the base-collector junction, where they will be drawn across this junction by the applied voltage. Thus, nearly all the emitter electron current will add directly to the collector reverse current. The potentials

across the NPN transistor junctions with applied bias voltages are as shown in **Fig. 2.69e.**

The collector current becomes relatively independent of the collector voltage when the collector junction is biased in the reverse direction since the reverse current is unaffected by the potential gradient across the junction. As a result, the collector resistance r_c may be as high as several megohms. The incremental change in collector current ΔI_c resulting from an incremental change in emitter current ΔI_e is known as the current amplification factor α

$$\alpha = \frac{\Delta I_c}{\Delta I_e}\bigg|_{V_c = \text{constant}} \tag{2.63}$$

where V_c is the collector to base voltage

The value of α is always less than unity in the junction transistor, generally being in the range of 0.95 to 1.0. However, the ratio of change in emitter or collector current to base current is normally in the range of 20 to 100, resulting in high current gain when the junction transistor is operated in the grounded-emitter circuit. Because of the high ratio of output-to-input resistance of the junction transistor, it has a large power gain in the grounded-base circuit even though the emitter-to-collector current gain is less than unity.

The operation of the PNP junction transistor is exactly analogous to operation of the NPN junction transistor except that the bias voltage polarities are reversed and the emitter current carriers are holes instead of electrons.

High-frequency Effects. The operation of junction transistors at high frequencies is adversely affected by three principal factors: collector-junction capacitance, phase shift, and charge-carrier diffusion.

The charge carriers, i.e., electrons and holes, of both base and collector are drawn away from the collector-junction region by the inverse potential, leaving a distribution of charge across the junction due to the donor and acceptor atoms. The density of the charge carriers increases with distance from the junction and decreases with increasing voltage. Thus the junction charge due to donor and acceptor atoms increases with voltage, and the junction exhibits capacitance. The collector-junction capacitance is a function of the junction voltage, current, area and the type of junction. Capacitances in the range of 5 $\mu\mu f$ to 50 $\mu\mu f$ are typical. Because of the high collector resistance, this capacitance severely limits the gain-bandwidth product of the transistor.

Because there is no potential gradient across the base to accelerate the emitter current carriers from the emitter junction to the collector junction, the emitter current carriers diffuse relatively slowly through the base to the collector resulting in an appreciable time delay between emitter and collector currents. This time delay represents a phase shift of the collector currents relative to the emitter current at the signal frequency, and in the grounded-emitter circuit this phase shift may cause a considerable reduction in base-to-collector current amplification β at relatively low frequencies for values of α close to unity. This is the case since $\beta = \alpha/(1 - \alpha)$.

Due to the differences in diffusion paths taken by the emitter current carriers in traveling through the base from emitter to collector as a result of their thermal energies, there is a dispersion in the arrival time at the collector of carriers which left the emitter at the same time. This dispersion results in an increasing degree of cancellation of the signal as the spread in arrival times becomes a large fraction of one cycle of the signal frequency. As a result, α decreases with increasing frequency.

Junction Transistor Construction. Three different manufacturing processes are employed in the construction of junction transistors. In the grown-junction process, a suitable amount of impurity is added to a container of molten germanium to form either P-type or N-type germanium. A seed crystal is lowered into the molten

material and gradually raised as the crystal grows. When the crystal is of sufficient size, enough of the opposite impurity is added to balance out the original impurity and reverse the impurity type. After the crystal has grown a short time further, the impurity balance is again reversed and the crystal allowed to grow until the new layer is of sufficient size. Either PNP or NPN transistors can be constructed by this process. A grown-junction NPN transistor is illustrated in Fig. 2.70a.

In the diffusion process, an impurity such as arsenic is added to molten germanium to create N-type germanium. The solidified metal is then cut into thin disks. By soldering (melting) a small amount of indium or gallium to opposite sides of the disk, some indium or gallium is diffused into the N-type germanium changing it to P-type and forming a PNP transistor. A PNP diffused-junction transistor is illustrated in Fig. 2.70b.

In the etched process, two streams of an electrolyte are directed at opposite sides of an impurity crystal. A current is passed through the two electrolyte streams which attracts the crystal particles and allows them to be washed away. When the crystal is sufficiently thin, the polarity of the two streams is reversed, which allows some

(a) N-P-N GROWN JUNCTION TRANSISTOR

(b) CROSS SECTION OF P-N-P DIFFUSED JUNCTION TRANSISTOR
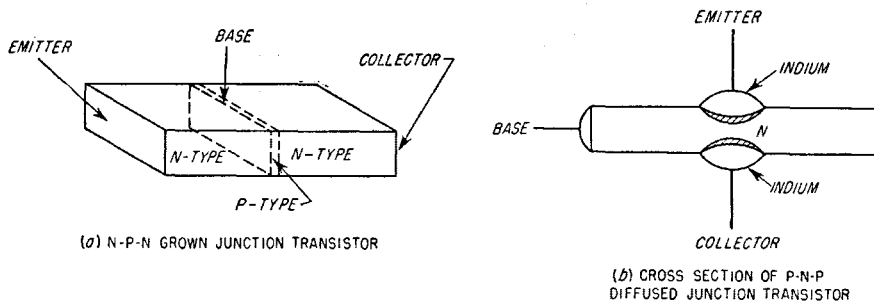
FIG. 2.70. Basic forms of junction transistors.

material of opposite impurity to be deposited. By this technique, transistors having very thin base layers can be constructed with precisely controlled characteristics. The use of a thin base layer greatly improves the high-frequency response of the transistor, but reduces the allowable power dissipation.

Static Characteristics. There are four interrelated parameters which specify the characteristics of transistors and determine the operating point for any particular circuit configuration. These are emitter current, emitter base voltage, collector current, and collector-base voltage. The specification of any two of these parameters determines the other two.

Although there are a number of possible ways to plot transistor characteristics, the two generally accepted families of curves are (1) collector current versus collector voltage for constant values of emitter current, and (2) emitter current versus emitter voltage for constant values of collector current. From these two sets of characteristic curves, all the operating conditions can be determined. Because of the wide use of transistors in the grounded-emitter circuit, manufacturers' data also often include curves of collector current versus collector voltage for constant values of base current. This greatly simplifies the determination of the operating point and load line in grounded-emitter amplifier circuits. Characteristic curves for a typical junction transistor are shown in Fig. 2.71. The collector voltage versus collector current curves in Fig. 2.71a are seen to be very straight, nearly parallel to the voltage axis and quite evenly spaced. These factors indicate very high output resistance and excellent linearity for large signals.

2.9c. Point-contact Transistors. The earliest type of transistor was the point-contact. The construction of a typical PNP point-contact transistor is shown in
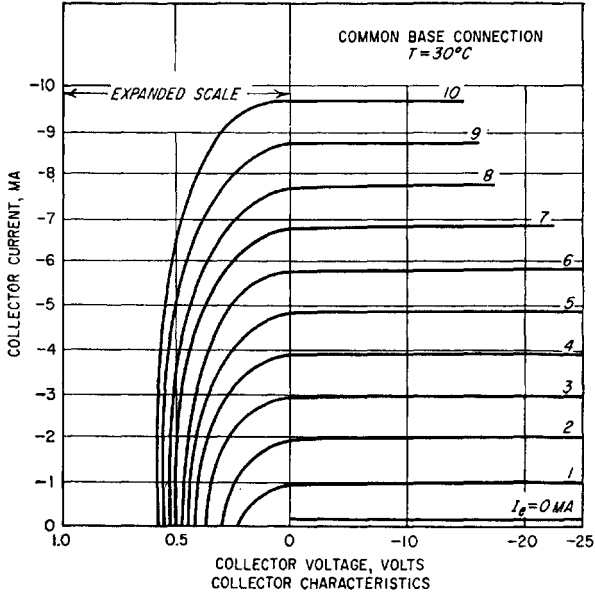
(*a*) COMMON BASE CONNECTION
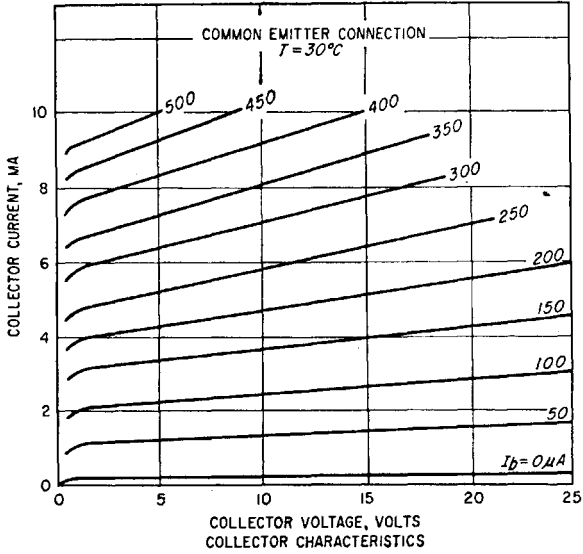
(*b*) COMMON EMITTER CONNECTION

FIG. 2.71. Characteristic curves of a typical junction transistor.

Fig. 2.72. Two closely spaced electrodes having fine points are pressure-mounted against a small block of N-type germanium which forms the base of the transistor. In the manufacturing process a small hemispherical area in the base around each electrode is converted to P-type germanium. One of these areas is the emitter and the other the collector. In the forming process, a relatively large current is passed through the collector electrode for a short period of time to create the desired imperfections in the base material surrounding the collector.

The P-type germanium collector forms a PN junction with the block of N-type germanium. Biasing the collector negatively increases the potential gradient across this junction and reduces the forward current to a value less than the reverse current I_r. When a positive bias is applied to the emitter, current will flow between the emitter and base. This current consists principally of holes moving from emitter to base.

FIG. 2.72. Point-contact transistor construction.

Because the base-collector junction is very close to the emitter-base junction, the holes are drawn to the collector by the potential gradient across the collector junction. Although the exact mechanism of current amplification is not well understood, the result is that current amplification greater than unity is obtained in the point-contact transistor. Values of α of 2 to 3 are typical.

Static Characteristics. The static characteristics of a typical point-contact transistor are given in Fig. 2.73. Comparison with the junction-transistor characteristics of Fig. 2.71 shows that the collector resistance r_c is considerably lower in the point-contact transistor, typically being in the order of 50,000 ohms. The input resistance is quite low, varying from several hundred ohms at low emitter current to approximately 50 ohms at high current levels. Power gains in the order of 20 to 30 db are typical for point-contact transistors.

Point-contact Transistor Frequency Response. The same general effects limit the frequency response of point-contact transistors that limit the frequency response of junction transistors (Sec. 2.9b). However, the lower collector resistance r_c and a smaller collector capacitance due to a smaller collector-junction area make the effects of collector-junction capacitance less serious than for the junction transistor. The effects of carrier dispersion and carrier transit time between emitter and collector limit the frequency response by reducing α with increasing frequency. However, the transit time of minority carriers in the base is less in the point-contact transistor than in the junction transistor because the electric field of the collector extends into the base region in the point-contact transistor.

In point-contact transistors, the transit time of electrons and holes through the germanium semiconductor is given by[1]

$$T = \frac{S^3 \sigma}{\mu I_e} \qquad (2.64)$$

where T = time in seconds required for the electrons or holes to travel from the emitter to the collector

S = spacing between the emitter and collector point contacts, cm

σ = conductivity of the germanium, (ohm-cm)$^{-1}$

[1] W. Shockley, "Electrons and Holes in Semi-conductors," D. Van Nostrand Company, Inc., New York, 1950.

μ = mobility of the holes or electrons, cm^2/volt-sec

$$= \frac{\text{drift velocity of electrons or holes}}{\text{applied electric-field gradient}}$$

I_e = emitter current, amp

From Eq. (2.64) it can be seen that transit time, and consequently the high-frequency response in point-contact transistors, is extremely dependent upon the close spacing of the emitter and collector electrodes. With contact spacings of about 0.003 in., the current amplification factor cutoff frequency f_c is about 1.5 Mc, and for a contact spacing of 0.0005 in., f_c is about 200 Mc. Because of electrical-stability considerations and the difficulty of mechanical construction, most point-contact transistors have an f_c of 2 to 5 Mc.

2.9d. *Transistor Circuit Analysis.* Transistors can be represented schematically as illustrated in Fig. 2.74. The arrow identifies the emitter, and the direction of the arrow indicates the direction of conventional current flow in the emitter.

Transistor Equivalent Circuit. The equivalent circuit of a transistor is shown in Fig. 2.75. The base resistance r_b is common to both the emitter and collector voltage loops. The collector loop contains an equivalent voltage generator $i_e r_m$, where i_e is the incremental emitter current change and r_m is a mutual resistance analogous to mutual conductance in a vacuum tube. The emitter has a dynamic resistance r_e, and the collector has a dynamic resistance r_c.

FIG. 2.73. Characteristic curves of typical point-contact transistor.

The characteristics of transistors may be plotted as families of curves for four combinations of transistor parameters (see Fig. 2.73). If a transistor is considered as a generalized four-terminal network (see Sec. 23.14) the following general equations can be written

$$V_e = i_e R_{11} + i_c R_{12} \qquad (2.65)$$
$$V_c = i_e R_{21} + i_c R_{22} \qquad (2.66)$$

Each of the impedance terms in these equations corresponds to the slope of one of the characteristic curves of the transistor. For small signals, the slopes of the curves may be taken as constants at the appropriate operating point, and from these constants the values for each element in the transistor equivalent circuit can be determined. A comparison of the equivalent circuit of Fig. 2.75 with Eqs. (2.65) and (2.66) yields the following relationships:

$$r_e = R_{11} - R_{12} \qquad (2.67)$$
$$r_b = R_{12} \qquad (2.68)$$
$$r_c = R_{22} - R_{12} \qquad (2.69)$$
$$r_m = R_{21} - R_{12} \qquad (2.70)$$

In addition, the current amplification factor α is given by

$$\alpha = \frac{R_{21}}{R_{22}} = \frac{r_b + r_m}{r_b + r_c} \tag{2.71}$$

The complete equivalent circuits for a transistor including external circuit impedances in the three possible amplifier configurations are shown in Fig. 2.76. The
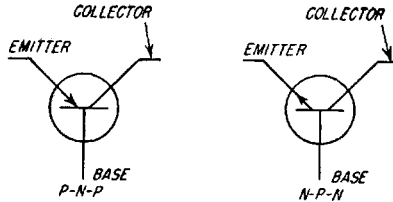
FIG. 2.74. Schematic representation of transistors.
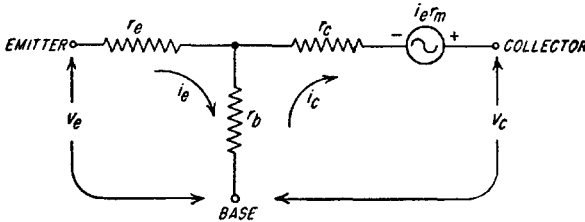
FIG. 2.75. Equivalent circuit of a transistor.

equations for various impedances and gains of these circuits are given below. It should be noted that if there is any external impedance in series with the electrodes shown grounded in Fig. 2.76, it can be taken into account in the equations which follow by adding the impedance directly to the dynamic resistance of that electrode wherever it appears. The image impedances represent the source and load impedances which will match the input and output impedances of the transistor in the indicated circuit.

CASE I—GROUNDED BASE

1. Input impedance:

$$Z_i = r_e + r_b - \frac{r_b(r_m + r_b)}{r_b + r_c + R_L} \tag{2.72}$$

2. Output impedance:

$$Z_o = r_b + r_c - \frac{r_b(r_m + r_b)}{R_s + r_e + r_b} \tag{2.73}$$

3. Forward voltage gain:

$$A = \frac{R_L(r_b + r_m)}{(r_b + r_c + R_L)(r_e + r_b) - (r_b + r_m)r_b} \tag{2.74}$$

4. Forward power gain:

$$G = \frac{A^2 Z_i}{R_L} \tag{2.75}$$

5. Image input impedance:

$$Z_{11} = (r_b + r_e)\sqrt{1 - \frac{\alpha r_b}{r_e + r_b}} \tag{2.76}$$

(a) GROUNDED BASE

(b) GROUNDED EMITTER

(c) GROUNDED COLLECTOR
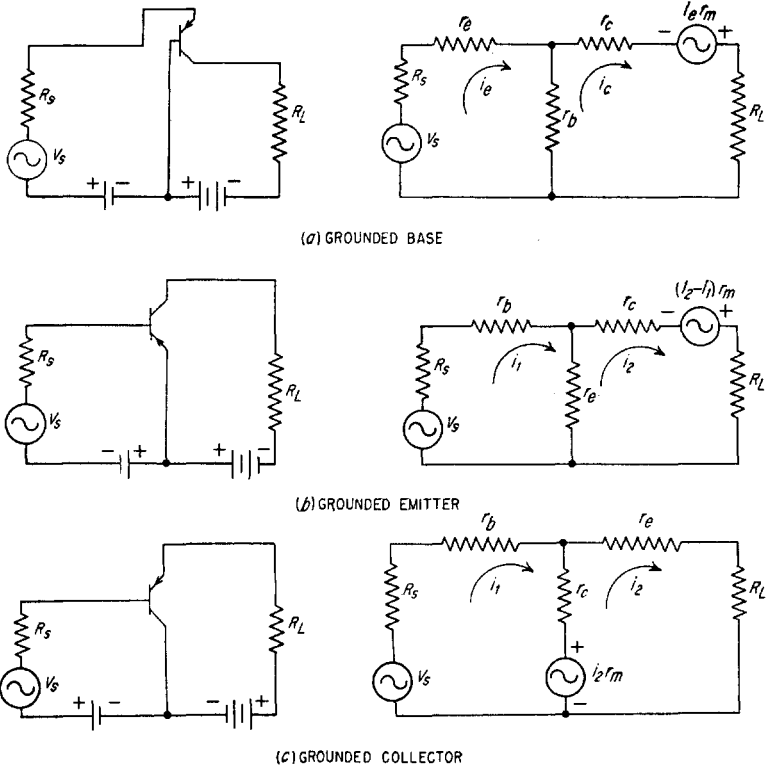
FIG. 2.76. Three basic transistor amplifier equivalent circuits. The arrow at the emitter indicates the conventional current direction. The bias polarities shown are correct for P-N-P type point-contact transistors and P-N-P junction transistors. The current direction arrows and bias polarities are reversed for N-P-N junction transistors.

6. Image output impedance:

$$Z_{12} = (r_b + r_c) \sqrt{1 - \frac{\alpha r_b}{r_e + r_b}} \qquad (2.77)$$

CASE II—GROUNDED EMITTER

1. Input impedance:

$$Z_i = r_b + \frac{r_e(r_c + R_L)}{r_e + r_c + R_L - r_m} \qquad (2.78)$$

2. Output impedance:

$$Z_o = r_c + \frac{(r_e - r_m)(r_b + R_s)}{r_e + r_b + R_s} \qquad (2.79)$$

3. Forward voltage gain:

$$A = \frac{-R_L}{\dfrac{(r_b + r_e)(r_c + R_L)}{r_m - r_e} - r_b} \qquad (2.80)$$

4. Forward power gain:

$$G = A^2 \frac{Z_i}{R_L} \qquad (2.81)$$

5. Image input impedance:

$$Z_{11} = \sqrt{(r_b + r_e)\left(r_b + \frac{r_e r_c}{r_e + r_c - r_m}\right)}$$ (2.82)

6. Image output impedance:

$$Z_{12} = \sqrt{\left[r_c + \frac{(r_e - r_m)r_b}{r_e + r_b}\right](r_c + r_e)}$$ (2.83)

CASE III—GROUNDED COLLECTOR

1. Input impedance:

$$Z_i = r_b + \frac{r_c(r_e + R_L)}{r_c + r_e + R_L - r_m}$$ (2.84)

2. Output impedance:

$$Z_o = r_e + \frac{(r_b + R_s)(r_c - r_m)}{r_b + R_s + r_c}$$ (2.85)

3. Forward voltage gain:

$$A = \frac{r_c R_L}{(r_b + r_c)(r_e + R_L) + r_b(r_c - r_m)}$$ (2.86)

4. Forward power gain:

$$G = A^2 \frac{Z_i}{R_L}$$ (2.87)

5. Image input impedance:

$$Z_{11} = \sqrt{\left(r_b + \frac{r_c r_e}{r_c + r_e - r_m}\right)(r_b + r_c)}$$ (2.88)

6. Image output impedance:

$$Z_{12} = \sqrt{\left[r_e + \frac{r_b(r_c - r_m)}{r_b + r_c}\right](r_e + r_c - r_m)}$$ (2.89)

The variation of input resistance Z_i as a function of R_L is illustrated in Fig. 2.77 for the three basic transistor circuits. The characteristics of a number of junction transistors are given in Table 2.6, and the characteristics of several point-contact transistors are given in Table 2.7.

TABLE 2.6. CHARACTERISTICS OF SEVERAL JUNCTION TRANSISTORS

Tran- sistor	Type	Power gain, db	r_e, ohms	r_b, ohms	r_c, ohms	Base current amplification*	Noise figure, db†	Cutoff freq. f_c, Mc
CK722	PNP	39	25	250	2 megohms	22	25	0.6
CK721	PNP	41	25	700	2 megohms	45	22	0.8
CK725	PNP	42	25	1,500	2 megohms	90	20	1.2
CK727	PNP	36	50	500	2 megohms	25	12	0.8
2N99	NPN	47	...	.....	....	40	20	3.5
2N100	NPN	53	...	.....	....	140	20	5
2N114	PNP	33	...	.....	....	65	...	20

* Ratio of collector current to base current in grounded emitter circuit.
† Measured in a 1-cycle band at 1,000 cycles.

Feedback in Point-contact Transistors. An analysis of the equivalent circuit of a transistor shows that the internal positive current feedback between emitter and

TABLE 2.7. TYPICAL POINT-CONTACT TRANSISTOR CHARACTERISTICS

Type no.	Power gain, db	r_b, ohms	r_e, ohms	α	r_c, ohms	Cutoff freq. f_c, Mc	Noise figure,* db
2N32	21	40	260	2.2	31K	2.7	40
2A	22	500	300	2	10K	2	55
2N52	22	...	...	1.5	....	1	
GE-G11	17	200	275	2.2	22K	2	57

* Noise figure measured in a one-cycle band at 1,000 cycles.

collector will cause oscillation at any frequency where the conditions of the following equation are met.

$$\frac{Z_e + r_e}{Z_b + r_b} \leq \frac{r_m - r_e - Z_e}{Z_c + r_c} - 1 \qquad (2.90)$$

where Z_e, Z_b, Z_c = external circuit impedances in series with the emitter, base, and collector electrodes, respectively

The significance of Eq. (2.90) is twofold. First, if the external impedances are properly chosen so that the left term of the equation is equal to or less than the right
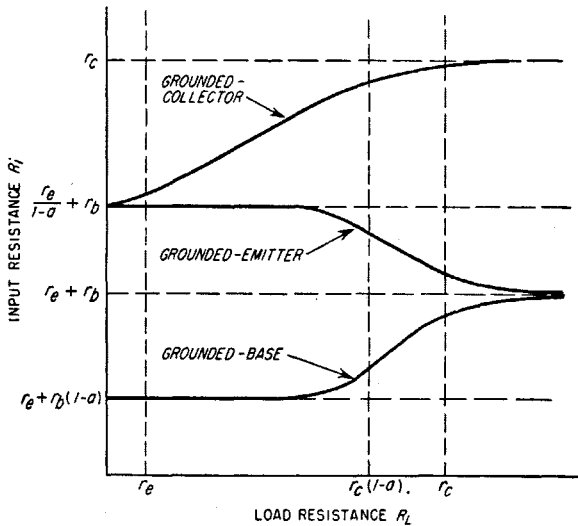
FIG. 2.77. Variation of transistor input resistance with load resistance.

term at any frequency where the current amplification factor α is greater than unity, the transistor will oscillate at that frequency without any external feedback. With a transistor of specified characteristics this can be accomplished by a sufficiently low value of Z_e and Z_c and a sufficiently large value of Z_b. Typical oscillator circuits for a point-contact transistor are shown in Fig. 2.78. When the current amplification factor α is reduced below unity, r_m is reduced in magnitude to the point where the right side of Eq. (2.90) is negative even when Z_e and Z_c are made zero. The transistor cannot oscillate under this condition.

The second ramification of Eq. (2.90) is that a point-contact transistor operating as an amplifier will be unstable at any frequency including zero frequency (d-c) where Eq. (2.90) is satisfied. In oscillator circuits the d-c stability condition must be satisfied. Therefore sufficient resistance must be present in series with either the emitter or the collector to ensure stable operation and prevent possible damage to the transistor in any circuit application.

2.9e. Bias and Stabilization Considerations. In contrast with vacuum tubes, transistors are extremely temperature-sensitive. One effect of a temperature increase is an increase in the *zero-emitter collector current* i_{co}, that is, the collector current with zero-emitter current, and another effect is a decrease in the emitter to base voltage V_{eb}. The increase in i_{co} causes an increase in the total collector current i_c which in turn causes additional collector heating and a further increase in i_{co}. In the extreme case, the transistor will either "run away" and destroy itself, or the drift will be so excessive that the collector-to-base voltage will be reduced nearly to zero, thereby preventing a symmetrical voltage output swing. Similarly, the decrease in V_{eb} causes an increase in the collector current with the same undesirable effects which are associated with the increase in i_{co}. The terms S and S' have been adopted to express the stability of the collector current with respect to changes in i_{co} and V_{eb} respectively and are discussed in the material which follows.

In an a-c amplifier, the problem of temperature stabilization is less severe than in a direct-coupled amplifier since the d-c drift in each stage is not amplified in the following stage or stages. In a high-gain direct-coupled amplifier, a very small amount of drift in a low level stage is amplified and may reach saturation level in the output stage. Stabilization of the individual stages reduces the magnitude of the drift; however, it is almost always necessary to employ over-all amplifier compensation.

(a) PARALLEL RESONANT   (b) SERIES RESONANT   (c) SERIES RESONANT
   BASE CIRCUIT   EMITTER CIRCUIT   COLLECTOR CIRCUIT
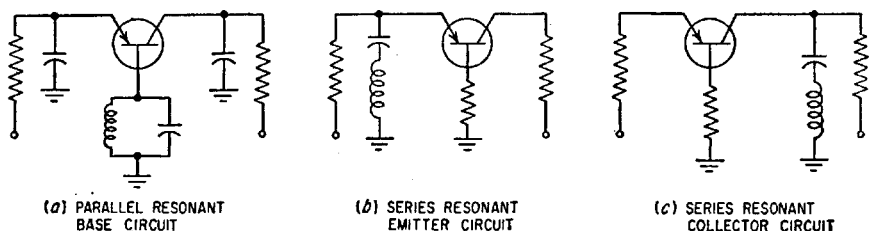
FIG. 2.78. Oscillator circuits for point-contact transistors.

Stability Factor S. The zero-emitter collector current increases approximately 5 to 10 per cent for every degree centigrade increase in temperature. This is equivalent to an increase of i_{co} by a factor of approximately 50 to 200 for an 80°C rise. Associated with the increase in i_{co} is an even greater increase in the total collector current i_c. The amount of increase in i_c is a function of the circuit configuration. In order to express the effect the change in i_{co} has on i_c, the term *stability factor S* has been originated. S is defined as the ratio of the incremental change in the total collector current to an incremental change in the zero-emitter collector current.

$$S = \frac{\partial i_c}{\partial i_{co}} \tag{2.91}$$

As the value of S is decreased, the circuit becomes increasingly stable; however, associated with the increase in stability is a decrease in circuit efficiency. For a given transistor amplifier, it is possible to establish the circuit values which are required to obtain a particular value of S. Ordinarily, resistance in the base lead is to be avoided since it increases S, thus contributing to temperature instability. If

the external base resistance is equal to zero, S is very nearly equal to 1.   In grounded-emitter and grounded-collector circuits, it is possible to have extreme temperature instability due to the external base resistance.   In these applications, it is advisable to establish the circuit parameters which provide satisfactory values of S.
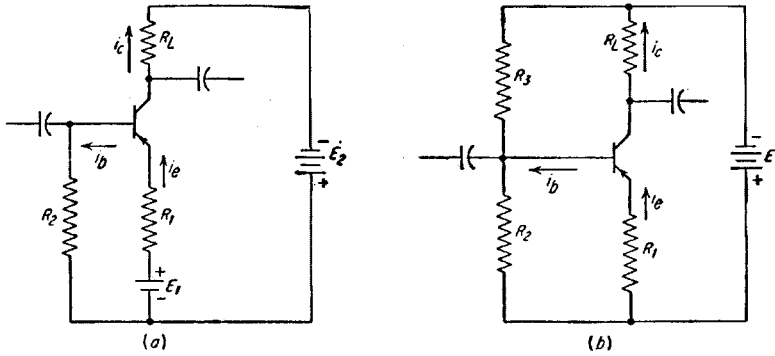
FIG. 2.79. Typical grounded emitter bias configuration.

The method of determining the stability factor of a circuit is best delineated by considering a simple example.   Assuming α to be constant over the temperature range, Eqs. (2.92) and (2.93) can be written for Fig. 2.79a.

$$i_c = i_e - i_b \tag{2.92}$$

$$i_c = \alpha i_e + i_{co} \tag{2.93}$$

$$\therefore i_b = \frac{i_c - i_{co}}{\alpha} - i_c \tag{2.94}$$

$$i_e = \frac{i_c - i_{co}}{\alpha} \tag{2.95}$$

Also

$$E_1 = R_1 i_e + R_2 i_b + V_{eb} \tag{2.96}$$

$$E_2 = R_L i_c - R_2 i_b + V_c \tag{2.97}$$

Substituting the values for i_b and i_e in Eq. (2.96) and solving for i_c results in

$$i_c = \frac{\alpha(E_1 - V_{eb}) + i_{co}(R_1 + R_2)}{R_1 + R_2(1 - \alpha)} \tag{2.98}$$

and from Eq. (2.91)

$$S = \frac{\partial i_c}{\partial i_{co}} = \frac{R_1 + R_2}{R_1 + R_2(1 - \alpha)} \tag{2.99}$$

Equation (2.98) can be rewritten in terms of S.

$$i_c = S i_{co} + \frac{E_1(S - 1)}{R_2} \tag{2.100}$$

From Eqs. (2.100) and (2.99)

$$E_1 = \frac{R_2(i_c - S i_{co})}{S - 1} \tag{2.101}$$

$$R_2 = \frac{E_1(S - 1)}{i_c - S i_{co}} \tag{2.102}$$

$$R_1 = \frac{E_1[1 - S(1 - \alpha)]}{i_c - S i_{co}} \tag{2.103}$$

From Eqs. (2.94) and (2.97)

$$E_2 = i_c \left[R_L + \frac{R_2(\alpha - 1)}{\alpha} \right] + \frac{R_2 i_{co}}{\alpha} + V_c \qquad (2.104)$$

The power P_{dc} delivered by the two batteries is given by Eq. (2.105).

$$P_{dc} = E_1 i_e + E_2 i_c$$
$$= \frac{E_1(i_c - i_{co})}{\alpha} + E_2 i_c \qquad (2.105)$$

The power dissipated in the transistor is very nearly equal to $V_c i_c$.

The equations for S, i_c, R_1, R_2, E_1, and E_2 are interdependent since only four of these variables can be specified independently. Consequently, it is necessary to specify four of these parameters in order to establish the remaining two. For example, for a particular transistor and value of S, the value of E_1 can be determined from Eq. (2.101) for any desired value of R_2 since i_{co} can be established from the transistor characteristics and the desired value for i_c can be specified. In addition, α can be established from the transistor characteristics, and V_c can be specified. The remaining steps consist of calculating R_1 from Eq. (2.103) and E_2 from (Eq. 2.104). The power dissipated in the transistor is equal to $V_c i_c$, and the total power taken from the batteries is given by Eq. (2.105). As the value of S is decreased, additional power must be delivered by the batteries for stabilization.

As the temperature is increased, the zero-emitter collector current increases by an amount Δi_{co}. Therefore, the total collector current i_c' at the elevated temperature is given by

$$i_c' = i_c + S \Delta i_{co} \qquad (2.106)$$

where i_c = initial collector current

Having established the value of i_c' at the maximum operating temperature, it is possible to determine the value of V_c', that is, the collector-to-base voltage at the maximum operating temperature. For the circuit shown in Fig. 2.79a, V_c' is given by Eq. (2.107).

$$V_c' = E_2 - i_c' \left[R_L + \frac{R_2(\alpha - 1)}{\alpha} \right] - \frac{R_2 i_{co}'}{\alpha} \qquad (2.107)$$

where i_{co}' = zero-emitter current at the elevated temperature

The power dissipated in the transistor at the elevated temperature is equal to $V_c' i_c'$. The value of S is based on the allowable collector-current drift, allowable maximum power dissipation, and the permissible power dissipation in the stabilizing network. In general, a value of S between 3 and 5 is found to be adequate for a temperature increase of 80°C when using germanium transistors.

Since the zero-emitter collector current is considerably lower for silicon transistors than for germanium transistors, larger values of S can be used in circuits utilizing silicon transistors. In either case, there is no general rule and each circuit must be considered as an individual problem.

A summary of the equations applicable to the circuit shown in Fig. 2.79b are given below.

$$S = \frac{1 + R_1/R_2 + R_1/R_3}{1 - \alpha + R_1/R_2 + R_1/R_3} \qquad (2.108)$$

$$R_1 = \frac{\alpha(E - V_c - R_L i_c)}{i_c - i_{co}} \qquad (2.109)$$

$$R_2 = \frac{S - 1}{\dfrac{(1 - S + \alpha S)(i_c - i_{co})}{\alpha(E - V_c - R_L i_c)} - \dfrac{i_c - S i_{ce}}{E}} \qquad (2.110)$$

$$R_3 = \frac{E(S - 1)}{i_c - S i_{co}} \qquad (2.111)$$

$$i_c = Si_{co} + \frac{E(S-1)}{R_3} \tag{2.112}$$

$$V_c = E\left[1 - \frac{S-1}{R_3}\left(\frac{R_1}{\alpha} + R_L\right)\right] - i_{co}\left[\frac{R_1(S-1)}{\alpha} + SR_L\right] \tag{2.113}$$

$$P_{dc} = Ei_c + \frac{(V_c + R_L i_c)(i_c - Si_{co})}{S-1} \tag{2.114}$$

As in the case of the circuit shown in Fig. 2.79a, the initial steps ordinarily consist of specifying the desired value of S and establishing the desired values for V_c and i_c from the transistor characteristics. From the transistor curves, the values of both α and i_{co} can also be determined. For a particular gain, the value of R_L will also be specified. It is then possible to establish the relationship between E and R_1 and between E and R_2 [Eqs. (2.109) and (2.110)]. Therefore the assignment of a value to R_1 establishes E and R_2. The value of R^1 can then be determined from Eq. (2.111). The collector-to-base voltage at an elevated temperature can be determined from Eq. (2.113) by substituting i'_{co} for i_{co}.

A value of S larger than unity arises as a result of the change in i_{co} in the base lead causing a shift in the bias voltage across the emitter-base junction because of resistance R_b in the base lead. This increase in forward bias with increasing temperature causes an increase in I_e, $(1 - \alpha)I_e$ of which appears in the base lead in a direction opposite to Δi_{co}. As a result, the shift in forward bias across the emitter-base junction is nearly compensated for, but at the expense of an increase in I_e and I_c. The introduction of a resistance R_e into the emitter lead causes the increase in I_e arising from Δi_{co} to develop a counteracting bias across R_e, which also compensates for the original change in bias caused by Δi_{co} flowing in R_b. Since ΔI_e is $1/(1 - \alpha)$ times greater than the net current change in the base lead, the presence of resistance in the emitter lead greatly reduces the current change required to return the emitter-base junction bias to the original value. The stability factor S is related only to the ratio R_b/R_e and α, and is given by Eq. (2.115) *for any circuit configuration.*

$$S = \frac{1 + R_b/R_e}{1 + (1-\alpha)R_b/R_e} \tag{2.115}$$

where R_b = effective d-c resistance in base lead, e.g., R_2 in parallel with R_3 in Fig. 2.79b
$\quad\quad R_e$ = effective d-c resistance in emitter lead

Stability Factor S'. The *stability factor S'* is the ratio of the incremental change in collector current to an incremental change in the emitter to base voltage. From Eq. (2.98),

$$S' = \frac{\partial i_c}{\partial V_{eb}} = \frac{-\alpha}{R_b(1-\alpha) + R_e} \tag{2.116}$$

where R_b and R_e are defined in Eq. (2.115). The incremental change in the collector current resulting from a change in V_{eb} can be determined from Eq. (2.117).

$$\Delta i_c = S'K\Delta T \tag{2.117}$$

where ΔT = incremental temperature change in degrees centigrade
$\quad\quad K = -0.0018V/C°$ for silicon
$\quad\quad\quad = -0.0014V/C°$ for germanium

Note that a large value of R_b reduces S' [see Eq. 2.116)] and increases S [see Eq. (2.115)]. *In general, only in those applications using low-power silicon transistors is it necessary to calculate S' and determine the drift in collector current due to variations in V_{eb}.* For other transistors, the effects of i_{co} variations are far more significant.

Methods of Stabilization. As previously stated, more and more power must be dissipated in the biasing network as the stability factor is decreased. In the power output stage of an a-c coupled amplifier, the amount of power which must be dis-

sipated in the biasing network to achieve the desired stability factor is often so large that other techniques must be employed. It frequently is advantageous to utilize an additional transistor (which also provides additional gain) or some nonlinear element such as a diode or thermistor for temperature stabilization.

With reference to Fig. 2.80a, the transistor $T1$ has been added to stabilize $T2$. An increase in temperature would cause the collector voltage on both $T1$ and $T2$ to become more positive when considered individually. However, because of the coupling between $T1$ and $T2$, the incremental voltage change at the collector of $T1$ is amplified by $T2$ and appears larger in magnitude and with a negative sign at the collector of $T2$. If $T1$ and $T2$ have similar characteristics, this signal will ordinarily more than compensate for the positive signal appearing at the collector of $T2$ due to the temperature increase in $T2$. If a variable resistor is placed in series with the interstage coupling lead, it is possible to establish experimentally an optimum resistor value for tem-
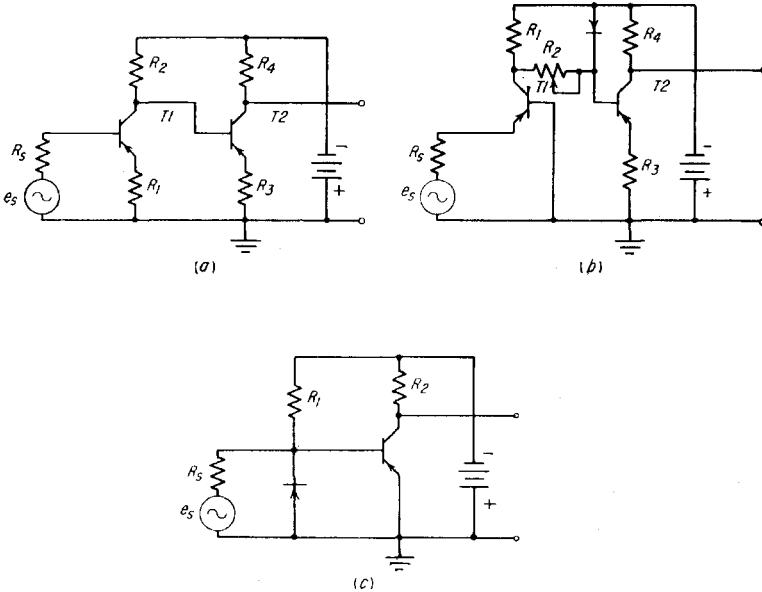
FIG. 2.80. Circuit configurations for temperature stabilization.

perature stabilization. This type of stabilization is not considered to be too practical because of the limited temperature range over which satisfactory stabilization can be achieved. It is of importance to note that whenever two PNP or NPN transistors are d-c coupled, temperature stabilization cannot be achieved unless one of the following circuit configurations is used.

1. Grounded emitter to grounded emitter
2. Grounded base to grounded emitter
3. Grounded emitter to grounded collector
4. Grounded base to grounded collector
5. Grounded collector to grounded base

Good temperature stabilization can be achieved utilizing nonlinear elements. In Fig. 2.80b, the interstage coupling network consists of a diode and a series resistor. If a PN junction diode is used, it is possible to approximate closely the temperature behavior of a junction transistor and thereby achieve better compensation. The value of R_2 should be experimentally adjusted for optimum stabilization.

Another example of the use of a nonlinear element for stabilization is shown in Fig. 2.80c. In the absence of the diode, there is a negative incremental increase in the base-to-ground voltage and a positive incremental increase in the collector-to-ground voltage associated with an increase in temperature. With the junction diode added, an increase in temperature causes the reverse diode current to increase, thereby introducing a positive incremental signal to the base. The positive signal at the base causes the collector current to decrease, thereby tending to compensate for the increase in the transistor zero-emitter collector current.

2.9f. Noise in Transistors. The randomness of motion of the electrons and holes in a transistor gives rise to minute fluctuations in the transistor currents. These fluctuations are called noise currents. The magnitude of the noise current in a transistor is dependent upon the average current in the transistor and the frequency and bandwidth of the measuring device. When there are no bias currents flowing in the transistor, the noise output of the transistor is equal to the thermal noise that would be obtained from a resistor of equal resistance. When a biasing current is present, the transistor noise at low frequencies may increase to as much as 10^4 times the equivalent resistor thermal noise. The noise power per unit bandwidth varies approximately inversely as the frequency of operation up to approximately 100 kilocycles where it has a value several times that of the resistor thermal noise. It remains relatively constant for higher frequencies and is always several times larger than thermal noise.

The noise characteristics of transistors are specified in terms of the transistor noise figure (see Sec. 7.2j). The noise figure for a transistor is usually specified for a one-cycle bandwidth at 1,000 cycles. This measured value can be extrapolated to other frequencies by the relation

$$F_{db} \simeq F_{1,000} + 10 \log_{10} \frac{1,000}{f} \qquad (2.118)$$

where F_{db} = transistor noise figure in decibels for a one-cycle bandwidth at the new frequency f

$F_{1,000}$ = transistor noise figure in decibels for a one-cycle band at 1,000 cps

f = new frequency, cps

The noise figures of junction transistors are considerably lower than those of point-contact transistors. The noise figures of several junction and point-contact transistors are listed in Tables 2.6 and 2.7, respectively.

Example 2.4

Design a class A common-emitter transistor power amplifier utilizing a 2N83 (NPN) transistor. Use a circuit of the type shown in Fig. 2.79a and assume that the source has a resistance of 100 ohms and is a-c coupled to the base. Also assume that transformer output coupling is used rather than the resistance load shown in the figure and that the transformer primary has negligible resistance. Determine R_1, R_2, a-c load impedance, E_2, E_1, maximum power output, and power dissipation, assuming a maximum operating temperature of 70°C. The common-emitter characteristics of the 2N83 transistor are given in Fig. 2.81. Shown in Fig. 2.82a and b, respectively, are the critical runaway voltage V_{crit} (see Fig. 2.82 for definition) and dissipation-derating curves. Note: Construction of load lines and determination of operating points are discussed in detail in Sec. 3.

Solution

1. Determine the maximum allowable transistor dissipation

In Fig. 2.82b, it is seen that the maximum power dissipation at 70°C is 40 per cent of the 25°C power-dissipation rating of the transistor, provided $V_c/V_{crit} \leq 0.01$. Assume that this condition will be satisfied. The 25°C rating of the 2N83 (assuming typical heat sink) is 5 watts (see Fig. 2.81); therefore, the maximum allowable power dissipation $P_{d(max)}$ is

$$P_{d(max)} = 0.4 \times 5 = 2.0 \text{ watts}$$

2. Plot the 2-watt power-dissipation curve on the common-emitter characteristic curves for the 2N83 as shown in Fig. 2.81.

3. Establish a quiescent operating point and construct an a-c load line which provides reasonable linearity for large base-current swings. The quiescent operating point should be on or below the 2-watt power-dissipation curve.

For the load line shown in Fig. 2.81, the operating point is defined by $I_c = 100$ ma and $V_c = 20$ volts. The value of the a-c load resistance R_L is established by the negative inverse of the a-c load-line slope, i.e.,

$$R_L = \frac{20}{0.035} = 571 \text{ ohms}$$

4. Determine the required stability factor S (disregard S').


```
                          MAXIMUM RATINGS

COLLECTOR VOLTAGE : COMMON EMITTER D-C CIRCUIT     -30  VOLTS
(AT 25°C)               COMMON BASE D-C CIRCUIT    -45  VOLTS
                        INSTANTANEOUS              -60  VOLTS

COLLECTOR CURRENT                                 -1000 MA
EMITTER VOLTAGE                                     -6  VOLTS
TOTAL POWER DISSIPATION AT 25°C
     WITH RADIATING FIN F-1, FREE AIR              2.0  WATTS
                          CIRCULATING AIR          5.0  WATTS
     WITH TYPICAL HEAT SINK                        5.0  WATTS
     WITH INFINITE HEAT SINK                      10.0  WATTS
JUNCTION ( AND STORAGE ) TEMPERATURE ( T MAX )     100  °C
                     I_co                          100  μA AT -10V
                                                        AND 25°C
```

Fig. 2.81. Common-emitter characteristics of 2N83 power transistor at 25°C.

With reference to Fig. 2.82a, it can be seen that the larger the value of V_{crit}, the lower must be the stability factor. From Fig. 2.82b, the smaller the value of V_{crit}, the less power can be dissipated in the transistor. Ordinarily, in a power amplifier, a value for V_{crit} which places V_c/V_{crit} at the break point of the curves in Fig. 2.82b is considered to be satisfactory; i.e., operation at this point provides maximum allowable power dissipation with the maximum circuit stability factor for any given operating temperature. Therefore,

$$V_{crit} = \frac{V_c}{0.01} = \frac{20}{0.01} = 2,000 \text{ volts}$$

For a dissipation of 5 watts at 25°C, this requires a circuit stability factor of 5.3 or less (see Fig. 2.82a). Therefore, make $S = 5$.

5. Determine the necessary value of emitter resistance R_1. From Eq. (2.99)

$$S = \frac{R_1/R_2 + 1}{R_1/R_2 + (1 - \alpha)}$$

or

$$\frac{R_1}{R_2} = \frac{1 - S(1 - \alpha)}{S - 1}$$

In the region of the operating point, a 100-ma change in emitter current causes a 90-ma change in collector current (see Fig. 2.83). Therefore, $\alpha = 0.9$ and

$$\frac{R_1}{R_2} = \frac{1 - 5(0.1)}{4} = 0.125$$

To minimize loading of the source, R_2 should be approximately $10R_s$. Therefore, let $R_2 = 1,000$ ohms. Then $R_1 = 125$ ohms.

(a)

(b)
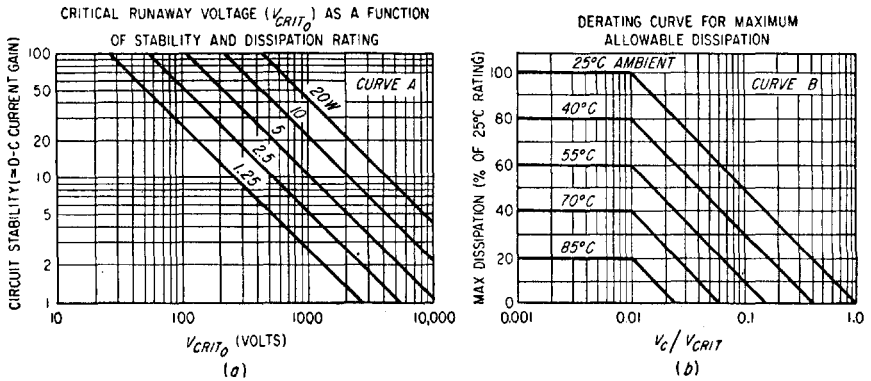
A. Find V_{crit_0} from curve A. This is the critical collector voltage ($I_{co} = 100\mu a$) at which the transistor will run away at negligible dissipation at 25°C. It is a function of transistor dissipation rating and circuit stability. The circuit stability S equals dI_c/dI_{co} or the change in collector current for a change in cutoff current. It is approximately equal to the d-c current gain of the circuit.

B. Determine V_{crit} from the following relation

$$V_{\text{crit}} = V_{\text{crit}_0} \frac{100}{I_{co}(\mu a) \text{ at } 25°C}$$

C. Use V_{crit} in curve B to find the per cent of rated dissipation allowed for the given ambient temperature and collector voltage.

FIG. 2.82. Critical runaway voltage and derating curves for germanium transistors. The value of $V_{\text{crit}} = V_{\text{crit}_0}$ for a 2N83, since I_{co} at 25°C is equal to $100\mu a$. (*a*) Critical runaway voltage. (*b*) Maximum allowable dissipation. The dissipation limits must be derated when the transistor is operated at elevated ambient temperatures. In addition, it is necessary to reduce further the allowable dissipation to prevent a condition known as "runaway." This condition, common to all germanium transistors, results when the dissipation, due to the product of V_c and ΔI_c is sufficient to increase I_c through self-heating rapidly enough to cause unstable equilibrium. The effect is a rapid build-up of collector dissipation beyond the allowable limit. Curves A and B permit a calculation of the maximum allowable dissipation for given ambient and operating conditions.

6. Determine the bias supply voltages E_1 and E_2

Since emitter-base characteristics for the 2N83 transistor are not given, it is assumed that the emitter-base potential necessary to obtain the quiescent-operating-point conditions is zero volts. The error in this assumption will not be greater than 0.1 or 0.2 volt for a germanium transistor and about 0.6 volt for a silicon transistor.

From the characteristic curves of Fig. 2.81, the quiescent conditions of operation are

$$I_c = 100 \text{ ma}$$
$$I_b = 7 \text{ ma}$$
$$V_c = 20 \text{ volts}$$
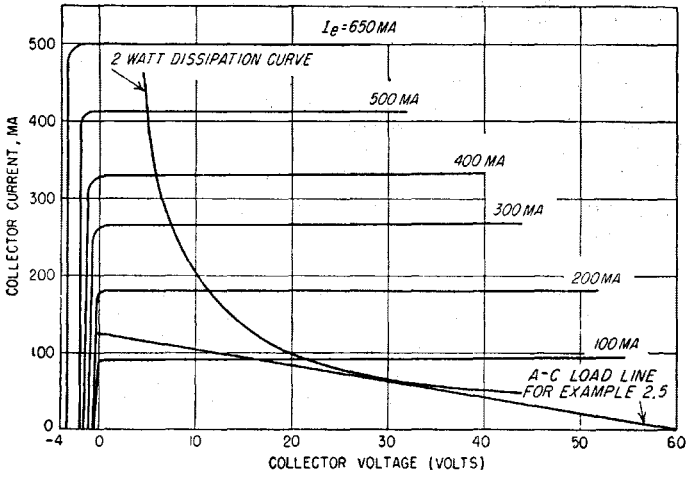
FIG. 2.83. Common-base characteristics of 2N83 power transistor.

(a)   (b)
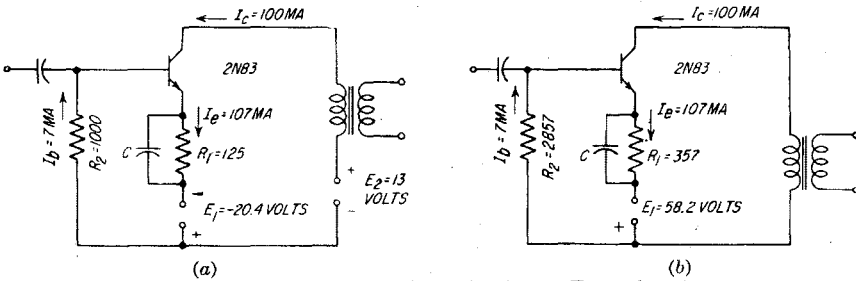
FIG. 2.84. Transistor amplifier circuits for Example 2.4.

Since these currents are much larger than I_{co}, the emitter current is given closely by

$$I_e = I_b + I_c$$
$$= 107 \text{ ma}$$

From the circuit of Fig. 2.79a and remembering that V_c is the collector-to-base voltage, it is seen that the following two equations can be written to describe the static conditions of operation (direction of I_1 and I_2 as shown in Fig. 2.79a).

$$E_2 = V_c + I_b R_2$$
and
$$E_1 = I_e R_1 + I_b R_2$$
Therefore
$$E_2 = 20 + 1,000(-7 \times 10^{-3})$$
$$= 13 \text{ volts}$$
$$E_1 = 125(-107 \times 10^{-3}) + 1,000(-7 \times 10^{-3})$$
$$= -20.4 \text{ volts}$$

7. Determine the circuit input power.

$$P_{dc} = E_2 I_c + E_1 I_e = 13 \times 0.10 + 20.4 \times 0.107$$
$$= 3.48 \text{ watts}$$

8. Determine the a-c output power and the circuit efficiency.

$$P_o = \frac{(V_c - V_{c(\min)})(I_{c(\max)} - I_c)}{2}$$

$$= \frac{(20 - 0.5)(0.135 - 0.100)}{2}$$

$$= 0.341 \text{ watts}$$

$$\eta_c = \frac{P_o}{P_{dc}} = \frac{0.341}{3.48} = 9.8\%$$

The complete circuit of the transistor amplifier is shown in Fig. 2.84a. This circuit requires two bias supplies of odd voltage values. By altering the values of R_1 and R_2, keeping their ratio constant to maintain the required value of stability factor S, the values of E_1 and E_2 can be varied to more suitable voltages. If R_2 cannot be reduced in value, any change from the circuit of Fig. 2.84a will result in decreased circuit efficiency because of increased power loss in R_1 as its value increases. The circuit of Fig. 2.84b where E_2 has been reduced to zero illustrates the possible range of supply-voltage variation for a constant value of S. The supply voltage E_1 can be further increased to any desired value by also increasing R_1 to maintain the emitter-base voltage approximately equal to zero. Resistor R_2 cannot be changed from 2,857 ohms without a collector circuit bias supply. Therefore, increasing E_1 and R_1 will reduce S, increasing the d-c stability of the circuit at the expense of increased power dissipation in R_1.

Example 2.5

Design a class A common-base transistor power amplifier utilizing the 2N83 (NPN) transistor. Assume that the source resistance is 100 ohms and is capacitively coupled to the emitter. Assume that transformer output coupling is used and that the transformer primary has negligible resistance. Determine R_1, R_L, E_1, E_2, maximum power output, and power dissipation. Assume a maximum operating temperature of 70°C. The common-base characteristics of the 2N83 transistor are given in Fig. 2.83.

Solution

1. Determine the maximum transistor dissipation.

From step 1 of the previous example, the maximum allowable power dissipation is

$$P_d = 2.0 \text{ watts} \qquad \text{assuming } V_c/V_{\text{crit}} \leq 0.01$$

2. Plot the 2-watt dissipation curve on the common-base characteristic curves of the 2N83 as shown in Fig. 2.83.

3. Construct a load line and determine an operating point which provides good linearity for large emitter-current swings.

For the load line shown in Fig. 2.83 the operating point (point of tangency) is $I_c = 62.5$ ma, $V_c = 30$ volts

$$R_L = \frac{30 \text{ volts}}{62.5 \text{ ma}} = 480 \text{ ohms}$$

4. Determine the required stability factor S.

The necessary value of V_{crit} (see Step 4 of Example 2.4) is

$$V_{\text{crit}} = \frac{V_c}{0.01} = \frac{30}{0.01} = 3,000 \text{ volts}$$

For a dissipation of 5 watts at 25°C, this requires a circuit stability factor of 3.5 or less.

For the common-base circuit with zero external base resistance ($R_2 = 0$) $S = 1$ [see Eq. (2.99)]. Therefore, stability is assured.

5. Determine the value of R_1

The circuit is shown in Fig. 2.85. To minimize the loading of the source, R_1 should be approximately $10R_s$ or greater. Therefore, let $R_1 = 1,000$ ohms.

6. Determine the required bias supply voltages E_1 and E_2.

From the circuit of Fig. 2.85 and remembering that the emitter-base voltage is very closely equal to zero, the equations determining the bias voltages E_1 and E_2 are (current direction as shown in Fig. 2.85)

FIG. 2.85. Grounded-base circuit of Example 2.5.

$$E_1 = -I_e R_1$$
$$E_2 = V_c$$

From the characteristic curves of Fig. 2.83, it is seen that $I_e = 0.067$ amperes. Therefore,

$$E_1 = -0.067 \times 1,000$$
$$= -67 \text{ volts}$$
$$E_2 = +30 \text{ volts}$$

7. Determine the bias supply input power.

$$P_{dc} = E_2 I_c + E_1 I_e$$
$$= 1.88 + 4.49$$
$$= 6.37 \text{ watts}$$

8. Determine the a-c output power and the collector circuit efficiency.

$$P_o = \frac{(V_c - V_{c(min)})(I_{c(max)} - I_c)}{2}$$

$$= \frac{(30 - 0)(0.125 - 0.0625)}{2}$$

$$= 0.94 \text{ watts}$$

$$\eta_c = \frac{P_o}{P_{dc}} = \frac{0.94}{6.37} = 14.8\%$$

A comparison of the 2N83 in the two circuits illustrates the advantages of each circuit type

Parameter	Common-base	Common-emitter
P_d	6.37 watts	3.48 watts
P_o	0.94 watts	0.34 watts
η	14.8%	9.8%
linearity	Excellent	Fair

It should be noted that the use of RC coupling into the transistor and the 1,000 ohm value of the coupling resistor have resulted in greatly reduced circuit efficiency. Transformer input coupling is much more desirable from an efficiency standpoint.

2.9g. Transistor h Parameters. In order to conveniently define the characteristics of transistors in common-base, common-emitter, and common-collector circuits, six families of curves are required. The necessary families are:

Common-base Circuit

1. Input—V_e versus I_e for constant values of output terminal voltage V_c
2. Output—I_c versus V_c for constant values of input current I_e (see Fig. 2.83)

Common-emitter Circuit

3. Input—V_b versus I_b for constant values of output terminal voltage V_c
4. Output—I_c versus V_c for constant values of input current I_b (see Fig. 2.81)

Common-collector Circuit

5. Input—V_b versus I_b for constant values of output terminal voltage V_e
6. Output—I_e versus V_e for constant values of input current I_b

The family of curves (see Fig. 2.83) which defines the transistor output characteristics for a common-base circuit can be compared to the standard E_b-I_b characteristic curves for a vacuum tube. In each case, the slope of the curves is equal to the output admittance. The vertical spacing between the curves represents the change in output current for an incremental change in bias, viz., the incremental change in emitter current in the transistor and the incremental change in grid voltage in the vacuum tube. For a transistor in which V_c is held constant, the vertical separation of the curves defines the current amplification α between the emitter and the collector, that is, $\alpha = \partial I_c / \partial I_e \big|_{V_c}$. In the case of a vacuum tube in which E_b is held constant, the vertical spacing defines the g_m of the tube, that is, $g_m = \partial I_b = \partial E_c \big|_{E_b}$. The significant observation from the preceding is that the information contained in the family of curves defining the output characteristics for a transistor in a common-base circuit is primarily (1) the transistor *output admittance* for the input a-c open-circuited (note that a constant emitter current denotes an infinite impedance source

which is an effective input a-c open circuit); and (2) the *forward current amplification* from the emitter to the collector for the output a-c shorted (note that a constant collector voltage denotes a zero impedance collector voltage source which is an effective output a-c short circuit). The symbols which have been recommended by RETMA for these two characteristics are h_{ob} and h_{fb}, respectively. In a similar manner, 10 additional *h parameters* are utilized to describe the characteristics of the five other families of curves. Most manufacturers are now supplying transistor data in terms of these parameters. A complete tabulation of the *h* parameters is as follows:

Common-base Circuit (b)

h_{ib} = input resistance with output a-c short-circuited
h_{rb} = reverse voltage gain with input a-c open-circuited
h_{fb} = forward current gain with output a-c short-circuited $= \alpha$
h_{ob} = output admittance with input a-c open-circuited

where
$$r_b = \frac{h_{rb}}{h_{ob}} \tag{2.119}$$

$$r_c = \frac{1 - h_{rb}}{h_{ob}} \tag{2.120}$$

$$r_e = h_{ib} - \frac{h_{rb}(1 - |h_{fb}|)}{h_{ob}} \tag{2.121}$$

$$r_m = \frac{|h_{fb}| - h_{rb}}{h_{ob}} \tag{2.122}$$

Common-emitter Circuit (e)

h_{ie} = input resistance with output a-c short-circuited
h_{re} = reverse voltage gain with input a-c open-circuited
h_{fe} = forward current gain with output a-c short-circuited $= \beta$
h_{oe} = output admittance with input a-c open-circuited

where
$$r_b = h_{ie} - \frac{h_{re}(1 + h_{fe})}{h_{oe}} \tag{2.123}$$

$$r_c = \frac{1 + h_{fe}}{h_{oe}} \tag{2.124}$$

$$r_e = \frac{h_{re}}{h_{oe}} \tag{2.125}$$

$$r_m = \frac{h_{fe} + h_{re}}{h_{oe}} \tag{2.126}$$

Common-collector Circuit (c)

h_{ic} = input resistance with output a-c short-circuited
h_{rc} = reverse voltage gain with input a-c open-circuited
h_{fc} = forward current gain with output a-c short-circuited
h_{oc} = output admittance with input a-c open-circuited

where
$$r_b = h_{ic} - |h_{fc}|(1 - h_{rc}) \tag{2.127}$$

$$r_c = \frac{|h_{fc}|}{h_{oc}} \tag{2.128}$$

$$r_e = \frac{1 - h_{rc}}{h_{oc}} \tag{2.129}$$

$$r_m = \frac{|h_{fc}| - h_{rc}}{h_{oc}} \tag{2.130}$$

The terms r_b, r_c, r_e, and r_m can be determined from Eqs. (2.119) to (2.130). The circuit input and output impedances, gain, etc., can then be determined from Eqs. (2.72) to (2.89).

Voltage Amplifiers

3.1. Classes of Operation and Types of Coupling. In general, the use of an amplifier is necessitated by the requirement for reproducing a given signal at either (1) a higher voltage level or (2) a higher power level. Although in many cases both results are accomplished, the two objectives are separated for design considerations into (1) voltage amplifiers and (2) power amplifiers. The special design considerations for power amplifiers are treated in Sec. 4.

3.1a. Amplifier Classifications. The tube used as either a voltage or power amplifier can be employed in any one of the following classes of operation:[1]

Class A Amplifier. The grid bias and alternating grid voltages are such that plate current in a specific tube flows at all times.

Class AB Amplifier. The grid bias and alternating grid voltages are such that plate current in a specific tube flows for more than half but less than the entire electrical cycle.

Class B Amplifier. The grid bias is approximately equal to the cutoff value so that the plate current is approximately zero when no exciting grid voltage is applied and so that plate current in a specific tube flows for approximately one-half of each cycle when an alternating grid voltage is applied.

Class C Amplifier. The grid bias is appreciably greater than the cutoff value so that the plate current in each tube is zero when no alternating grid voltage is applied and so that plate current in a specific tube flows for appreciably less than one-half of each cycle when alternating grid voltage is applied.

To denote that grid current does not flow during any part of the input cycle, the suffix 1 can be added to the letter or letters of the class identification, e.g., class AB1. The suffix 2 can be used to denote that grid current flows during some part of the cycle.

3.1b. Types of Coupling. A tube used for voltage amplification is usually operated class A and at low to moderate plate supply voltages and plate currents. The types of amplifiers, denoted by the types of coupling employed, fall into four general categories:

1. *The resistance-coupled amplifier,* shown in Fig. 3.1a, is characterized by the resistive plate load R_b, grid resistor R_c, and coupling capacitor C_c. This form of voltage amplifier is the most widely used because of the ease with which broad frequency coverage can be achieved and the economy in the cost, size, and weight of the components required.

2. *The transformer-coupled amplifier,* shown in Fig. 3.1b, can provide much more gain per tube than the corresponding resistance-coupled amplifier since a voltage gain can be derived from a step-up ratio in the transformer. Amplification of very low and very high frequencies is difficult because of the inherent limitations of the transformer.

3. *The impedance-coupled amplifier,* shown in Fig. 3.1c, provides a method of obtaining a high value of plate-load impedance with a low d-c voltage drop. With the exception that there is no transformation ratio, most of its operating characteristics are the same as those of the transformer-coupled stage.

4. *The direct-coupled amplifier,* shown in Fig. 3.1d, is the only type that has a frequency response without lower limit.

[1] As defined by the Institute of Radio Engineers, Standards on Electronics, 1938.

3.2. Amplifier Equivalent Circuits. The performance of tube circuits can be analyzed by the use of equivalent circuits. For relatively small input signals,[1] the tube can be assigned a linear transfer coefficient in the equivalent circuit. The transfer coefficient can be (1) in terms of μ and relate the equivalent generator voltage to the input (signal) voltage or (2) in terms of g_m and relate the equivalent generator current to the input voltage.

It must be emphasized that this method of analysis treats only the varying (a-c or signal) components of current and voltage, not the quiescent values.

(a) RESISTANCE COUPLING   (b) TRANSFORMER COUPLING

(c) IMPEDANCE COUPLING   (d) DIRECT COUPLING
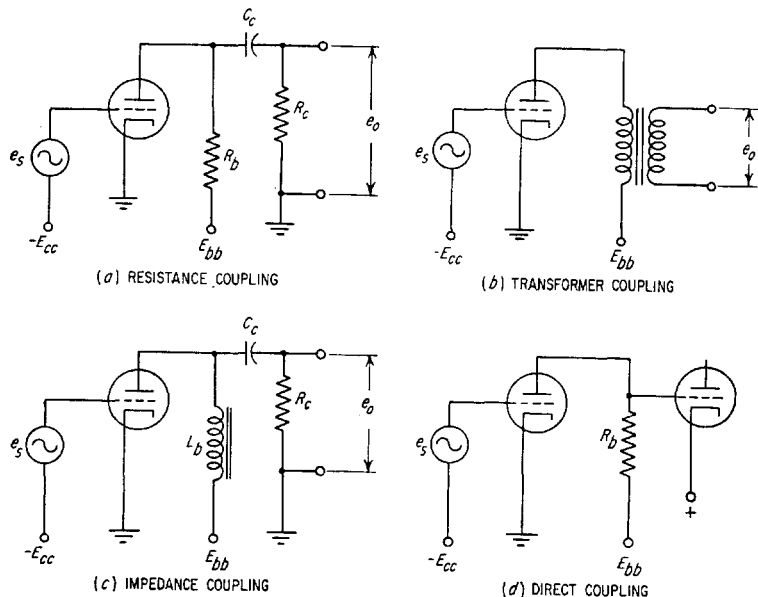
FIG. 3.1. Types of coupling in amplifiers.

(a)   (b)
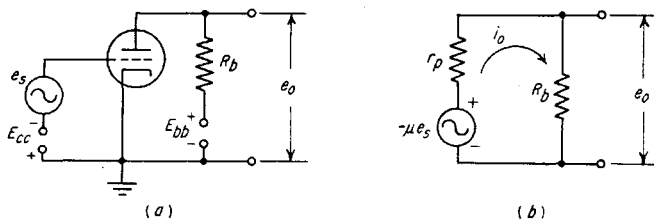
FIG. 3.2. Amplifier and equivalent circuit.

3.2a. Equivalent Voltage or Current Generator. The circuit shown in Fig. 3.2a can be replaced by the simple equivalent plate circuit of Fig. 3.2b. The grid circuit and all the factors pertinent to the quiescent condition only (E_{cc}, E_{bb}) are dropped out in the equivalent diagram. The vacuum tube as a voltage amplifier can always[2] be replaced by a generator whose open-circuit voltage is μ times the grid-cathode signal voltage and whose internal resistance is equal to the tube dynamic plate resistance r_p.

---

[1] The small signal limitation is to minimize the variation in tube constants with applied signal.

[2] This equivalent-circuit discussion assumes compliance with the small signal limitation. Also, the frequency of operation has been assumed to be in a region where tube capacitances, transit time, and other high-frequency effects need not be considered.

Using the equivalent circuit, the output signal current i_o and the output signal voltage e_o can be computed:

$$i_o = -\frac{\mu e_s}{r_p + R_b} \tag{3.1}$$

$$e_o = i_o R_b = -\frac{\mu e_s R_b}{r_p + R_b} \tag{3.2}$$

The load resistor R_b in Fig. 3.2a can be replaced with any type of load (inductor, tuned circuit, etc.) if a path of d-c continuity is provided for the flow of plate current. A generalized load impedance Z_b can be substituted for R_b in the equivalent circuit and equations. The impedance of reactive loads varies as a function of frequency; therefore any numerical solution of the equations will be for a specified frequency. If the load Z_b is composed of two or more parallel current paths, the solution will be expedited if this parallel load is converted to the series equivalent by the methods of complex algebra. If the load Z_b has a reactive component, i.e., a phase angle other than 0° at the specified frequency, i_o and e_o will each have a phase angle with respect to the signal voltage e_s.[1]

3.2b. Constant Voltage Generator. Examination of Eq. (3.2) indicates that if R_b were very large compared to r_p, the output voltage would be essentially equal to
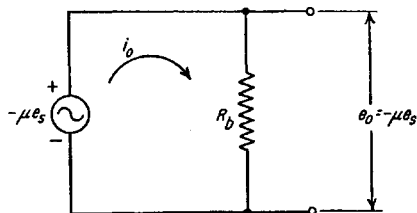
FIG. 3.3. Constant voltage generator equivalent circuit for the condition where $R_b \gg r_p$.

$-\mu e_s$ and almost independent of the exact value of R_b. In this special case the tube is equivalent to a constant voltage generator. This concept is useful in the evaluation of circuits with a varying load resistance provided that the minimum value of R_b is very much larger than r_p. If the amplifier in Fig. 3.2a satisfies the large R_b criterion, an equivalent circuit could be drawn as in Fig. 3.3. This is the constant voltage generator equivalent circuit in which e_o is equal to the generator voltage, and the gain of the amplifier e_o/e_s is equal to $-\mu$. When the constant voltage criterion is applicable, the phase angle of the output current is equal to but has the opposite sign from the load phase angle.

This special equivalent circuit is often applicable to an amplifier having an inductor or an unloaded interstage transformer in the plate circuit.

3.2c. Constant Current Generator. Examination of Eq. (3.1) indicates that if R_b were very small compared to r_p, the output current would be given by Eq. (3.3).

$$i_o \simeq -\frac{\mu e_s}{r_p} \qquad \text{when } R_b \ll r_p \tag{3.3}$$

For this particular condition, the output current is virtually independent of the value of the load resistance or impedance, and the tube is equivalent to a constant current generator. In circuits with tubes having a high value of dynamic plate resistance, pentodes in particular, the condition stipulated for Eq. (3.3) is frequently satisfied and the tube can be considered as being a constant-current generator.

A fundamental relationship between tube constants is given by Eq. (3.4).

$$\frac{\mu}{r_p} = g_m \tag{3.4}$$

---

[1] In general, discussions of phase shift in an amplifier will disregard the normal 180° relationship between the grid and plate voltages of each tube and will treat only the departures from this relationship.

By making use of Eq. (3.4) and the relationship given by Eq. (3.3), it is possible to obtain expressions for the output in terms of g_m for the case where $r_p \gg R_b$.

$$i_o \simeq -e_s g_m \tag{3.5}$$

and
$$e_o \simeq -e_s g_m R_b \tag{3.6}$$

Figure 3.4 illustrates the form of the constant current generator equivalent circuit which is valid if r_p is very much larger than the load resistance. The output current i_o is assumed to flow from a constant current generator of infinite internal resistance and is indicated by the arrow on the generator symbol in the drawing. When the constant current generator principle is applicable, the phase angle of the output voltage, with respect to the grid voltage, is equal to 180° plus the phase angle of the load.
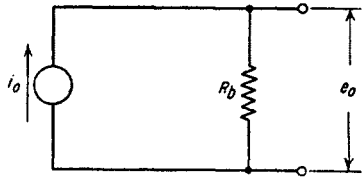
Fig. 3.4. Constant current generator equivalent circuit for the condition where $r_p \gg R_b$.

3.2d. Equivalent Circuits. Neglecting the effects of shunting and interelectrode capacitances, lead inductances, and electron transit time within the tube (see Secs.
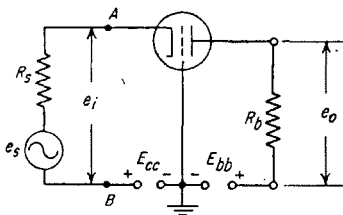
(a) GROUNDED-GRID AMPLIFIER AND DRIVING GENERATOR
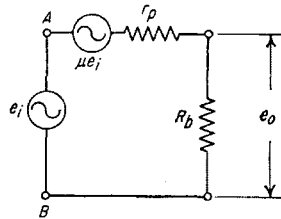
(b) EQUIVALENT AMPLIFIER CIRCUIT
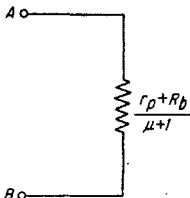
(c) EQUIVALENT INPUT IMPEDANCE
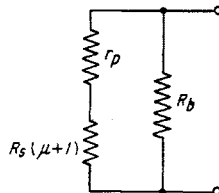
(d) EQUIVALENT OUTPUT IMPEDANCE

Fig. 3.5. Grounded-grid amplifier and equivalent circuits.

3.8b to 3.8f), simple equivalent circuits can be drawn for the operation of most vacuum-tube circuits. In the majority of applications these simplified equivalent circuits are satisfactory; however, at high frequencies it is necessary to consider the effects of shunt capacitances, transit time, etc.

The three basic circuit configurations are shown in Figs. 3.5, 3.6, and 3.7. In addition, the equivalent amplifier circuits and the circuits representing the input and output impedances are included. In the low- and high-frequency cases, these circuits must be modified by those factors which are discussed in paragraph 4 of Sec. 3.2e.

In an amplifier which has its grid stabilized to ground, all plate-circuit impedances (r_p, R_b, etc.) as viewed from the cathode are multiplied by the term $1/(\mu + 1)$.

Under similar conditions, the cathode impedance when viewed from the plate circuit is multiplied by the term $(\mu + 1)$.

1. Grounded-grid Amplifier. An amplifier with a signal source connected to the cathode, as in Fig. 3.5a, will have an equivalent circuit as shown in Fig. 3.5b and input and output impedances as shown in Fig. 3.5c and d.

2. Grid-driven Amplifier. If the driving source is connected to the control grid and the output signal taken from the plate, Fig. 3.6a and b is applicable. If the

(a) GROUNDED CATHODE

(b) GROUNDED CATHODE EQUIVALENT AMPLIFIER AND OUTPUT IMPEDANCE CIRCUIT

(c) GROUNDED PLATE

(d) GROUNDED PLATE EQUIVALENT AMP-LIFIER AND OUTPUT IMPEDANCE CIRCUIT
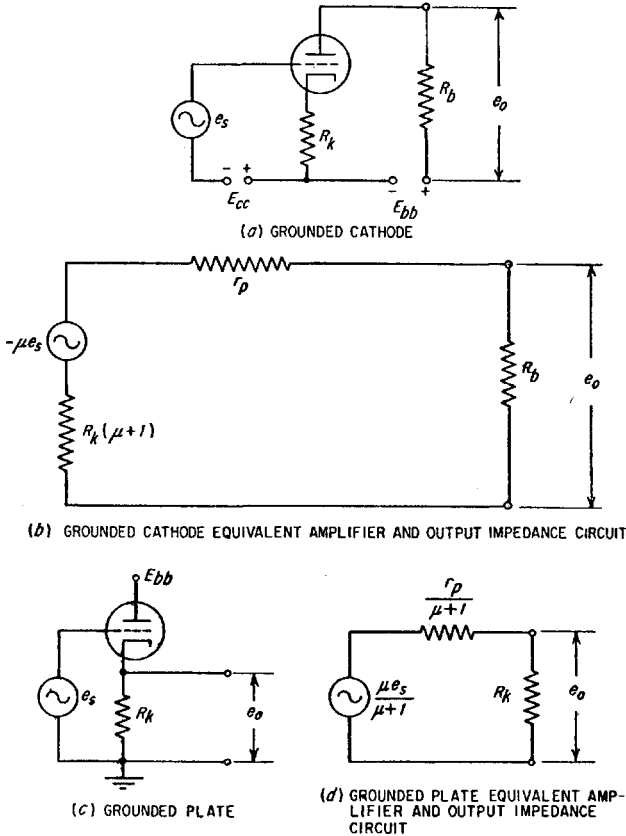
FIG. 3.6. Grid-driven amplifier and equivalent circuit.

driving source is connected to the control grid and the output signal taken from the cathode as shown in Fig. 3.6c, the equivalent circuit of Fig. 3.6d is applicable.

3. Plate-driven Tube. If a driving signal is injected in the plate circuit as shown in Fig. 3.7a, the equivalent amplifier circuit in Fig. 3.7b is applicable.

The impedance into which the driving generator must function is given by the circuit in Fig. 3.7c. It should be noted that R_1, across which an opposing grid signal is developed because of the flow of signal current, is effectively multiplied by the term $(\mu + 1)$.

3.2e. Additional Considerations for Equivalent Circuit Analysis. The following factors must be considered in establishing equivalent circuits:

1. Small Signal Restriction. Equivalent circuit analysis is based on small signal conditions and the tubes being in continuous conduction. Under these conditions

it is possible to obtain solutions with an accuracy which is limited primarily by variations in tube characteristics and circuit components from their nominal values.

   2. Tube Coefficients. The tube coefficients (μ, r_p, g_m) must be valid for the actual operating conditions. There is considerable danger in applying "typical" coefficients without careful regard for the specific application. Many tube data sheets supply the proper values as a function of the tube voltages or current. If this information

(a) AMPLIFIER WITH SIGNAL INJECTED IN PLATE CIRCUIT

(b) EQUIVALENT AMPLIFIER CIRCUIT

(c) EQUIVALENT INPUT IMPEDANCE

(d) EQUIVALENT OUTPUT IMPEDANCE
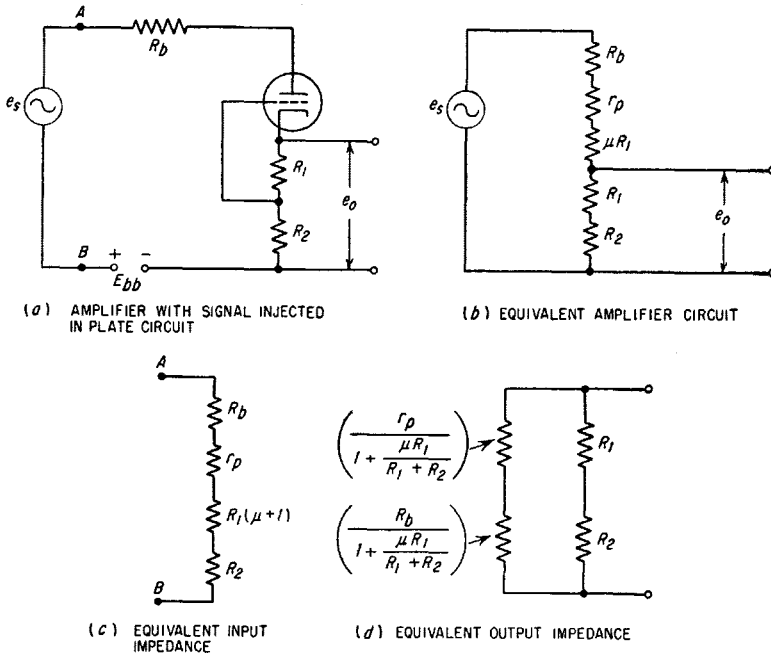
FIG. 3.7. Plate-driven tube and equivalent circuits.

is not available, it should be determined graphically from the tube characteristics in the region of operation.

   3. Linearity of Elements. Some common nonlinear circuit elements are as follows:

   a. Amplifier tubes or diodes in which the operating region passes into both the conducting and nonconducting regions or through the region of excessive curvature of the characteristic curves at low current
   b. Thyrite resistors, inductors, capacitors, or other materials in which the impedance varies as a function of current or voltage
   c. Motors or other electromagnetic devices in which the voltage-current relationship changes as a function of load, saturation, or other factors

   4. Operating Frequency. The principal factors pertinent to the frequency of operation which must be considered are:
   At low frequencies,

   a. Reactance of capacitors used for bypassing and coupling
   b. Reactance of inductors or transformer windings

At high frequencies,

a. Interelectrode and shunting capacitances
b. Inductances due to wire lengths inside and outside the tube
c. Transit time and the associated loading and phase shift
d. Losses due to core materials and/or dielectrics
e. Leakage reactance in transformers

If the equivalent network contains more than one current loop, one of the many conventional network theorems may be used to simplify the problem.

When the restriction concerning small signals is violated, the accuracy of the analysis is progressively reduced as the signal amplitude is increased. Examination of the tube dynamic transfer characteristic (see Sec. 3.3*b*) will indicate the actual limits for good accuracy. These limits are indicated by the range over which the dynamic transfer characteristic is essentially a straight line.

If nonlinear circuit elements are employed, the equivalent circuit analysis can still be used. The method becomes one of reaching a separate solution for each discrete operating condition. In a circuit having two different operating states, e.g., a diode conducting and nonconducting, a separate solution for each state will be valid. For a continuously nonlinear device a number of point solutions must be obtained in the region of interest unless the equation for the nonlinear function is known.

3.3. Graphical Analysis of Triode Amplifiers. Graphical solutions of vacuum-tube amplifier circuits provide data on quiescent operating conditions, dynamic characteristics, and distortion as a function of signal amplitude.

3.3a. Quiescent Operating Conditions. Plate characteristics as supplied by the manufacturer indicate the plate current I_b as a function of plate voltage E_b and grid bias E_c. E_b and E_c are the voltages applied to the tube elements, not the supply voltages.

The use of load lines provides a graphical simultaneous solution for the quiescent tube current, plate voltage, grid voltage, and the voltage drops across the resistances in the plate and cathode circuits.
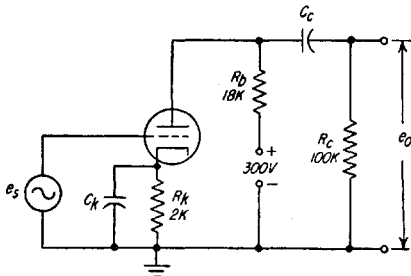
The amplifier in Fig. 3.8 will be used as an example to show the steps in the solution of a resistance-coupled amplifier circuit.[1] The E_b-I_b characteristics for the tube are shown in Fig. 3.9.

FIG. 3.8. Resistance-coupled amplifier.

1. D-C Plate-load Line. The d-c plate-load line can be drawn on the tube characteristics as shown in Fig. 3.9. This line defines the division of the plate supply voltage E_{bb} between the tube and the resistors in the plate and cathode circuits. The load line has a slope equal to $-1/(R_k + R_b)$. Its position is readily established by determining the intercepts on the voltage and current coordinates as follows: (1) if the plate current is zero the plate voltage is E_{bb} and (2) if the voltage drop across the tube is zero, the current must have a value as to produce a voltage drop equal to E_{bb} in the resistors. This current value is equal to $E_{bb}/(R_k + R_b)$.

These two considerations establish the two points between which the d-c plate-load line is drawn, point *A* (300 volts, zero current) and point *B* (zero volts, 15 ma). This line is the locus of the possible d-c operating points for this specific circuit.

2. Bias Line. This line defines the relation between the cathode voltage E_k and the cathode current. This line is plotted against the plate-current scale and the

[1] An amplifier having external grid bias and cathode degeneration is given in Example 3.1.

grid-voltage curves. At any current I_b the product $I_b R_k$ determines the value of E_k, which is also the bias $-E_c$ for the case of the grid returned to ground. In constructing the bias line, select various values of plate current and determine the corresponding cathode voltages. For example, if $R_k = 2,000$ ohms and $I_b = 3$ ma, $E_k = 6$ volts (point C), and when $I_b = 5$ ma, $E_k = 10$ volts (point D). A line passing through points C and D intersects the d-c plate-load line at point O. This is the quiescent operating point.

Because one of the coordinates against which the bias line was plotted is curved, the bias line is also curved. This is particularly noticeable at the low-current values. It is evident that for accuracy the points C and D should be selected near, and preferably bracketing, the d-c plate-load line.
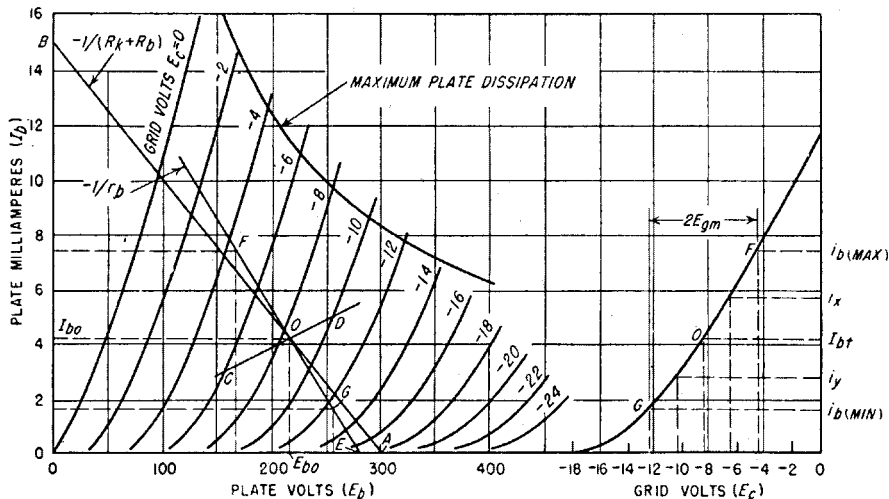
FIG. 3.9. Graphical analysis of a triode amplifier.

3. Quiescent Circuit Values. The point O and its projection on the current and plate voltage coordinates make it possible to tabulate the following data:

$$I_{bo} = 4.2 \text{ ma}$$
$$E_{bo} = 216 \text{ volts}$$
$$E_{co} = -E_{ko} = -I_{bo}R_k = -8.4 \text{ volts}$$
$$\text{Plate dissipation } P_p = I_{bo}E_{bo} = 0.91 \text{ watt}$$
$$\text{Voltage drop in } R_b = I_{bo}R_b = 75.6 \text{ volts}$$
$$\text{Dissipation in } R_b = I_{bo}{}^2 R_b = 0.32 \text{ watt}$$

4. Maximum Plate Dissipation Curve. The maximum allowable plate dissipation expressed in watts can be restated in terms of a maximum steady-state current at any given plate voltage. If these corresponding currents and voltages are plotted on the plate characteristic curves, the resulting curve defines the limits of the permissible quiescent operating area. The maximum plate dissipation curve shown in Fig. 3.9 has been drawn for a maximum plate dissipation rating of 2.5 watts. Point O must always lie below the curve. When the quiescent point is well below this limit, short-duration signal excursions above the curve are permissible.

3.3b. Dynamic Operating Conditions. An amplifier such as the one shown in Fig. 3.8 can be evaluated graphically to determine the gain, maximum signal amplitude, and distortion as demonstrated in the following procedure. The dynamic transfer characteristic is useful in determining the limits of the linear region.

1. A-C Plate-load Line.   The d-c load line was drawn for the quiescent condition. If, for the specific signal frequencies involved, the reactances of C_k and C_c are negligible, the a-c load resistance r_b seen by the tube, from plate to ground, consists of R_b paralleled by R_c.   The a-c load resistance is therefore equal to $R_b R_c/(R_b + R_c)$, or

$$\frac{18 \times 10^3 \times 10^5}{(18 \times 10^3) + 10^5} = 15{,}250 \text{ ohms}$$

The a-c load line does not necessarily pass through the quiescent operating point if the a-c plate-load resistance r_b is not equal to the d-c plate-load resistance R_b. The displacement is a function of the distortion introduced due to the nonlinear tube characteristics.   It should be noted that when distortion does exist, the amount of displacement of the a-c load line from the quiescent point is also a function of the
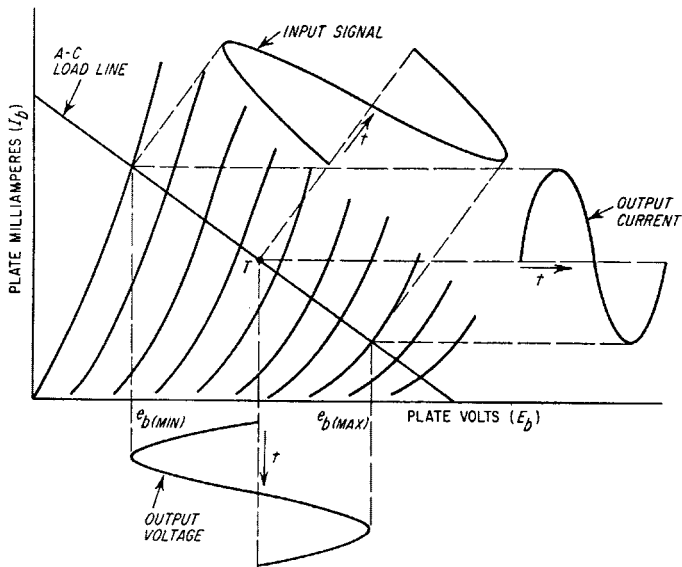
FIG. 3.10. Output voltage and current relationships for a sine-wave input signal.

output signal amplitude.   For very small output signal amplitudes, the displacement approaches zero and the a-c load line can be considered as intersecting the d-c load line at the quiescent point.   In this case, or if the tube characteristics are considered to be linear, the method of plotting the new load line is to first find the value of $I_{bo}r_b$ and add this voltage to E_{bo}.   This new voltage is equal to $0.0042 \times 15{,}250 + 216$, or 280 volts, and establishes the point of intersection between the a-c load line and the zero current axis.   In Fig. 3.9 this has been plotted as point E.   A line drawn from E through O represents the path of operation for a-c signals.

2. Determination of Gain.   Figure 3.10 illustrates the voltage and current relationships during one cycle of the input signal.   The applied signal is represented by an excursion along the a-c load line originating from point T, the effective bias under a-c conditions (identical to the quiescent operating point when no distortion is present), and is measured by the E_c voltage scale.   The input signal has been projected point-for-point onto the plate current and plate voltage scales.

In Fig. 3.9 a signal with peak values of ± 4 volts will swing the instantaneous operating point between points F and G.   The projection of points F and G on the

plate voltage scale determine $e_{b(\min)}$ and $e_{b(\max)}$, the plate voltages at the peak signal excursions from E_{bo}. The gain A is expressed by

$$A = \frac{e_{b(\max)} - e_{b(\min)}}{2E_{gm}} = \frac{257 - 165}{8} = 11.5$$

where E_{gm} = peak value of the input signal

To determine the gain of an amplifier having cathode degeneration, refer to part 3 of Example 3.1.

3. Dynamic Transfer Characteristic. On the right side of Fig. 3.9 the scale along the abscissa has been changed to E_c, that is, the voltage between grid and cathode. From the plate characteristics each intersection of the a-c load line and the grid voltage E_c curves can be used to establish a point to be used in the construction of another curve which is called the dynamic transfer curve. Each point is translated
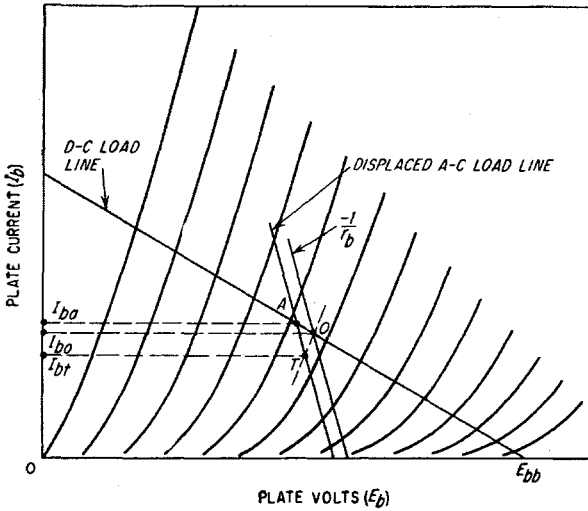
FIG. 3.11. Example of shift from quiescent operating point due to nonlinearity in the tube characteristic.

horizontally, maintaining the same current, and located above the corresponding value of E_c.

The curve in this graph is the dynamic transfer characteristic for the amplifier shown in Fig. 3.8. The upper portion of this plot is virtually a straight line, and operation restricted to this area would be essentially distortionless.

4. Shift in the Operating Point under Dynamic Conditions. In an amplifier that is not perfectly linear, a change in the average plate current will occur when a signal is present. This departure from the quiescent operating point will be related to the degree of nonlinearity and to the signal amplitude. This shift of operating point, sometimes termed "rectification effect," is usually negligible in small-signal voltage amplifiers but frequently of importance in power amplifiers.

The shift from the quiescent operating point under dynamic conditions due to the nonlinearity in the tube characteristic curves results in the establishment of two new and distinct operating points and the displacement of the a-c load line (see Fig. 3.11). One of these is point A on the d-c load line and represents the average plate current with the signal present. The other is point T on the displaced a-c load line and indicates the effective bias with the signal present and consequently the

plate current at the instant the input signal has zero amplitude. The a-c load line is displaced to a position which passes through point A on the d-c load line and is parallel to its position through the quiescent operating point O. If the tube has an external bias supply, point T' will be determined by the intersection of the displaced a-c load line and the same bias voltage curve as at quiescence. If the tube has self-bias, point T will be established by the intersection of the displaced a-c load line and the bias voltage curve determined by the product of the cathode resistance and the average value of the plate current under dynamic conditions.

Point A is located along the d-c load line at a value of plate current I_{ba} given by Eq. (3.7). The process of determining I_{ba} is not direct, however, since Eq. (3.7) contains terms which can be determined only after the a-c load line has been constructed through point A. Consequently, different values for point A must be assumed until the values obtained from the resulting a-c load line and substituted in Eq. (3.7) produce a calculated value of I_{ba} which agrees with the assumed value.

5. Determination of Harmonic Distortion. The harmonic analysis of the output signal can be accomplished by graphically determining plate current for five specific values of a given sine-wave input signal and applying the proper equations. The five input-signal amplitudes and the associated plate-current terms to be read from the a-c load line are as follows:

$$
\begin{array}{ll}
-E_{gm} & i_{b(min)} \\
-0.5E_{gm} & i_y \\
0 & I_{bt} \text{ (current at point } T') \\
+0.5E_{gm} & i_x \\
+E_{gm} & i_{b(max)}
\end{array}
$$

where E_{gm} = maximum peak input signal.
These quantities are shown in Fig. 3.9.

Equation (3.7) permits the determination of the average plate current I_{ba} with the signal present.

$$I_{ba} = 0.167(i_{b(max)} + 2i_x + 2i_y + i_{b(min)}) \tag{3.7}$$

With a sinusoidal signal applied to the input of an amplifier, sinusoidal current and voltage waveforms usually are desired at the output. In Fig. 3.9 the input signal swinging between points F and G produces current variations extending from $i_{b(max)}$ to $i_{b(min)}$ (no shift in operating point is assumed). The asymmetry of these current values with respect to I_{bt} (I_{bo} if no shift in the operating points is considered)[1] indicates the presence of 2nd-harmonic distortion. The peak value of the 2nd harmonic I_{H2} in the current waveform is given by

$$I_{H2} = 0.25(i_{b(max)} + i_{b(min)} - 2I_{bt}) \tag{3.8}$$

for the example given:

$$i_{b(max)} = 7.4 \text{ ma} \qquad i_{b(min)} = 1.7 \text{ ma} \qquad I_{bt} = 4.2 \text{ ma}$$

Therefore

$$I_{H2} = 0.25(7.4 + 1.7 - 8.4) = 0.175 \text{ ma}$$

The percentage of 2nd harmonic is

$$\text{Per cent } H_2 = \frac{3I_{H2}}{i_{b(max)} + i_x - i_y - i_{b(min)}} \times 100 \tag{3.9}$$

$$= \frac{3(0.175)}{8.6} \times 100 = 6.1\%$$

---

[1] In most instances it is possible to determine the harmonics from the nondisplaced a-c load line without introducing appreciable errors.

where $i_x = 5.7$ ma

$\quad i_y = 2.8$ ma

The peak value of the 3rd harmonic is given by

$$I_{H3} = 0.167(i_{b(\max)} - 2i_x + 2i_y - i_{b(\min)}) \qquad (3.10)$$
$$I_{H3} = 0.167(7.4 - 11.4 + 5.6 - 1.7) = -0.0167 \text{ ma}$$

The negative sign indicates that the harmonic subtracts from the fundamental when the fundamental reaches its positive crest value.

The percentage of 3rd harmonic is

$$\text{Per cent } H_3 = \frac{3I_{H3}}{i_{b(\max)} + i_x - i_y - i_{b(\min)}} \times 100 \qquad (3.11)$$
$$= \frac{0.05}{8.6} \times 100 = 0.58\%$$

Similarly, the 4th harmonic may be figured from

$$I_{H4} = 0.083(i_{b(\max)} - 4i_x + 6I_{bi} - 4i_y + i_{b(\min)}) \qquad (3.12)$$
$$\text{Per cent } H_4 = \frac{3I_{H4}}{i_{b(\max)} + i_x - i_y - i_{b(\min)}} \times 100 \qquad (3.13)$$

3.3c. Graphical Analysis of a Degenerative Amplifier Having Both External and Cathode Bias. A problem frequently encountered is the amplifier with a bias determined by a cathode resistor as well as a d-c potential applied to the grid. The quiescent operating condition can be found by a simultaneous solution involving the two bias sources. This can be done graphically, and the method is demonstrated in Example 3.1.

Example 3.1

In Fig. 3.12, the grid can be made either positive or negative with respect to ground. Assuming the tube characteristics to be those shown in Fig. 3.13, determine the quiescent operating points for each bias supply and also the gain.

Solution

1. Positive Grid Potential (Switch in the y Position)

If the external bias supply is a positive voltage, the first step is to establish the value of current through the cathode resistor required to provide zero bias for the tube. This value of current is equal to $E_{cc}/R_k = 20/4000$, or 5 ma, and is plotted as point U in Fig. 3.13.

Additional points to be plotted on Fig. 3.13 are obtained by determining the required plate current for different assumed values of bias. For example, if the bias E_c is assumed to be equal to -6 volts, the required cathode voltage would be equal to 26 volts since the external bias supply voltage is 20 volts. The value of current

FIG. 3.12. Circuit for Example 3.1 to illustrate the graphical solution for the operating point where both external and cathode bias are used.

required for a cathode voltage of 26 volts is $26/R_k$, or 6.5 ma. This value is plotted as point V in Fig. 3.13. A line drawn through several points located in this manner defines the locus of grid-bias voltages associated with values of plate current. The intersection of this bias line with the d-c load line determines the quiescent operating point.

2. Negative Grid Potential (Switch in the x Position)

If the external bias supply is a negative voltage, the first step is to locate point R on Fig. 3.13 which is the bias (external supply voltage) which would exist if no plate current were flowing, i.e., $E_c = -8$ volts. The method of establishing additional points to be used in the plotting of the bias line is to assume values for the bias voltages which are more negative than that of the external supply voltage and determine the required associated values of plate current. For example, if the bias voltage is to be -12 volts, the cathode voltage

must be equal to 4 volts. The required plate current is therefore equal to $4/R_k$, or 1 ma, and is plotted as point S in Fig. 3.13. As in step A, a line can be drawn through similarly located points, and the point of intersection between the constructed line and the d-c load line defines the quiescent point.

3. Determination of Gain

The gain of an amplifier having cathode degeneration can be determined by assuming incremental changes in the external bias supply and noting the incremental change in plate current ΔI_b. The product $\Delta I_b r_b$ represents the incremental change in output voltage.
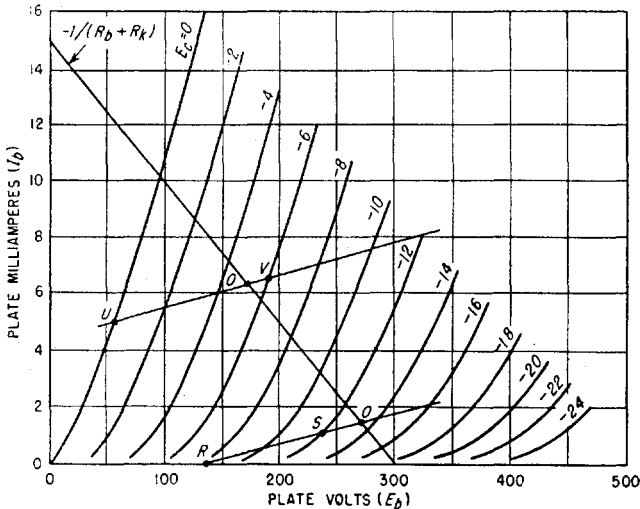
FIG. 3.13. Construction of bias and plate load lines and determination of quiescent operating points.

For example, a 28-volt change in the input voltage ($+20V$ to $-8V$) in this particular problem resulted in a 4.9 ma (6.3 to 1.4 ma) change in plate current. The associated output voltage Δe_o is

$$\Delta e_o = \Delta I_b r_b$$
$$= 4.9 \times 10^{-3} \times 16{,}000$$
$$= 78.4 \text{ volts}$$

and the gain A is

$$A = \frac{\Delta e_o}{\Delta e_s} = \frac{78.4}{28} = 2.8$$

3.4. Graphical Analysis of Tetrode or Pentode Amplifiers. To analyze graphically the operation of a tetrode or pentode amplifier it is necessary to perform the same operations that are required for the analysis of triode amplifiers. The principal difference in the analysis is the necessity of considering the screen current.

3.4a. D-C Plate-load Line. The d-c plate-load line is drawn with a slope of $-1/(R_b + KR_k)$. The factor K is the ratio of the total cathode current, i.e., plate current plus screen-grid current, to the plate current. Note that K is fairly constant over most of the tube characteristics; however, the evaluation of K should be at the approximate operating point. The load line is a straight line drawn between the plate supply voltage E_{bb} on the abscissa and $E_{bb}/(R_b + KR_k)$ on the plate current coordinate.

3.4b. D-C Transfer Characteristic. The d-c transfer characteristic is constructed by plotting the d-c plate current as a function of bias voltage as determined from the d-c plate-load line.

3.4c. Bias Line. The bias line cannot be drawn on the E_b-I_b characteristics as was possible in the case of triode amplifiers. Instead, the bias line must be drawn on the plot of the d-c transfer characteristic. The procedure is to assume various values of plate current I_b and determine the associated product $I_b KR_k$. The self-

bias, $-I_b K R_k$, must then be added to any external bias so that the actual bias E_c associated with the assumed value of plate current I_b can be established and plotted on the d-c transfer characteristic. A series of assumed values of plate current permit the plotting of the bias line which defines the locus of bias values associated with different values of plate current.

3.4d. Quiescent and Dynamic Operating Points. The intersection of the bias line and the d-c transfer characteristic establishes the quiescent values of plate current I_{bo}, plate voltage E_{bo}, and bias E_{co}.

In Sec. 3.3b, the considerations in constructing the a-c load line for a triode were presented. The procedure is identical for a pentode. It was stated that the a-c plate load line does not pass through the quiescent operating point if the d-c plate-load resistance R_b is not equal to the a-c plate-load resistance r_b and provided distortion exists. The a-c plate-load line passes through point A (point of average plate current with signal present) on the d-c load line and has a slope of $-1/r_b$ (see Fig. 3.11). The time axis of the output signal is determined by the location of point T (effective bias with the signal present) on the a-c load line. The determination of points A and T is somewhat complicated, and the procedure recommended in Sec. 3.3b should be followed. In general, the a-c plate-load line will very nearly pass through the quiescent operating point.

3.4e. Dynamic Transfer Characteristic. After the construction of the a-c load line, the dynamic transfer characteristic should be drawn. The procedure is similar to that for the d-c transfer characteristic described in Sec. 3.4b. The dynamic transfer characteristic permits a visual presentation of the amplifier linearity and in addition indicates the instantaneous plate current as the signal changes the instantaneous grid bias between zero volts and cutoff.

3.4f. Determination of Gain. If the amplifier is not degenerative, i.e., if it has negligible impedance in the cathode and screen circuits, the gain can easily be determined by taking an increment of output voltage as measured along the abscissa directly below the a-c plate-load line and dividing this value by the associated change in bias voltage.

If cathode degeneration is present, the gain must be determined by making use of the dynamic transfer characteristic. It is necessary to construct an additional bias line in accordance with the method described in Sec. 3.4c and with the assumption that the external bias supply has had some incremental increase or decrease in voltage. If there is no external bias supply, it is necessary to arbitrarily assume some bias supply voltage such as 1 or 2 volts. The increment of plate current ΔI_b, which is equal to the difference in plate currents as established by the intersection of the two bias lines and the dynamic transfer characteristic, represents the plate current change which can be expected for the assumed incremental change in the input signal voltage Δe_s. The incremental change in output voltage Δe_o is equal to $\Delta I_b r_b$. The gain A is equal to $\Delta e_o/\Delta e_s$.

FIG. 3.14. Pentode resistance-coupled amplifier.

Example 3.2

Determine the gain of the circuit shown in Fig. 3.14, assuming the tube characteristics to be those given in Fig. 3.15.

Solution

1. Draw the d-c plate-load line.

The d-c plate-load line is drawn with a slope of $-1/(R_b + K R_k)$. The value of K is determined as in Sec. 3.4a. Examination of the tube characteristic shown in Fig. 3.15 indicates that at the arbitrarily selected plate voltage of 160 volts and zero bias, the plate

current is 9 ma and the screen current is 2.6 ma.   The value of K in this instance is 11.6/9, or 1.3.   This value of K will be valid for all bias voltages provided that the plate voltage is 160 volts.   Actually, the value of K is not critical with changes in plate voltage unless the plate voltage is lowered to and below values associated with the knees of the various bias curves.

$$R_b + KR_k = 29,000 + (1.3)(1,000)$$
$$= 30,300 \text{ ohms}$$

In Fig. 3.15 the d-c plate load line is drawn as a straight line between the plate supply voltage, 300 volts, on the abscissa and $E_{bb}/(R_b + KR_k)$, or 9.9 ma, on the ordinate.
    2. Construct the d-c transfer characteristic.
    The d-c transfer characteristic is drawn as stipulated in Sec. 3.4b and is shown in Fig. 3.15.
    3. Draw the bias line.
    The bias line is drawn with the same coordinates as the d-c transfer characteristic. Since there is no external bias supply, the effective bias is equal to $-I_b KR_k$.   The values of E_c are therefore equal to $-1,300 I_b$.   The intersection of this line and the d-c transfer characteristic defines the quiescent operating point.   In this example, $E_{co} = -3.3$ volts, $E_{bo} = 222$ volts, and $I_{bo} = 2.5$ ma.
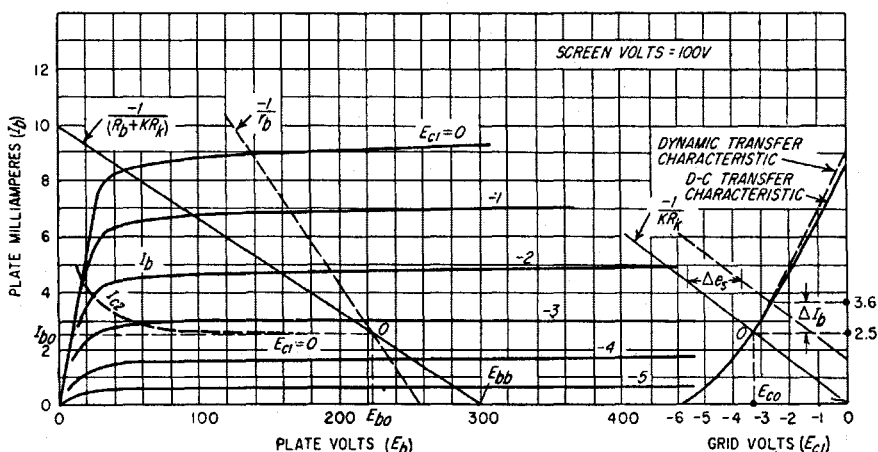
Fig. 3.15. Graphical analysis of the circuit shown in Fig. 3.14.

4. Draw the a-c load line and dynamic transfer characteristic.
    It was stated in Sec. 3.4d that the a-c load line does not necessarily pass through the quiescent operating point if R_b does not equal r_b and if distortion exists.   To determine the exact location of the a-c load line, follow the procedure described in Sec. 3.3b.   Normally the analysis is sufficiently accurate if the a-c load line is drawn through the quiescent operating point provided that it is not necessary to make a harmonic analysis and provided that the plate current does not approach cutoff or exceed that value associated with zero bias.   Therefore, in this example, the a-c load line has been drawn through the quiescent point.   The a-c load r_b is determined as follows:

$$r_b = KR_k + \frac{R_b R_c}{R_b + R_c} = (1.3 \times 1,000) + \frac{29,000 \times 20,000}{29,000 + 20,000}$$
$$= 13,140 \text{ ohms}$$

The value of $I_{bo} r_b$ should be added to E_{bo}.   This voltage,

$$(0.0025 \times 13,140) + 222 = 254.8 \text{ volts}$$

establishes the point of intersection between the a-c load line and the abscissa.   A straight line drawn from this point through the quiescent operating point therefore locates the a-c plate-load line.
    The dynamic transfer characteristic can then be drawn as stated in Sec. 3.4e.
    5. Determine the amplifier gain.
    It is necessary to draw a new bias line based on the addition of a d-c signal in the grid

circuit. The procedure is given in Sec. 3.4c, and in this example a +2 volt signal has been assumed. Construct a line parallel with the original bias line but displaced horizontally by an amount equivalent to the value of the assumed signal. The difference in plate currents associated with the intersections of the two bias lines and the dynamic transfer characteristic represents the incremental a-c current change associated with the assumed increment of signal input. The gain A is determined as follows:

$$A = \frac{\Delta e_o}{\Delta e_s} = \frac{\Delta I_b R_b R_c}{\Delta e_s (R_b + R_c)}$$

$$= \frac{(0.0036 - 0.0025)29,000 \times 20,000}{2(29,000 + 20,000)}$$

$$= 6.51$$

3.5. Gain and Bandwidth. Resistance-coupled Amplifiers. Equations for the gain of a resistance-coupled amplifier with and without degeneration are given in Secs. 3.5a and 3.5b, and the equations for determining the lower and upper frequency limits are given in Secs. 3.5c and 3.5d, respectively.

3.5a. Voltage Gain. The voltage gain can be calculated from the known tube characteristics and the circuit resistances or impedances. The midband gain is generally determined with the assumption that the plate-load impedance is

FIG. 3.16. Resistance-coupled amplifier.

equal to the resistive component r_b. For this case the gain of a triode or multielement tube without degeneration can be expressed by Eq. (3.14) or (3.15).

$$A = \frac{e_o}{e_s} = \frac{-\mu r_b}{r_p + r_b} \tag{3.14}$$

$$= \frac{-g_m r_b}{1 + \dfrac{r_b}{r_p}} \tag{3.15}$$

$$\simeq -g_m r_b \qquad \text{when } r_p \gg r_b$$

where r_b = a-c plate-load resistance = $\dfrac{R_b R_c}{R_b + R_c}$ (see Fig. 3.16)

μ, r_p, g_m = constants of the tube under specific operating conditions

3.5b. Gain with Degeneration. In an amplifier, the signal which is amplified is that portion of the input signal which exists between the grid and the cathode. An amplifier which has a cathode circuit impedance in series with the input signal as shown in Fig. 3.17a will have reduced gain since the cathode circuit voltage, caused by cathode current through the cathode resistance, introduces a signal voltage which opposes the input voltage and, consequently, reduces the grid to cathode voltage. A circuit of this type is said to be degenerative. In general, this form of degeneration will be present whenever an unbalanced (one terminal grounded) generator is employed with an amplifier having a cathode impedance. The gain of an amplifier A' with cathode degeneration is given by Eq. (3.16) or (3.17).

$$A' = \frac{-\mu r_b}{r_p + r_b + R_k(\mu + 1)} \tag{3.16}$$

$$= \frac{-g_m r_b}{1 + g_m R_k + \dfrac{R_k + r_b}{r_p}} \tag{3.17}$$

$$\simeq \frac{-g_m r_b}{1 + g_m R_k} \qquad \text{if } r_p \gg (R_k + r_b)$$

In the case of a tetrode or pentode amplifier, Eq. (3.17) is based on the screen grid being bypassed to the cathode. If the screen grid is bypassed to ground or in the event there is no screen voltage dropping resistor in a tetrode or pentode amplifier having cathode circuit degeneration, Eq. (3.19) is applicable if R_s is made equal to zero.

Not all amplifiers which have an impedance in the cathode circuit are degenerative. The amplifier with the transformer-coupled input shown in Fig. 3.17b is an example if it is assumed that the impedance formed by R_2 in parallel with C_1 is negligible at the signal frequency. The signal voltage drop across R_1 does not alter the grid-to-cathode voltage and consequently does not cause degeneration. Resistor R_2 is to provide the proper bias and does not cause degeneration since it is bypassed. This technique of lifting the generator by a voltage resulting from the generator output as shown in Fig. 3.17b is called *bootstrapping*. In bootstrapping amplifiers, the same gain can be realized with a resistance R_1 in the cathode circuit as that realized with the same resistor in the plate circuit. By bootstrapping, the low output resistance normally associated with the cathode is lost, e.g., the output resistance in Fig. 3.17b is the same as it would be if R_1 were in the plate circuit.

(*a*) DEGENERATIVE   (*b*) NOT DEGENERATIVE
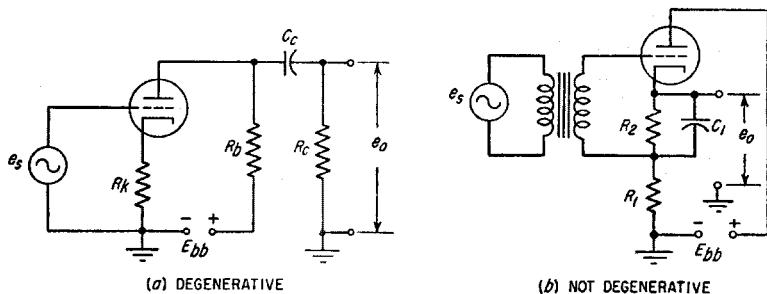
FIG. 3.17. Amplifiers with resistance in the cathode lead.

A tetrode or pentode amplifier with an unbypassed resistance or impedance in the screen circuit will cause degeneration. The computation of amplifier gain A'' with the screen circuit degeneration is complicated by the requirement of tube constants not normally supplied by the manufacturers.

$$
\begin{aligned}
A'' &= -\frac{r_b}{1 + r_b/r_p}\left(g_m - \frac{g_{g2g1}g_{pg2}R_s}{1 + R_s/r_{g2}}\right) \\
&= -\frac{g_m r_b}{1 + r_b/r_p}\left(1 - \frac{1}{1 + r_{g2}/R_s}\right)
\end{aligned}
\tag{3.18}
$$

where g_{g2g1} = transconductance, control grid to screen grid
   g_{pg2} = transconductance, screen grid to plate
   R_s = resistance, or signal impedance, in screen circuit
   r_{g2} = dynamic resistance of the screen grid

If the cathode circuit is degenerative, any bypass for the screen circuit should return to the cathode, not to ground.

The gain A''' with both screen and cathode degeneration is given by Eq. (3.19).

$$
A''' \simeq \frac{-g_m r_b\left(r_{g2} - \dfrac{R_k}{g_{pg2}r_p}\right)}{R_s + r_{g2} + R_k\left[g_m\left(r_{g2} + \dfrac{1}{g_{pg2}}\right) + 1 + g_{pg2}r_{g2}\right]}
\tag{3.19}
$$

3.5c. Low-frequency Response. At some low frequency f_1, the gain of a resistance-coupled amplifier will be 3 db less than the gain at midband as given by Eq. (3.14) or (3.15). This -3-db frequency, or lower half-power frequency, is nominally taken as the lower limit of bandwidth. For an amplifier of the general form shown in Fig.
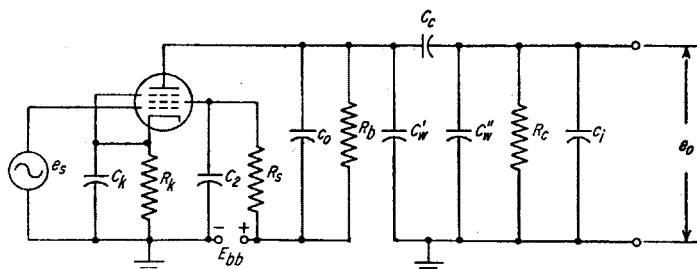
FIG. 3.18. Pentode resistance-coupled amplifier with circuit capacities shown.
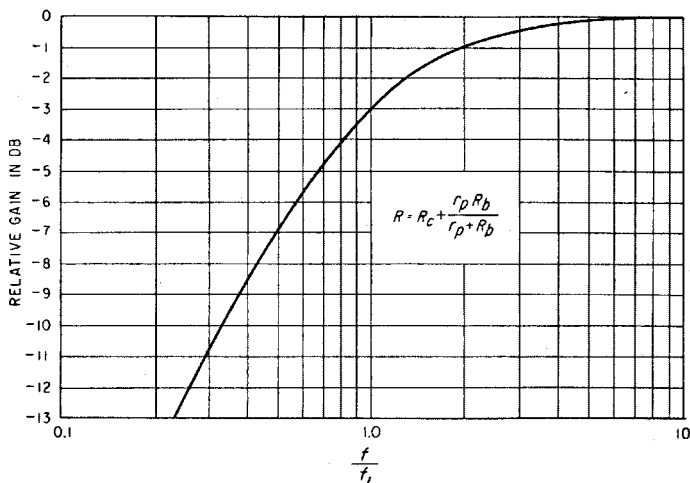
FIG. 3.19. Low-frequency loss characteristics due to R-C_c interstage coupling.

3.18 and using either triode or multielement tubes, f_1 will be given by Eq. (3.20) provided that no degeneration exists in either the screen or cathode circuits.[1]

$$f_1 = \frac{1}{2\pi C_c \left(R_c + \dfrac{r_p R_b}{r_p + R_b} \right)} \tag{3.20}$$

In many circuits R_c will be very much larger than r_p in parallel with R_b. When this is true Eq. (3.20) may be simplified to Eq. (3.21).

$$f_1 \simeq \frac{1}{2\pi C_c R_c} \qquad \text{when } R_c \gg \frac{r_p R_b}{r_p + R_b} \tag{3.21}$$

The low-frequency attenuation and phase characteristics are shown in Figs. 3.19 and 3.20, respectively.

[1] The effects of degeneration in the cathode and screen circuits are treated in Secs. 3.10a to 3.10c.

3.5d. High-frequency Response.   At some high frequency f_2, the gain of a resistance-coupled amplifier will also be 3 db less than the gain at midband as given by Eq. (3.14) or (3.15).   This -3-db frequency, or upper half-power frequency, is nominally
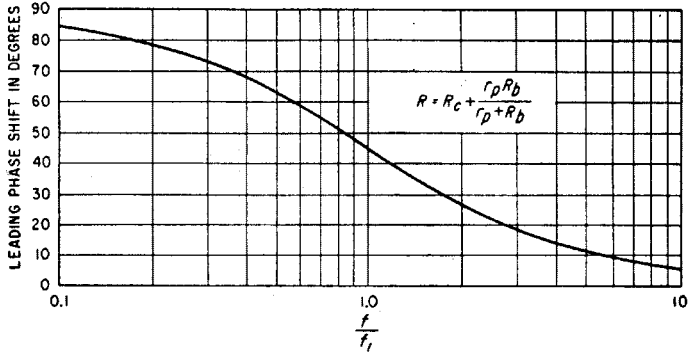
Fig. 3.20. Low-frequency phase characteristics due to R-C_c interstage coupling.

taken as the upper limit of bandwidth.   For an amplifier of the general form shown in Fig. 3.18 and using either triode or multielement tubes, f_2 will be given by Eq. (3.22).

$$f_2 = \frac{1}{2\pi C_t r_b'} \tag{3.22}$$

where C_t = total capacitance shunting load = $c_o + C_w' + C_w'' + c_i$
   c_o = output capacitance of amplifier tube
   C_w' = wiring capacitance on the plate side of the coupling capacitor
   C_w'' = wiring capacitance on the grid side of the coupling capacitor
   c_i = input capacitance of the following amplifier tube or terminal device

$$r_b' = \frac{1}{\frac{1}{r_p} + \frac{1}{R_b} + \frac{1}{R_c}} = \frac{r_p R_b R_c}{R_c R_b + r_p R_c + r_p R_b}$$

In wideband or video amplifiers, R_b will often be very much smaller than either r_p or R_c.   When this is true Eq. (3.22) can be simplified to Eq. (3.23).

$$f_2 = \frac{1}{2\pi C_t R_b} \qquad \text{when } R_b \ll r_p \text{ and } R_b \ll R_c \tag{3.23}$$

The high-frequency attenuation and phase characteristics are shown in curve 5 of Figs. 3.31 and 3.32, respectively.

Example 3.3
   Determine the gain and bandwidth of an amplifier similar to that shown in Fig. 3.18 and having the circuit constants listed below:

$g_m = 3,900 \ \mu$mhos	$C_w' = 4 \ \mu\mu$f
$r_p = 500,000$ ohms	$C_w'' = 6 \ \mu\mu$f
$R_b = 10,000$ ohms	$c_i = 29 \ \mu\mu$f
$R_c = 100,000$ ohms	$C_c = 1,000 \ \mu\mu$f
$c_o = 16 \ \mu\mu$f	

1. Solve for the amplifier gain by use of Eq. (3.15).

$$r_b = \frac{10,000 \times 100,000}{10,000 + 100,000} = 9,100 \text{ ohms}$$

$$A = \frac{-g_m r_p r_b}{r_p + r_b} = \frac{-0.0039 \times 500,000 \times 9,100}{500,000 + 9,100} = -34.9$$

2. Solve for f_1, with Eq. (3.20).

$$f_1 = \cfrac{1}{2 \times 3.14 \times 0.001 \times 10^{-6}\left(100,000 + \cfrac{500,000 \times 10,000}{500,000 + 10,000}\right)}$$

$$= 1,450 \text{ cycles}$$

3. Solve for f_2 with Eq. (3.22).

$$f_2 = \cfrac{1}{2 \times 3.14 \times 55 \times 10^{-12} \times 8,930}$$
$$= 324,000 \text{ cycles}$$

3.6. Gain and Bandwidth. Impedance-coupled Amplifiers. Impedance coupling in an amplifier is shown in Fig. 3.21. Inductor L_1 replaces the plate-load resistor R_b used in resistance-coupled amplifiers. An equivalent circuit is shown in Fig. 3.21b, where r_1 is the resistance of the inductor and C_d is the inductor distributed capacitance. In parallel with C_d are the output capacitance of the amplifier, c_o, the distributed wiring capacitance, C_w, and the input capacitance of the following amplifier, c_i. In

(a)   (b)
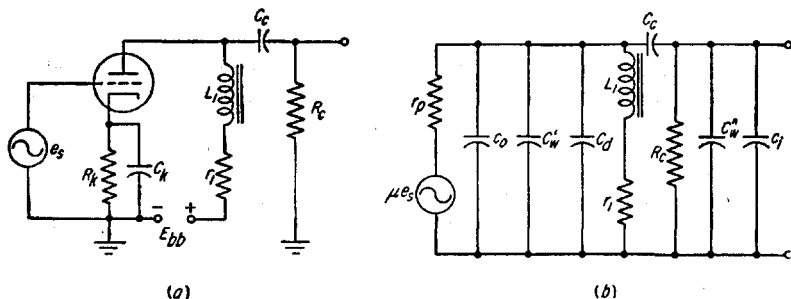
FIG. 3.21. Impedance-coupled amplifier and equivalent circuit.

the figure, C_w has been divided into two parts, C_w' and C_w'', which are shown located on either side of C_c.

Impedance coupling is used where it is desirable to have a high-load impedance with a small d-c voltage drop. It is possible to develop output signals with almost twice the amplitude as that from the same tube and power supply used in a resistance-coupled circuit. The impedance-coupled amplifier has many of the disadvantages of transformer coupling without the advantage of additional gain from step-up winding ratios. Some of these disadvantages are the possibility of magnetic saturation due to d-c current, variation of inductance with applied voltage, high distributed capacitance, high cost, and large size.

3.6a. Gain. The assumptions in computing the midband gain of an impedance-coupled amplifier are that the reactance of L_1 is very large compared to R_c in parallel with r_p and that the reactance of C_c is negligible compared to both R_c and r_p. When these assumptions are valid and there is no degeneration, the gain is given by Eq. (3.24) or (3.25).

$$A = \frac{-\mu R_c}{r_p + R_c} \qquad\qquad (3.24)$$
$$\simeq -\mu \qquad \text{when } R_c \gg r_p$$
$$A = \frac{-g_m R_c}{1 + R_c/r_p} \qquad\qquad (3.25)$$
$$\simeq -g_m R_c \qquad \text{when } r_p \gg R_c$$

The gain with cathode degeneration A' is given by Eq. (3.26) or (3.27).

$$A' = \frac{-\mu R_c}{r_p + R_c + R_k(\mu + 1)} \tag{3.26}$$

$$= \frac{-g_m R_c}{1 + g_m R_k + \dfrac{R_c + R_k}{r_p}} \tag{3.27}$$

At low frequencies where the reactance of L_1 may not be large enough to satisfy the assumption made for the equations above, the gain may be computed by substituting Z_b for R_c in Eqs. (3.24) to (3.27). This is valid when the reactance of C_c is negligible as compared to r_p and Z_b.

$$Z_b = R_c \sqrt{\frac{r_1{}^2 + \omega^2 L_1{}^2}{(r_1 + R_c)^2 + \omega^2 L_1{}^2}} \tag{3.28}$$

In the low-frequency case where the reactance of C_c becomes large compared to the sum of R_c and r_p and the shunting reactance of L_1 is still high compared to r_p and R_c, the low-frequency response can be computed as in a resistance-coupled amplifier.

$$f_1 = \frac{1}{2\pi C_c(R_c + r_p)} \tag{3.29}$$

where f_1 = low frequency at which the amplifier gain is 3 db less than at midband

The -3-db lower frequency, as limited by the shunting inductance only, will be the frequency at which ωL_1 is equal to R_c in parallel with r_p. This is based on the assumptions that the reactance of C_c is negligible compared to r_p and R_c and that r_1 is small compared to ωL_1.

The -3-db upper frequency is reached when the reactance of the sum of all the shunting capacitances, that is, $c_o + C_w + C_d + c_i$, is equal in magnitude to R_c in parallel with the output impedance of the tube which is equal to r_p in the case of no degeneration.

3.6b. Graphical Determination of the Quiescent and Dynamic Operating Conditions. The quiescent operating point for an impedance-coupled amplifier is fixed by the power-supply voltage and the resistors r_1 and R_k in the plate and cathode circuits (see Fig. 3.21a). The a-c plate load for midband frequencies is equal to R_c when the reactance of L_1 is high and is represented by a load line which normally can be assumed to very nearly pass through the quiescent operating point (see Sec. 3.3b). The procedure for the graphical solution is given in the following example.

Example 3.4

For the impedance-coupled amplifier shown in Fig. 3.21a and the circuit constants listed below, determine the quiescent operating condition, the maximum-amplitude sinusoidal output signal, the gain, and the nominal lower and upper limits of frequency response. Assume the tube characteristics to be those given in Fig. 3.22.

$$
\begin{aligned}
L_1 &= 30 \text{ henrys} & R_c &= 100 \text{ kilohms} \\
r_1 &= 450 \text{ ohms} & R_k &= 1 \text{ kilohm} \\
E_{bb} &= 250 \text{ volts} & C_d &= 50 \ \mu\mu\text{f} \\
C_k &= 50 \ \mu\text{f} & c_o &= 8 \ \mu\mu\text{f} \\
C_c &= 1.0 \ \mu\text{f} & C_w &= 20 \ \mu\mu\text{f} \\
& & c_i &= 10 \ \mu\mu\text{f}
\end{aligned}
$$

1. Construct the bias and d-c load lines (see Sec. 3.3).

The d-c plate-load line is plotted for the resistance $R_k + r_1$ and is shown in Fig. 3.22. The bias line is drawn for a resistance of 1,000 ohms. The intersection of the two lines at point O determines the quiescent operating conditions:

$$
\begin{aligned}
I_{bo} &= 7.75 \text{ ma} \\
E_{co} &= -7.75 \text{ volts} \\
E_{bo} &= 238 \text{ volts}
\end{aligned}
$$

2. Construct the a-c load line.

Assuming no distortion, this line will have a slope determined by R_c and will pass through point O. To construct this line, assume some excursion from the quiescent current and determine the corresponding voltage change. For a 1-ma change in current through the a-c load of 100 kilohms, a 100-volt change in plate voltage will occur. Therefore, at $I_b = 6.75$ ma, the corresponding value of E_b is 338 volts. A straight line through this point and point O will establish the a-c plate-load line.

3. Determine the maximum amplitude sinusoidal output signal and the midband gain.

The maximum excursion for a symmetrical signal from point O is first limited in the direction of reduced plate voltage along the a-c plate-load line. The limit is at the $E_c = 0$ curve where $E_b = 90$ volts. For a sinusoidal input signal, the grid excursion in the opposite direction must be identical, that is, -7.75 volts. Therefore the maximum positive excursion of plate voltage will be that plate voltage measured along the a-c load line for an instantaneous grid voltage of -15.5 volts. This value is approximately 382 volts. The
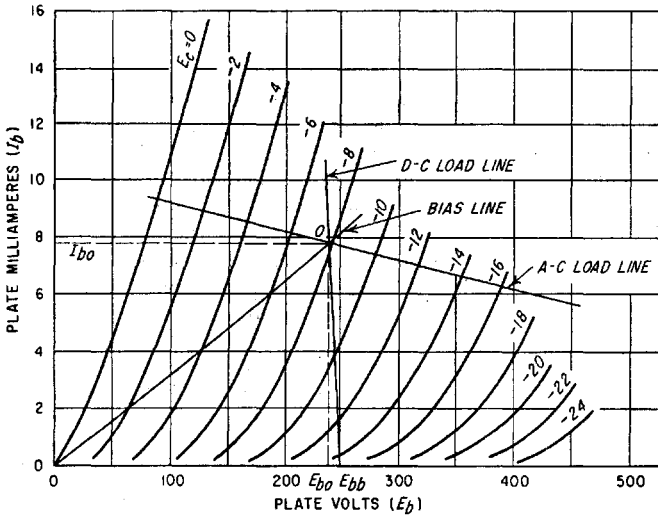
Fig. 3.22. Graphical solution for Example 3.4.

peak-to-peak output signal is equal to $382 - 90$, or 292 volts, and the gain for an output signal of this amplitude is given by

$$A = \frac{\Delta e_o}{\Delta e_s} = \frac{292}{15.5} = 18.8$$

4. Determine f_1, the -3-db lower frequency.

From the tube curves, $r_p = 7,500$ ohms. The value of the parallel combination of R_c and r_p is given by

$$\frac{R_c r_p}{R_c + r_p} = \frac{100,000 \times 7,500}{100,000 + 7,500} = 6,980 \text{ ohms}$$

Since r_1 is very small compared to 6,980 ohms it can be neglected in this solution. The frequency at which the reactance of L_1 equals 6,980 ohms is f_1. Therefore

$$f_1 = \frac{6,980}{2\pi L_1} = \frac{6,980}{2 \times 3.14 \times 30} = 37 \text{ cycles}$$

For this solution to be correct, the reactance of C_c at 37 cycles must be very small compared to $R_c + r_p$.

$$X_c = \frac{1}{2 \times 3.14 \times 37 \times 10^{-6}} = 4,300 \text{ ohms}$$

This value, less than 5 per cent of $R_c + r_p$, would have negligible effect on the value of f_1 which proves that the above solution is of reasonable accuracy. R_k is bypassed by a sufficiently large capacitor to prevent any appreciable reduction in gain at f_1 (see Sec. 3.10b).

5. Determine f_2, the -3-db upper frequency.

The sum of the capacitances shunting the load is

$$C_t = c_o + C_d + C_w + c_i = 8 + 50 + 20 + 10$$
$$= 88 \ \mu\mu\text{f}$$

where C_t = shunt capacitance

Therefore,

$$f_2 = \frac{1}{2\pi C_t \dfrac{r_p R_c}{R_c + r_p}}$$

$$= \frac{1}{2 \times 3.14 \times 88 \times 10^{-12} \times 6{,}980}$$

$$= 259 \text{ kc}$$

3.7. Gain and Bandwidth. Transformer-coupled Amplifiers.[1] Transformer coupling between stages of an amplifier or between an amplifier and load is usually used to obtain either a voltage gain, impedance transformation, d-c isolation, an ungrounded output, or a balanced (push-pull) connection. Transformers required to work over several octaves or more will have an iron-core and near-unity coupling. In applications requiring coverage of a small part of an octave the transformer will generally be of the air-core type with resonated windings. This section treats the use of iron-core transformer-coupled amplifiers only. For the treatment of air-core transformer-coupled amplifiers refer to Secs. 13.1 to 13.5 and for a discussion of the fundamentals of iron-core transformers refer to Sec. 14.2.
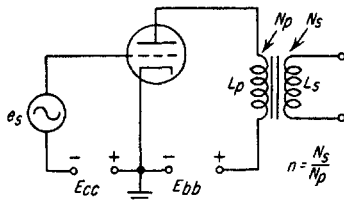
FIG. 3.23. Amplifier with unterminated transformer.

3.7a. Midband Gain. An amplifier employing transformer coupling without a resistive termination for the transformer is shown in Fig. 3.23. In such an amplifier the plate load is the reactance X_p of the transformer primary, that is, ωL_p, and at midband is generally very large compared to r_p. The gain at midband is therefore very nearly equal to μn where n is the ratio of secondary turns N_s to primary turns N_p.

Coupling transformers are more often resistively terminated, however, to improve the frequency response. The midband gain is lower than in the unterminated case and is given by Eq. (3.30) provided the winding resistances are assumed to be equal to zero.

$$A = \frac{\mu n R_L}{R_L + n^2 r_p} \tag{3.30}$$

where R_L = terminating resistance.

3.7b. Low-frequency Response. The -3-db lower frequency f_1 for an amplifier with an unterminated transformer is given by Eq. (3.31) provided the winding resistances are assumed to be equal to zero.

$$f_1 = \frac{r_p}{2\pi L_p} \qquad \text{(unterminated)} \tag{3.31}$$

where L_p is in henrys.

Equation (3.31) shows that the best low-frequency response is obtained when the dynamic plate resistance is small and the transformer primary inductance is large.

[1] See Sec. 4 for the graphical analyses of transformer-coupled amplifiers.

The complete low-frequency equivalent circuit is shown in Fig. 14.32, and in this figure R_G can be considered to be the plate output resistance.

Terminating the transformer with a resistance improves the low-frequency response. The frequency f_1 for the terminated transformer is given by Eq. (3.32).

$$f_1 = \frac{R_L r_p}{2\pi L_p(R_L + n^2 r_p)} \qquad \text{(terminated)} \tag{3.32}$$

In many cases it is necessary to state the low-frequency response as a function of signal level. This is due to the change of the core permeability and, consequently, L_p with changes in the level of the applied voltage.[1]

The resistive shunt-feed circuit shown in Fig. 3.24a is sometimes used to eliminate the d-c magnetization effect of the plate current on the transformer. The gain and
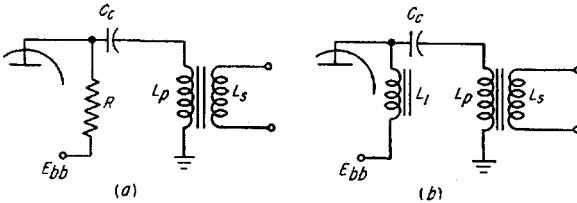
FIG. 3.24. Shunt-feed circuits for transformers.

low frequency limit f_1 of this amplifier can be computed from Eqs. (3.33) to (3.36) provided the winding resistance are assumed to be equal to zero.

$$A = \frac{\mu R n}{r_p + R} \qquad \text{(unterminated)} \tag{3.33}$$

$$A = \frac{\mu n R R_L}{R R_L + r_p R_L + n^2 R r_p} \qquad \text{(terminated)} \tag{3.34}$$

$$f_1 = \frac{r_p R}{2\pi L_p(r_p + R)} \qquad \text{(unterminated)} \tag{3.35}$$

$$f_1 = \frac{r_p R R_L}{2\pi L_p(n^2 r_p R + r_p R_L + R R_L)} \qquad \text{(terminated)} \tag{3.36}$$

A variation of the shunt feed is shown in Fig. 3.24b. In this circuit an inductor L_1 is used for a d-c current path. The midband gain of this circuit is the same as the gain with direct coupling to the transformer. With the assumption that the coupling capacitor C_c has negligible reactance, the low-frequency response will be determined by L_1 in parallel with L_p. This equivalent value of inductance, $(L_1 L_p)/(L_1 + L_p)$, can be substituted for L_p in Eq. (3.31) or (3.32).

Shunt-feed circuits are also used with coupling capacitor values that resonate with L_p within the useful passband. If C_c is adjusted for resonance at an optimum frequency somewhat below the normal lower-frequency limit f_1, the flat response of the amplifier can be extended for almost an octave. If the frequency of resonance is adjusted slightly above the normal lower-frequency limit f_1, a certain amount of low-frequency signal boost will result. The amount of signal boost obtained is a function of the primary winding Q. Normally, the primary winding Q of an iron core transformer will be small, i.e., about unity or less, at low frequencies unless there is a gap in the core (see Sec. 14.1i).

3.7c. High-frequency Response. There is no simple equation which can be used to solve for the high-frequency response (see Fig. 14.31c). In general, it will have a pronounced hump (a limited region of increased gain) if the transformer is not termi-

[1] For a discussion of this effect see Sec. 14.4.

nated or if terminated with an impedance much higher than the transformer output impedance. The appearance of the hump is contingent upon the source impedance, e.g., the dynamic tube resistance being small as compared to the transformer leakage reactance at f_r. The resonant frequency f_r is defined as the frequency at which the total leakage reactance L_p' is equal to the reactance of the sum of the primary and secondary circuit capacities in the 1:1 equivalent transformer. In Figs. 14.33 to 14.41 of Sec. 14.5, the high-frequency responses of terminated transformers having different circuit values have been plotted. It should be noted that it is necessary to first convert the actual transformer circuit into the equivalent circuit of a transformer having a 1:1 turns ratio. Figure 3.25 is applicable if the transformer is unterminated
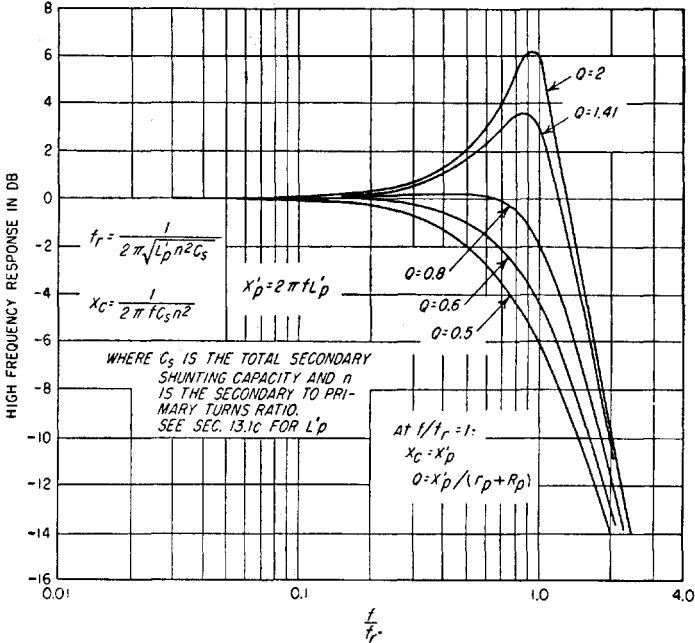
FIG. 3.25. High-frequency response of an unterminated transformer for the condition where $n^2 C_s \gg C_p$.

and if, in the 1:1 equivalent circuit, the secondary circuit capacitance is much larger than the primary circuit capacitance. In a step-up transformer this is usually the case, and the primary circuit capacitance can be neglected in the calculations.

Example 3.5

In a circuit similar to Fig. 3.23 and with constants listed below, determine the midband gain, the −3-db lower frequency f_1, f_r, and the −3-db upper frequency f_2.

$$\mu = 20 \qquad L_p' = 0.04 \text{ henry}$$
$$r_p = 7{,}000 \text{ ohms} \qquad C_p = 100 \ \mu\mu f$$
$$L_p = 12 \text{ henrys} \qquad C_s = 120 \ \mu\mu f$$
$$n = 3 \qquad R_p = \text{pri. resist.} = 300 \text{ ohms}$$
$$R_s = \text{sec. resist.} = 1{,}500 \text{ ohms}$$

1. Determine the midband gain.

$$A = \mu n = 20 \times 3 = 60$$

2. Determine f_1.
From Eq. (3.31)

$$f_1 = \frac{7{,}000}{2 \times 3.14 \times 12} = 93 \text{ cycles}$$

3. Determine f_r (see Fig. 3.25).

$$f_r = \frac{1}{2 \times 3.14 \sqrt{0.04 \times 3^2 \times 120 \times 10^{-12}}}$$
$$= 24,200 \text{ cycles}$$

4. Compute the circuit Q (see Fig. 3.25).

$$Q = \frac{X_p'}{r_p + R_p} = \frac{2 \times 3.14 \times 24,200 \times 0.04}{7,000 + 300} = 0.833$$

5. Determine the high frequency at which the response is 3 db less than at midband.

Reading from the $Q = 0.8$ curve of Fig. 3.25, the value of f/f_r corresponding to −3 db is approximately 1.1. Therefore, the frequency of −3-db response equals $1.1 \times 24,200$, or 26,600 cycles.

3.8. Grid-input Impedance of an Amplifier. If an a-c signal source is connected between the grid of an amplifier and ground, current will flow. The value of the applied signal voltage divided by this current is the grid circuit input impedance to the amplifier. The phase angle and the magnitude of the current indicate the values of the resistive and reactive components.

The factors contributing to the grid-input impedance are treated in the following paragraphs. Resistances in the grid circuit external to the tube are not considered.

3.8a. Steady-state Grid Current Loading with Negative Bias. The quiescent grid current, independent of signal, is an important design consideration whenever an amplifier is directly coupled to a very high impedance generator or storage device. The current that flows varies with electrode voltages, plate current, cathode temperature, residual gas in the tube envelope, and other less important factors. This current can vary greatly from tube to tube of the same type and manufacture, but some general rules are applicable. Grid current can be either positive or negative in direction of flow.

Positive grid current is defined as electron flow from the cathode to the grid within the tube. This current flows whenever electrons are emitted from the cathode with an initial velocity great enough to overcome the retarding field and reach the grid. Cathode temperature and cathode material determine the distribution of initial velocities. Low heater voltage, high negative grid bias, and high plate voltage (screen grid voltage in pentodes) tend to reduce the positive grid current for any given tube.

Negative grid current—defined as the flow of electrons from the grid into the plate current stream—is usually smaller but more erratic than positive grid current in voltage amplifier tubes. The primary reason for this current is the ionization of the gas in the tube by collision with high-velocity plate stream electrons. Some of the positive gas ions so formed flow to the negative grid where their charges are completed by those electrons supplied to the grid through the external grid circuit. This electron flow through the external grid circuit resistance causes a change in bias in the direction of zero bias which in turn causes an increase in plate current and results in an increase in the rate of ionization. Because this effect is regenerative, it is necessary to minimize the grid circuit resistance or employ cathode biasing if gassy tubes are used. This effect is most pronounced in power tubes because of the fact that the high operating temperatures cause the liberation of occluded gases. In any tube ionization can be minimized by operating at reduced plate and screen voltages and currents. In amplifier applications, a minor cause of negative grid current is secondary emission from the grid. This effect has been reduced by the selection of special metal alloys or by gold-plating the grid. Generally, in most small-voltage amplifier tubes, the negative grid current will be considerably less than one microampere.

Each of the two opposing current components becomes predominant in different regions of the applied grid voltage. At low values of bias, the positive current is the greater, and at high values the negative current is the greater. The intermediate point between these regions is a value of bias for which there is no net grid current. This value of bias can be established for a given operating condition by determining the potential of the "floating" grid. This is done by measuring the plate current with the grid floating and then duplicating this current while operating the grid from a variable-voltage source and observing the required bias. For high-μ triodes which are operated with moderate to high plate voltages, the zero grid current point will usually be in the region of -1 to -2 volts of bias.

3.8b. Effect of Positive Grid Signals on Input Resistance. A positive grid-cathode voltage E_{gk} causes grid current to flow and consequently represents a load on the grid signal source. For most receiving-type tubes with a positive grid signal, the value of R_{gk} is between 500 and 1,500 ohms. These values are representative only when the positive grid-cathode voltage is greater than a few tenths of a volt and when the screen or plate voltage is high compared to E_{gk}. As E_{gk} becomes very small, the value of R_{gk} becomes large, and when the grid-cathode voltage is zero or slightly negative, R_{gk} reaches a value which can usually be considered as being equal to one or more megohms. When E_{gk} approaches or exceeds the screen voltage in pentodes or the plate voltage in triodes, the minimum values of R_{gk} will be reached.

3.8c. Effect of Interelectrode Capacitance on Input Impedance. The phase angle between the grid signal voltage and the current that flows through the interelectrode capacitances is dependent on the magnitude and phase angle of the plate load impedance.

1. Resistively Loaded Amplifiers. In a resistively loaded amplifier capacitive currents will flow from the grid to the other electrodes through the interelectrode capacitances. If a signal voltage exists at any electrode other than the grid, the instantaneous grid input current due to capacitance between the grid and this electrode will be equal to the instantaneous difference in the two voltages divided by the reactance of the mutual capacitance. Multiple current paths are additive in determining the total input current.

An amplifier with a resistive load will have an effective input capacitance larger than the sum of c_{gk}, control-grid to screen-grid capacitance c_{g1g2}, and c_{gp}. This effect is sometimes referred to as "Miller effect." The input capacitance c_i of a resistively loaded amplifier with a gain A is given by Eq. (3.37).

$$c_i = c_{gk} + c_{g1g2} + c_{gp}(|A| + 1) \tag{3.37}$$

Equation (3.37) applies to amplifiers in which no signal exists at either the cathode or screen grid. In the case of an amplifier with a degenerative cathode, c_i is given by Eq. (3.38).

$$c_i = c_{gk}(1 - A_{gk}) + c_{g1g2} + c_{gp}(|A'| + 1) \tag{3.38}$$

where A' = grid-plate gain with cathode degeneration [see Eqs. (3.16) and (3.17)]

A_{gk} = grid-cathode gain [see Eq. (3.114)]

The superiority of the pentode is evident because the very small value of c_{gp} minimizes both the total input capacitance and the variations in input capacitance with variations in the grid-plate gain. If the screen circuit impedance is not bypassed, Eqs. (3.37) and (3.38) will apply if the term c_{g1g2} is modified by the factor $(1 + |A_{g1g2}|)$. The term A_{g1g2} is the voltage gain from the control grid to the screen grid.

In the case of the cathode follower where both the plate and screen grid are at signal ground, c_i is given by Eq. (3.39).

$$c_i = c_{gk}(1 - A_{gk}) + c_{g1g2} + c_{gp} \tag{3.39}$$

Example 3.6

Assume that a triode amplifier has a gain of 12 and has the circuit capacitances given below.   If the plate load is resistive and the cathode is bypassed at the signal frequency, what is the input capacitance?

$$c_{gp} = 4 \; \mu\mu f \qquad c_{gk} = 3 \; \mu\mu f$$

From Eq. (3.37)

$$c_i = 3 + 4(12 + 1) = 55 \; \mu\mu f$$

2. Reactively Loaded Amplifier.   An amplifier that operates into a reactive plate load will have a phase angle ϕ associated with the gain term, that is, A/ϕ.   If the plate load of an amplifier is capacitive, the input impedance has a positive resistance component.   Similarly, an inductive plate load produces a negative resistance component at the input to the amplifier.   A positive resistance in the input-impedance term indicates that the signal source will be required to supply power.   A negative-resistance component indicates that the amplifier will supply power to the external grid circuit.   The latter case is the basis of many oscillator circuits.

The input impedance, due to the interelectrode capacitances and reactive load, can be reduced to a resistive component R_x and a capacitive component c_i.

$$R_x = \frac{-1}{\omega c_{gp}|A| \sin \phi} \tag{3.40}$$

$$c_i = c_{gk} + c_{g1g2} + c_{gp}(1 + |A| \cos \phi) \tag{3.41}$$

where $\omega = 2\pi f$

ϕ = phase angle of output voltage caused by reactive load, that is, $\phi = \phi_1 - \phi_2$ where ϕ_1 is the phase angle of the load and ϕ_2 is the phase angle of the load impedance plus r_p

Equations (3.40) and (3.41) are valid only if there is neither cathode nor screen-grid degeneration.

Neutralization of tuned triode amplifiers is necessary at the frequency the load becomes inductive and feeds back enough power to the grid to cause oscillation. Neutralization consists of providing another path from the output to the input through which current can flow to compensate for the interelectrode current.

Example 3.7

Find the grid-input resistance and capacitance to an amplifier utilizing a triode tube and which has the following circuit constants:

$$\mu = 20 \qquad c_{gk} = 3 \; \mu\mu f$$
$$r_p = 8{,}000 \text{ ohms} \qquad c_{gp} = 4 \; \mu\mu f$$

Frequency = 2 Mc
Plate load = 12,000 ohms in series with 2.5 mh
1. Find X_L.

$$X_L = 2\pi f L = 6.28 \times 2 \times 10^6 \times 2.5 \times 10^{-3}$$
$$= 31{,}400 \text{ ohms}$$

2. Find Z_L.

$$Z_L = R_L + jX_L = 12{,}000 + j31{,}400$$
$$= 33{,}600/69.1° \text{ ohms}$$

3. Compute A.

$$A = -\frac{\mu Z_L}{r_p + Z_L} = -\frac{20 \times 33{,}600/69.1°}{8{,}000 + (12{,}000 + j31{,}400)}$$
$$= -18.1/11.6°$$

4. Compute R_x from Eq. (3.40).

$$R_x = \frac{-1}{6.28 \times 2 \times 10^6 \times 4 \times 10^{-12} \times 18.1 \times 0.201}$$
$$= -5{,}460 \text{ ohms}$$

5. Compute c_i from Eq. (3.41).

$$c_i = 3 + 4[1 + (18.1 \times 0.98)] = 78 \ \mu\mu f$$

3.8d. Effect of Transit Time on Input Impedance. In a vacuum tube the electron flow between the cathode and plate, which is initiated by the grid signal voltage, causes a current to flow in the external grid circuit. Electrons approaching the grid plane from the cathode cause electrons to flow from the grid to the cathode through the external grid circuit, and electrons passing from the grid plane to the plate cause an electron flow from the cathode to the grid through the external grid circuit. At low and moderate frequencies, the induced grid current appears as a predominantly capacitive load on the grid signal source. As the signal frequency is increased, the transit time, expressed in degrees of the applied signal, becomes larger and larger. This in turn causes the impedance loading the signal source to approach a pure resistance and to have a small enough value that it may severely load the input signal source. It has been shown by Llewellyn[1] that the grid input resistance due to transit time varies inversely as the square of frequency. An expression for input resistance due to electron transit time is given by Eq. (3.42).

$$R_t = \frac{1}{K g_m f^2 T^2} \tag{3.42}$$

where R_t = tube input resistance in ohms due to transit time

    K = a constant which is a function of cathode-grid and grid-plate transit times

    g_m = grid-plate transconductance, mhos

    f = frequency, cps

    T = transit time from cathode to grid, sec

For conventional voltage amplifier tubes, transit time loading usually becomes apparent in the region of 10 Mc and above. The degree of input loading becomes very pronounced for most of these tubes in the region of 50 Mc and above. Specially constructed tubes such as the lighthouse tube, which has closely spaced planar electrodes, have been designed to permit operation as high as 3,000 Mc and above.

Some typical values of input resistances at 100 Mc are given in Fig. 7.34. These values include the effects of cathode lead inductance (see Sec. 3.8f).

3.8e. Effect of Transconductance on Input Capacitance. In addition to the increase in grid-cathode capacitance due to electron transit time, c_{gk} is increased as the transconductance is increased because of the movement of the space charge toward the grid.

The incremental change Δc in the grid-cathode capacitance for a hot and cold tube represents the combined effects of transit time and transconductance. It is possible to introduce complete compensation for this effect by placing an unbypassed resistor R_k in the cathode circuit whose value is given by Eq. (3.43).

$$R_k = \frac{\Delta c I_b}{g_m c_{gk} I_k} \quad \text{ohms (tetrode or pentode)} \tag{3.43}$$

$$= \frac{\Delta c}{g_m c_{gk}} \quad \text{ohms (triode)}$$

where Δ_c = cold-to-hot increase in the tube grid-cathode capacitance, $\mu\mu f$

    g_m = grid-plate transconductance at the operating point, mhos

    c_{gk} = grid-cathode capacitance of the cold tube, $\mu\mu f$

    I_b = plate current

    I_k = cathode current

[1] F. B. Llewellyn, "Electron Inertia Effects," Cambridge University Press, New York, 1941.

3.8f. Effect of Cathode Lead Inductance on Input Capacitance. At high frequencies, the plate-signal current flowing through the cathode lead inductance may cause a significant voltage to exist at the cathode which is leading the grid-signal voltage. Because of the phase relationship between the applied grid voltage and the cathode voltage, the current which flows through the grid-cathode capacitance has a component which is in phase with the grid voltage. The value of the input resistance due to cathode lead inductance is given by Eq. (3.44).

$$R_y = \frac{1}{g_m \omega^2 L_k c_{gk}} \tag{3.44}$$

where R_y = tube input resistance in ohms due to cathode-lead inductance
   g_m = grid-plate transconductance, mhos
   $\omega = 2\pi f$
   L_k = cathode lead inductance, henrys
   c_{gk} = grid-cathode capacitance, farads

The value of cathode lead inductance for most miniature receiving-type tubes is between 0.01 and 0.02 μh. Equation (3.44) is based on the reactance of the cathode lead inductance being small as compared to the reactance of the grid-cathode capacitance. Compensation can be made for the effects of cathode lead inductance as described in Sec. 7.4h.

3.9. Gain-bandwidth Product, Pulse Rise Time, and Tube Selection for Wideband Amplifiers. The significant considerations in the choice of vacuum tubes for wideband amplifiers and the equations for gain-bandwidth product, pulse rise time, and figures of merit for comparing tubes are given in this section.

3.9a. Gain-bandwidth Product of a Resistively Loaded Amplifier. The gain of a resistively loaded amplifier stage can be determined from Eq. (3.14). At the −3-db upper frequency f_2, the circuit resistance as measured at the plate of the tube, that is, $r_p r_b / (r_p + r_b)$, is equal to the reactance of the shunt capacitance C_t. From these two relationships the absolute value of the midband gain $|A|$ can be expressed by Eq. (3.45).

$$|A| = \frac{g_m}{2\pi f_2 C_t} \tag{3.45}$$

C_t includes the stray wiring capacitances, the output capacitance of the amplifier tube, and the input capacitance to the following tube. From Eq. (3.45) it can be seen that the product of $|A|$ and f_2, that is, the *gain-bandwidth* product, is dependent only upon the tube constants and the circuit capacitances. This is based on the assumption that f_2 is very much larger than f_1.

$$\text{Gain} \times \text{bandwidth} = |A|f_2 = \frac{g_m}{2\pi C_t} \tag{3.46}$$

The gain-bandwidth product indicates that for an increase or decrease in gain there is a corresponding decrease or increase in bandwidth. If the bandwidth is increased until it equals the gain-bandwidth product, unity gain will be realized. The gain-bandwidth product can be increased, however, by video compensation in the form of series and/or shunt peaking. Feedback methods, including cathode compensation, will not increase this product.

3.9b. Figures of Merit for Tetrode and Pentode Tubes. To permit the comparative evaluation of tubes in wideband amplifiers, independent of wiring techniques, the *figure of merit* term F_{WA} is established from Eq. (3.47).

$$F_{WA} = \frac{g_m}{c_i + c_o} \tag{3.47}$$

where F_{WA} = figure of merit for a wideband amplifier tube

c_i = input capacitance to the tube, $\mu\mu$f

c_o = output capacitance of the tube, $\mu\mu$f

g_m = grid-to-plate transconductance, μmhos

Note that F_{WA} is evaluated with the input and output capacitances of the same tube and is therefore not readily converted to a gain-bandwidth product except where two or more identical tubes are used successively.[1] The figure of merit values for several tubes are given in Table 3.1.

TABLE 3.1. FIGURES OF MERIT FOR SEVERAL VACUUM TUBES

Tube type	g_m, μmho	c_i, $\mu\mu$f	c_o, $\mu\mu$f	F_{WA}
6AG5	5,000	6.5	1.8	603
6AH6	9,000	10	3.6*	661
6AK5	5,000	4	2.8*	735
6CB6	6,200	6.3	1.9	756
6AU6	5,000	5.5	5	476
807	6,000	12	7	316
829	8,500	14.5	7	395

* With shield.

Where wide bandwidth and a large output voltage are required simultaneously, different tube considerations apply. To develop a large voltage across a low-impedance load it is necessary to have large plate current excursions. For output signals that have equal positive and negative values, a high d-c plate current must flow. This current must be equal to or greater than the peak value of the output signal divided by r_b. In the case of a plate-loaded amplifier, negative output pulses only can be obtained without high steady-state plate current. For positive pulses, the use of a highly biased cathode follower will eliminate the need for a high quiescent current.

Maximum plate current rating and low output capacitance are usually the prime considerations in the choice of a tube for a large output voltage with wide bandwidth. This is based on the assumption that g_m and c_i are of less importance because the input signal can usually be supplied from some appropriate generator. When this is true, the figure of merit F_O for an output tube is given by Eq. (3.48).

$$F_O = \frac{I_p}{c_o} \qquad (3.48)$$

where F_O = figure of merit for output tube

I_p = maximum possible value of average plate current for desired plate load

For a given tube type, the value of F_O is not constant but is a function of the plate load. It permits, however, the comparison of different types of tubes. The practical considerations in supplying the drive signal for such a stage usually require that g_m and c_i enter into the tube selection. The final choice for this case will usually be a compromise between F_{WA} and F_O.

3.9c. Pulse Rise Time as a Function of Bandwidth. In some cases, pulses or other transient waveforms are to be amplified rather than sine waves. For these applications, it is usually more appropriate to consider the maximum rate of rise of output voltage that can be obtained rather than the frequency response to a sine-wave

---

[1] Since c_i in a triode is a function of stage gain, no figure of merit can be established for a triode.

input. Rise time is defined as the time required for the output-voltage waveform to rise from 10 to 90 per cent of its final value in response to a step-function input signal.[1] The following relationship for uncompensated resistively loaded amplifiers will give the rise time as a function of the circuit parameters.

$$\text{Rise time, sec} \simeq 2.2 r_b C_t \simeq \frac{2.2 A C_t}{g_m} \tag{3.49}$$

Equation (3.50) can be used to express the relationship between bandwidth and rise time for a single uncompensated stage.

$$\text{Rise time, sec} \simeq \frac{0.35}{\text{bandwidth, cps}} \tag{3.50}$$

To express the over-all rise time of a multistage amplifier, the composite rise time to a step function input is given by the square root of the sum of the squares of the individual rise times. This relationship assumes that the individual amplifiers do not have an overshoot that exceeds a few per cent. For input signals with rise times appreciably greater than that of the amplifier, the amplifier may be considered to have negligible waveform distortion of the leading edge. In the case where the input signal rise time approaches that of the amplifier for a step function input, the output signal will have a rise time equal to the square root of the sum of the squares of the input and amplifier rise times.

Example 3.8

In an uncompensated video amplifier using cascaded 6AK5 tubes, determine the maximum midband gain per stage and the plate-load resistances if the desired stage bandwidth is 4.25 Mc and the distributed wiring capacitance C_w is 6 $\mu\mu f$ per stage.

From Table 3.1

$$C_t = c_o + c_i + C_w = 2.8 + 4 + 6 = 12.8 \ \mu\mu f$$

From Eq. (3.45)

$$|A| = \frac{5 \times 10^{-3}}{2 \times 3.14 \times 4.25 \times 10^6 \times 12.8 \times 10^{-12}} = 14.6$$

In this problem the grid-to-plate capacitance of the succeeding tube (0.02 $\mu\mu f$) multiplied by ($|A| + 1$) is a negligible quantity and can be disregarded. In some cases it will be necessary to add this term to the other capacitances. To determine the value of the a-c plate load r_b consistent with the calculated value of A, the simple expression for midband gain can be used.

$$r_b \simeq \frac{|A|}{g_m} = \frac{14.6}{5 \times 10^{-3}} = 2,920 \text{ ohms} \quad \text{(assuming no degeneration)}$$

Example 3.9

An oscilloscope deflection amplifier must be capable of supplying a 110-volt rms sine wave to the deflection plates. Using an 807 tube, determine the operating conditions for 2-Mc bandwidth if the sum of the deflection plate input and wiring capacitances is 20 $\mu\mu f$.

From the tube manufacturer's data:

$$c_o = 7 \ \mu\mu f$$

The a-c plate-load resistance must be equal to or less than

$$r_b = \frac{1}{2\pi f_2 C_t}$$

$$= \frac{1}{2 \times 3.14 \times 2 \times 10^6 \times (20 + 7) \times 10^{-12}}$$

$$= 2,950 \text{ ohms}$$

---

[1] With reference to the input signal, a step function is an instantaneous change in the d-c value to some new d-c value.

and $\qquad e_o(\text{peak}) = \sqrt{2}\, e_{\text{rms}} = 1.414 \times 110 = 155.6$ volts

$$\text{Quiescent plate current } I_{bo} \geq \frac{155.6}{2,950} = 52.8 \text{ ma}$$

and $\qquad\qquad I_{b(\text{max})} \geq I_{bo} + \dfrac{155.6}{2,950} = 105.6 \text{ ma}$

If the 807 screen voltage is assumed to be equal to 250 volts, the quiescent plate voltage equal to 300 volts, and the quiescent control grid voltage equal to -12 volts, g_m will be approximately 6,000 μmhos. Under these conditions the quiescent plate current will be approximately 100 ma and the tube will not be driven into the positive grid region. Therefore,

$$|A| \simeq g_m r_b = 6 \times 10^{-3} \times 2,950$$
$$\simeq 17.7$$

3.10. Effect of Bypassed Cathode and Bypassed Screen-grid Resistors on the Frequency Response. In determining the low-frequency response, it is necessary to consider the effects of the interstage coupling as discussed in Secs. 3.5c, 3.6a, and 3.7b in addition to the effects of cathode and screen-grid circuit bypassing which are discussed in the following sections. Both the db loss and phase shift due to interstage coupling add to the db loss and phase shift due to the cathode and/or screen-grid bypassing. An amplifier employing screen-grid and cathode bypassing is shown in Fig. 3.26.
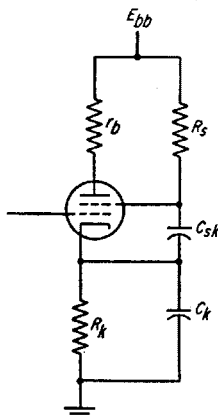

FIG. 3.26. Amplifier with both screen-grid and cathode bypassing.

3.10a. Low-frequency Gain and Phase Characteristics as Determined by Bypassing the Screen Grid to a Grounded Cathode. At high, medium, and relatively low frequencies, the screen-grid circuit which employs a dropping resistor can be stabilized with respect to the cathode by bypassing the screen grid to the cathode. At very low frequencies, however, the screen-grid circuit becomes degenerative since the reactance of the bypass capacitor becomes sufficiently large that it has negligible effect on the circuit. In Figs. 3.27 and 3.28 are the amplitude and phase characteristics of an amplifier as determined by screen-grid circuit bypassing as a function of f/f_s for several values of A_s/A_s' provided that the cathode is either grounded or very well bypassed at those frequencies in the region of f_s. The terms f_s and A_s/A_s' are determined from Eqs. (3.51) and (3.52).

$$f_s = \frac{1}{2\pi R_s C_{sk}} \qquad\qquad (3.51)$$

$$\frac{A_s}{A_s'} = 1 + \frac{R_s}{r_{g2}} \qquad\qquad (3.52)$$

where $r_{g2} =$ dynamic screen-grid resistance, ohms

$R_s =$ screen voltage dropping resistor

$C_{sk} =$ bypass capacitor between screen grid and cathode, farads

$f_s =$ frequency in cycles per second at which reactance of $C_{sk} = R_s$

$A_s =$ gain without screen circuit degeneration

$A_s' =$ gain with screen circuit degeneration

It is of interest to note that the total loss in gain due to the screen circuit degeneration at low frequencies is dependent only on the value of R_s/r_{g2}.

The frequency f_1, that is, the low frequency at which the amplifier gain is 3 db less than at midband, can be found as follows: (1) compute A_s/A_s', (2) from Fig. 3.27
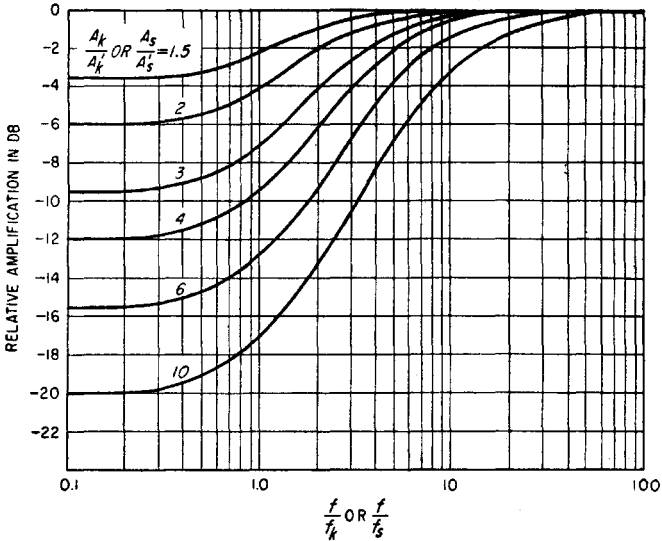
FIG. 3.27. Amplitude characteristics of an amplifier with a bypassed cathode resistor or a screen-grid resistor bypassed to the cathode.
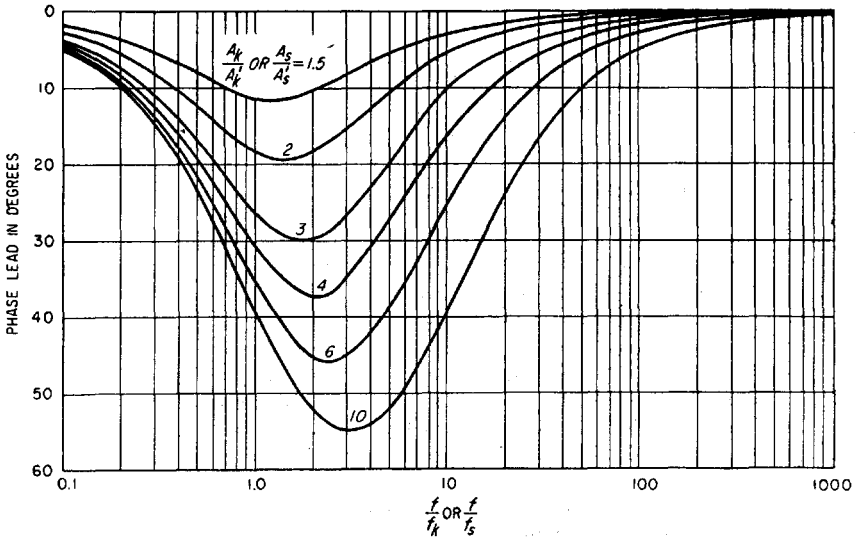

FIG. 3.28. Phase characteristics of an amplifier with a bypassed cathode resistor or a screen-grid resistor bypassed to the cathode.

determine the value of f/f_s which causes a 3-db loss, and (3) determine the value of f_1 from Eq. (3.53).

$$f_1 = \text{(value from step 2)} \times f_s \qquad (3.53)$$

3.10b. Low-frequency Gain and Phase Characteristics as Determined by a Bypassed Cathode Resistor. With the application of Eqs. (3.54) and (3.55) and Figs. 3.27 and 3.28, the low-frequency amplitude and phase characteristics of a triode can be deter-

mined. These same equations and figures apply to a tetrode or pentode amplifier provided that the screen-grid bypass capacitor to the cathode is large enough that a very small proportion of the a-c screen current flows through R_s and R_k at frequencies in the region of f_k. The terms f_k and A_k/A_k' are determined from Eqs. (3.54) and (3.55).

$$f_k = \frac{l}{2\pi R_k C_k} \tag{3.54}$$

$$\frac{A_k}{A_k'} = 1 + \frac{R_k(\mu + 1)}{r_p + r_b} \tag{3.55}$$

and if $r_p \gg r_b$,

$$\frac{A_k}{A_k'} \simeq 1 + g_m R_k \tag{3.56}$$

where R_k = cathode resistance, ohms

C_k = cathode bypass capacitor, farads

A_k = gain without cathode degeneration

A_k' = gain with cathode degeneration

f_k = frequency in cycles per second at which reactance of $C_k = R_k$

The value of f_1 is determined in a manner similar to the one described in Sec. 3.10a, except that A_k/A_k', f/f_k, and f_k are substituted for A_s/A_s', f/f_s, and f_s, respectively.

3.10c. Low-frequency Gain and Phase Characteristics as Determined by the Combination of Screen-grid and Cathode Bypassing. Equation (3.19) is applicable for the determination of gain if both cathode and screen-grid degeneration exist. The applicable equation for the case of the cathode being bypassed to ground and the screen circuit being bypassed to the cathode is considerably more complicated. This is partially caused by the fact that the screen-grid bypass capacitor is in parallel with the series combination of R_s and R_k and prevents the substitution of the term Z_{sk} (C_{sk} in parallel with the sum of R_s and R_k) for R_s in a manner similar to the possible substitution of the term Z_k (parallel combination of R_k and C_k) for R_k in Eq. (3.19). Because of the form of this new equation, it does not lend itself to usable simplification and consequently prevents the creation of a convenient and accurate means for determining the low-frequency amplitude and phase characteristics due to the combination of screen-grid and cathode bypassing.

A practical means of making a fairly good approximation as to the amplitude and phase characteristics is to assume that the attenuation in decibels and the phase shift in degrees can each be determined for screen-grid and cathode bypassing separately and then the results added so as to obtain the composite amplitude and phase characteristics. This method, although not exact, permits the determination of a response curve which usually does not deviate from the true response curve by more than about one decibel.

3.10d. Extension of the High-frequency Response by Cathode and/or Screen-grid Circuit Compensation. The high-frequency response of an amplifier which is degenerative, because of either an unbypassed cathode and/or an unbypassed screen-grid resistor, can be extended by reducing this degeneration at those upper frequencies where the plate load is reduced by the reactance of the shunt capacity. This can be accomplished by bypassing the appropriate resistor or resistors, viz., either the screen-grid resistor to the cathode and/or the cathode resistor to ground, with a small capacitor which is effective only at high frequencies. This permits the extension of the upper -3-db frequency by a factor equal to or larger than A/A' as shown in Fig. 3.29. The factor A/A' is equal to A_k/A_k' if only cathode circuit compensation is used, A_s/A_s' if only screen-grid compensation is used, and $A_k/A_k' + A_s/A_s'$ if both

cathode and screen-grid compensation are employed [see Eqs. (3.52) and (3.55)]. Note that this permits the realization of the same gain-bandwidth product that could be obtained by effectively bypassing the applicable resistor or resistors at all frequencies. The same gain-bandwidth product is obtained, however, with reduced gain and increased bandwidth.
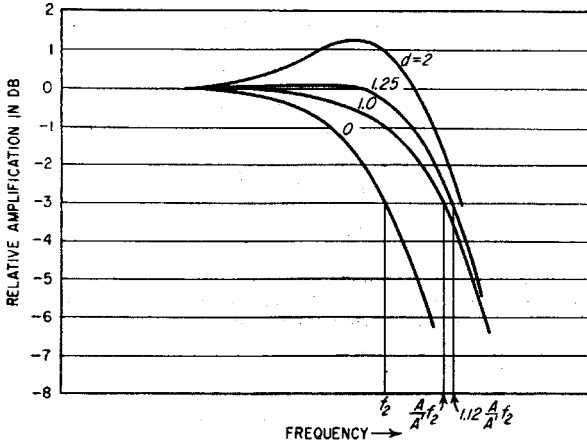

FIG. 3.29. High-frequency response of an amplifier as modified by cathode or screen-grid circuit compensation.

If only the cathode resistor is unbypassed, the high-frequency compensation which can be realized by bypassing the cathode resistor with a small capacitor can be determined from Fig. 3.29. The value of d should be determined from Eq. (3.57). In Fig. 3.29, the factor A/A' is equal to A_k/A_k' [see Eqs. (3.55) and (3.56)].

$$d = \frac{R_k C_k}{r_b C_t} \tag{3.57}$$

Similarly, if only the screen-grid resistor is unbypassed, A/A' in Fig. 3.29 is equal to A_s/A_s' [see Eq. (3.52)] and d should be determined from Eq. (3.58).

$$d = \frac{R_s C_{sk}}{r_b C_t} \tag{3.58}$$

If both the cathode and screen-grid resistors are unbypassed, the curves shown in Fig. 3.29 can also be utilized to establish the factor by which the upper -3-db frequency can be extended by properly bypassing these resistors. For this case, A/A' is equal to $A_k/A_k' + A_s/A_s'$. To obtain correct compensation, either Eqs. (3.57) and (3.59) or Eqs. (3.58) and (3.60) must be satisfied.

$$f_s = \frac{A_k f_k}{A_k'} \tag{3.59}$$

$$f_k = \frac{A_s f_s}{A_s'} \tag{3.60}$$

The curves shown in Fig. 3.29 show the high-frequency response of an amplifier as a function of frequency for four different values of d. The optimum value for d

is usually taken as 1 since with this value the bandwidth is increased by the factor A/A' without overcompensation. Where a slight amount of overcompensation is permissible, the value of d equal to 1.25 will cause the bandwidth to be increased by the factor $1.12A/A'$ as shown in the figure. Severe overcompensation results if the value of d is much greater than 1.25.

Example 3.10

Determine f_1, the -3-db lower frequency, for an amplifier which has the following circuit values:

$$\mu = 20 \qquad\qquad R_k = 2{,}000 \text{ ohms}$$
$$r_p = 10{,}000 \text{ ohms} \qquad r_b = 25{,}000 \text{ ohms}$$
$$C_k = 10 \ \mu\text{f}$$

1. Compute A_k/A_k'.

From Eq. (3.55)

$$\frac{A_k}{A_k'} = 1 + \frac{2{,}000(20 + 1)}{10{,}000 + 25{,}000}$$
$$= 2.2$$

2. Determine the value of f/f_k which causes a 3-db loss (see Fig. 3.27).

By interpolation between the curves for A_k/A_k' equal to 2 and 3, the value of f/f_k which causes a loss of 3 db is seen to be approximately equal to 1.7.

3. Determine f_1.

Determine f_k from Eq. (3.54).

$$f_k = \frac{1}{2 \times 3.14 \times 2{,}000 \times 10 \times 10^{-6}}$$
$$= 7.96 \text{ cycles}$$

and from Eq. (3.53)

$$f_1 = 1.7 \times 7.96$$
$$= 13.53 \text{ cycles}$$

NOTE: To determine the over-all low-frequency response of an amplifier it is also necessary to add the db loss and phase shift introduced by the interstage coupling as discussed in Secs. 3.5c, 3.6a, and 3.7b.

Example 3.11

A pentode amplifier with r_b of 2,000 ohms and an unbypassed R_k of 240 ohms is operating with a transconductance g_m of 5,200 μmhos. If the uncompensated bandwidth is 4 Mc, determine C_k for the greatest bandwidth without overcompensation.

At the upper -3-db frequency, the reactance of C_t shunting the load is equal to r_b. Therefore

$$C_t = \frac{1}{2\pi f_2 r_b} = \frac{1}{2 \times 3.14 \times 4 \times 10^6 \times 2 \times 10^3}$$
$$= 19.9 \ \mu\mu\text{f}$$

For maximum bandwidth without overcompensation $d = 1$, therefore

$$1 = \frac{R_k C_k}{r_b C_t}$$
$$C_k = \frac{2 \times 10^3 \times 19.9 \times 10^{-12}}{240}$$
$$= 166 \ \mu\mu\text{f}$$

The factor by which the bandwidth is increased is equal to A_k/A_k', therefore from Eq. (3.56)

$$\frac{A_k}{A_k'} \simeq 1 + g_m R_k = 2.25$$

$$\text{Actual bandwidth with compensation} \simeq 2.25 \times 4$$
$$\simeq 9 \text{ Mc}$$

3.11. High-frequency Compensation. In this section, several different types of compensating networks and their associated amplitude and phase characteristics are discussed. For each type of compensation it has been assumed that the a-c plate resistance r_p of the tube is much larger than the plate-load impedance. The phase-shift characteristics for several different types of compensating networks have been normalized so as to indicate the degree of phase distortion (see Figs. 3.32, 3.35, 3.37, 3.40, and 3.42). A network introduces phase distortion if it produces phase shift which is not directly proportional to frequency. Normalized phase-shift curves for networks introducing no phase distortion are horizontal lines. The vertical positions of these lines are determined entirely by the amount of phase shift at any given frequency whereas the fact that they are horizontal lines indicates that the phase shift is directly proportional to frequency. Note that it is the actual magnitude of the deviation from a uniform normalized phase-shift value rather than the percentage deviation which determines the magnitude of the phase distortion. The phase shift of a sine wave can be converted into an equivalent time delay as described on the referenced figures.

A repetitive signal having a complex waveform will be distorted if there is unequal amplitude response and/or phase distortion to the frequency components which make up the waveform.

In an amplifier having no phase distortion but having a nonuniform amplitude response as a function of frequency, the output waveform to a square wave input will be symmetrical. However, the top of the output waveform will not be flat. If the top of the output waveform tends to be convex, the high-frequency response is inadequate, and if the top of the output waveform is concave, the low-frequency response is inadequate.

Phase distortion, with or without a uniform amplitude response as a function of frequency, will cause a square wave to become nonsymmetrical. If the top of the output waveform has an upward slope, the low-frequency components are delayed more than the high-frequency components. This is caused by either insufficient phase lead at low frequencies or insufficient phase lag at high frequencies. If the top of the output waveform has a downward slope, the high-frequency components are delayed more than the low-frequency components. This is caused by either excessive phase lead at low frequencies or excessive phase lag at high frequencies.

A compensating network which provides critical damping will have a normalized phase-shift characteristic which decreases with increasing frequency. With reference to a step function input, critical damping will cause the output signal to have the minimum rise time without overshoot (see Fig. 18.4). In general, critical damping is not considered in video amplifiers since the rise time is usually considered to be excessive. In the multistage amplifier in which a certain per cent overshoot to a step function input is desired, the permissible overshoot per stage is inversely proportional to the number of stages.

For single input functions such as a step or for repetitive complex waveforms, an amplifier which has been compensated to optimize its phase characteristics, i.e., to minimize the deviations of the normalized phase shift from a constant value, is frequently considered to provide the best compromise in the reproduction of the referenced input functions.

If a single sine wave of varying frequency is to be amplified, the compensation which provides the best amplitude characteristics is usually used.

3.11a. Shunt Compensation. The conventional resistance-coupled amplifier without cathode or plate-circuit compensation will provide a specific value for the upper −3-db frequency f_2 for a given a-c plate-load resistance. Equation (3.61) shows that f_2 is the frequency at which the reactance of the shunt capacitance C_t is equal to r_b.

At this frequency, the impedance of the load is equal to $0.707r_b$ and continues to decrease with increasing frequency.

$$f_2 = \frac{1}{2\pi r_b C_t} \tag{3.61}$$

where f_2 = upper frequency in cycles per second at which resistance-coupled amplifier gain is 3 db less than at midband without compensation

C_t = total shunt capacitance in farads across plate-load resistor (includes input capacitance to following stage)

r_b = a-c plate-load resistance, ohms

If an inductor having the proper value is added in series with the plate-load resistor as shown in Fig. 3.30, a low-Q resonant circuit will be obtained in the region of f_2. The parallel resonant circuit so formed will present an increased plate-load impedance to the tube for a limited range of frequencies, thereby causing the amplifier upper -3-db frequency to be extended beyond f_2. The value of the inductor is given by Eq. (3.62).

$$L_b = k_b r_b^2 C_t \tag{3.62}$$

where L_b is in henrys, k_b is an arbitrary constant, r_b is in ohms, and C_t is in farads.

In Figs. 3.31 and 3.32 are the amplitude and normalized phase-shift characteristics of a shunt-compensated amplifier having several different values of compensation.

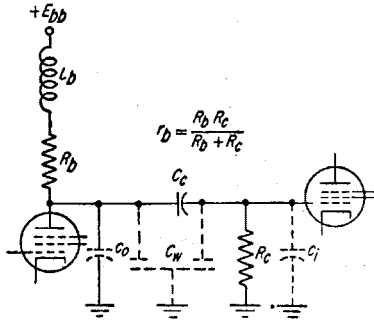The best phase and amplitude characteristics are obtained when k_b is approximately equal to 0.35 and 0.50, respectively. The maximum value that k_b can have without overcompensation is 0.44. Critical damping occurs when $k_b = 0.25$.

FIG. 3.30. Output circuit of a shunt-compensated amplifier.

3.11b. Series Compensation. The upper -3-db frequency limit can be extended beyond f_2 [see Eq. (3.61)] by the use of a series compensating circuit of the type shown in Fig. 3.33. The values of L_c and R_d can be determined by Eqs. (3.63) and (3.64), respectively.

$$L_c = k_c r_b^2 (C_1 + C_2) \tag{3.63}$$

where L_c is in henrys, k_c is an arbitrary constant, $C_1 = c_o +$ (wiring capacitance on plate side of L_c), and $C_2 = c_i +$ (wiring capacitance on grid side of L_c) (both C_1 and C_2 are in farads).

$$R_d = k_d r_b \tag{3.64}$$

where R_d and r_b are in ohms and k_d is an arbitrary constant.

The ratio of C_2 to C_1 is given by Eq. (3.65).

$$a = \frac{C_2}{C_1} \tag{3.65}$$

Amplitude and normalized phase-shift characteristics are plotted in Figs. 3.34 to 3.37 for various combinations of a, k_c, and k_d. Most vacuum tubes have an input capacitance which is approximately equal to twice the output capacitance, and, consequently, the value of a is frequently very nearly equal to 2.

Capacitor C_c can be located on either the plate or grid side of L_c so as to add its distributed capacitance to the side which will be most beneficial in obtaining the desired ratio C_2/C_1.
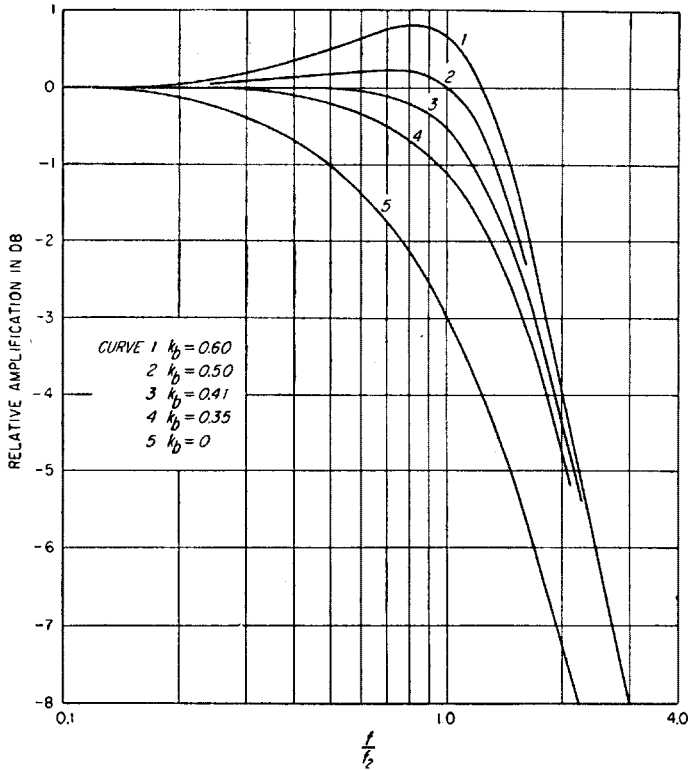
FIG. 3.31. Amplitude characteristics of a shunt-compensated amplifier as a function f/f_2 for several values of k_b.
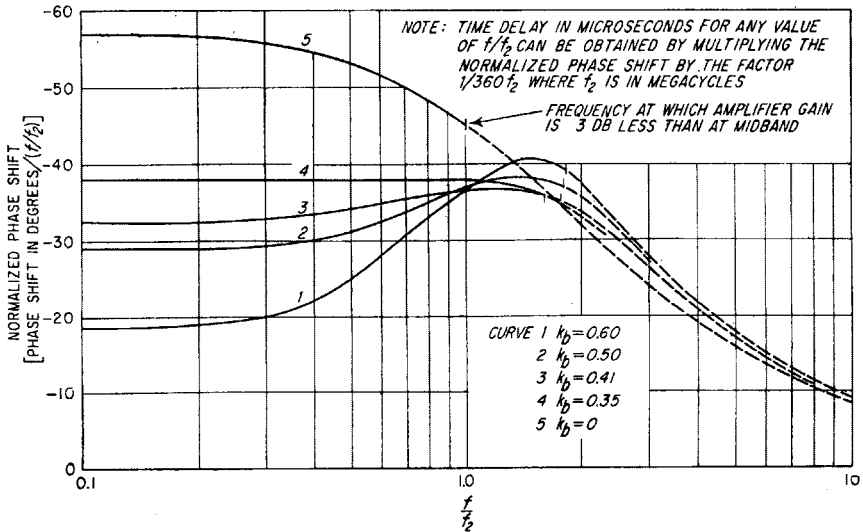

FIG. 3.32. Normalized phase shift in a shunt-compensated amplifier as a function of f/f_2 for several values of k_b.

Fig. 3.35. Normalized phase-shift characteristics of an output circuit of the type shown in Fig. 3.33 when $a = 1$.

Fig. 3.36. Amplitude characteristics of an output circuit of the type shown in Fig. 3.33 when $a = 2$.

to Fig. 16.11, $C_m = C_n = C/2$ and $2f_2 = f_c$. The normalized phase-shift characteristics are shown in Fig. 3.46.

The normalized phase-shift characteristics can be made fairly constant when the cutoff frequency f_c for the filter is made approximately equal to $2.44f_2$. In this case the calculated value of C from Fig. 16.11 will be less than C_t and the amplifier gain

will not be constant up to the frequency of cutoff. The -3-db frequency will be approximately equal to $2.3f_2$.

3.11e. Low-pass Filter as a Three-terminal Compensating Network. The three-terminal compensating network shown in Fig. 3.45 can also be employed to extend the frequency response. This type of compensation takes advantage of the fact that

CURVE 1 $k_c = 0.75$, $k_d = \infty$

2 $k_c = 0.67$, $k_d = \infty$

3 $k_c = 0.50$, $k_d = \infty$

4 $k_c = 0.58$, $k_d = 3.8$

FREQUENCY AT WHICH AMPLIFIER GAIN IS 3DB LESS THAN AT MIDBAND

NOTE: TIME DELAY IN MICROSECONDS FOR ANY VALUE OF f/f_2 CAN BE OBTAINED BY MULTI-PLYING THE NORMALIZED PHASE SHIFT BY THE FACTOR $1/360f_2$ WHERE f_2 IS IN MEGACYCLES

FIG. 3.37. Normalized phase-shift characteristics of an output circuit of the type shown in Fig. 3.33 when $a = 2$.

the output capacitance C_t is composed of two terms, C_1 and C_2 (do not confuse with C and C in the figure), which can be used as circuit elements in a constant-k low-pass filter. If C_1 is equal to C_2, the frequency response can theoretically be made absolutely constant up to $4f_2$. If C_2 is equal to $2C_1$, the upper limit is $3f_2$. In Fig. 3.45, the value of C must be determined from Fig. 16.11. If $C_1 = C_2$, the value of f_c is equal to $4f_2$ and $C = C_1 = C_2$. If $C_2 = 2C_1$, f_c is equal to $3f_2$ and

$$C = C_2 = 2C_1.$$

$R_c \gg R_b$

THEREFORE $r_b \approx R_b$
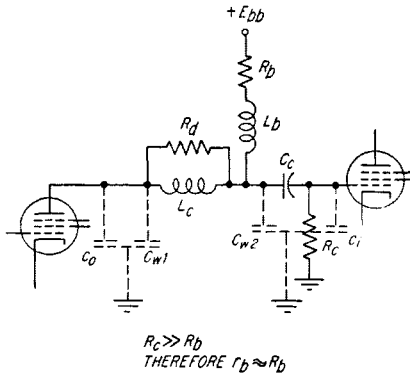
FIG. 3.38. Output circuit of an amplifier using both series and shunt compensation.

In this latter case, it is necessary to parallel C_1 with another capacitance of the same value. The normalized phase-shift characteristics are shown in Fig. 3.46.

3.11f. High-frequency Boost in Audio Amplifiers. In an audio amplifier, there is not the severe requirement for obtaining as exact compensation as in a video amplifier. The amplifier gain can deviate above or below the midband value by several decibels, and in addition there can be considerable deviation from a linear phase characteristic without causing the signal to be unpleasant to the human ear. Since many persons prefer high-frequency components to be emphasized, most audio amplifiers are designed with the provision for high-frequency boost.

High-frequency boost can be obtained by employing special equalizing networks of the type discussed in Sec. 17.2 or by employing impedance elements as plate loads
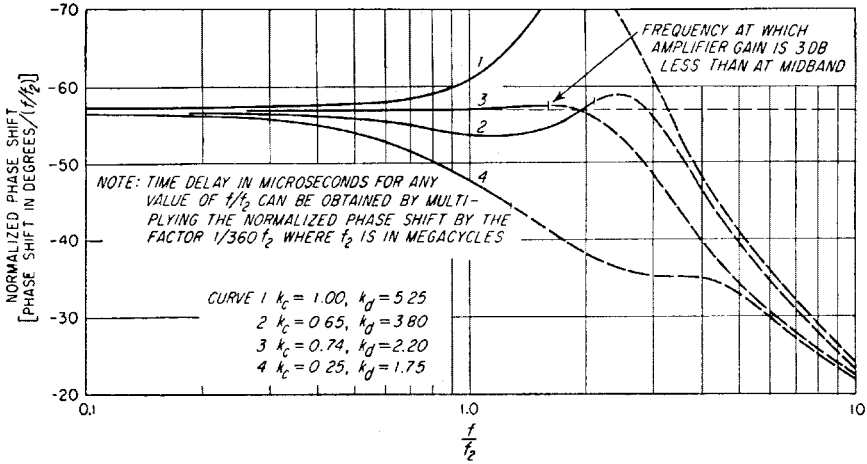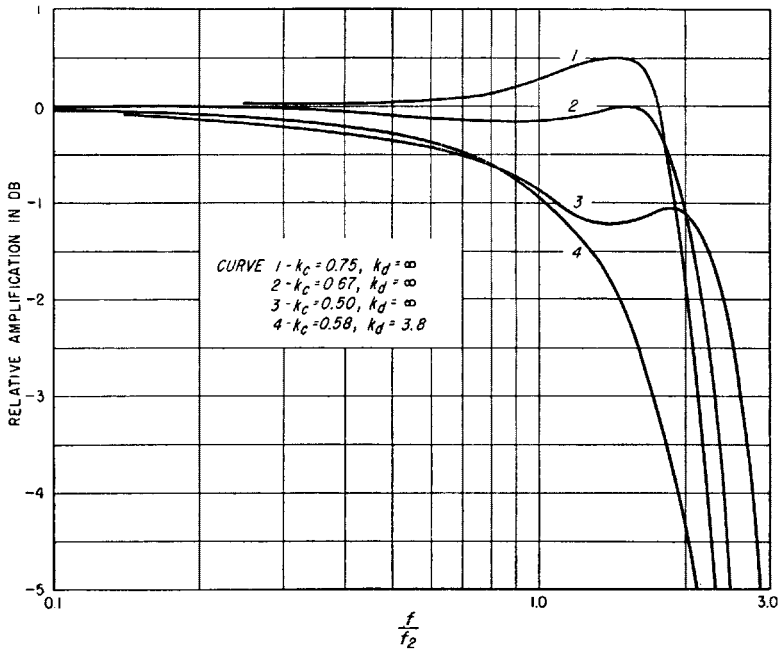
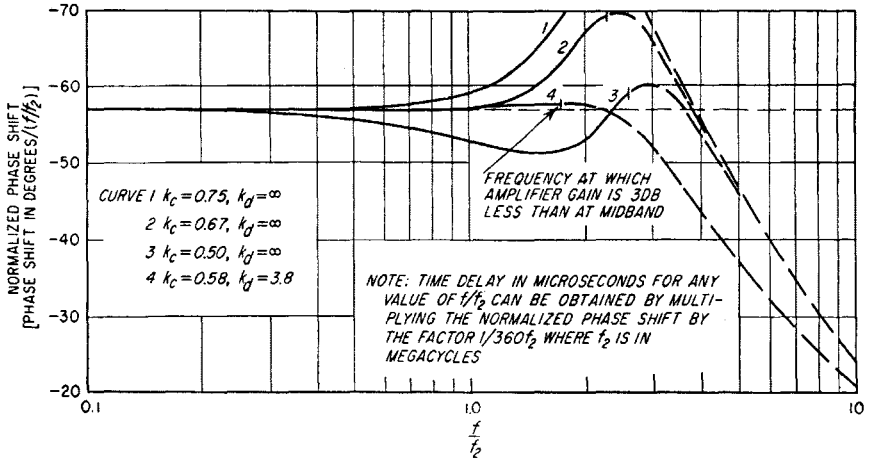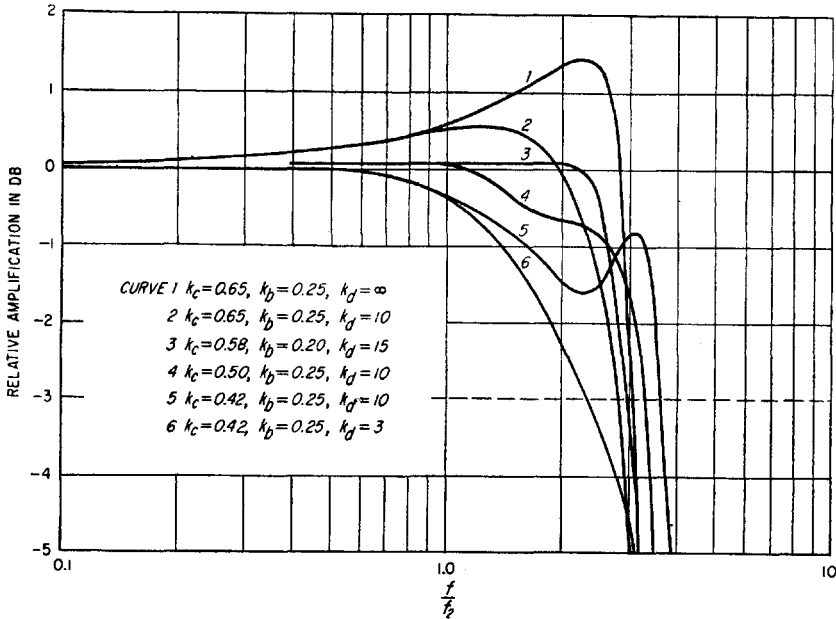FIG. 3.39. Amplitude characteristics of an output circuit of the type shown in Fig. 3.38 when $a = 1$.
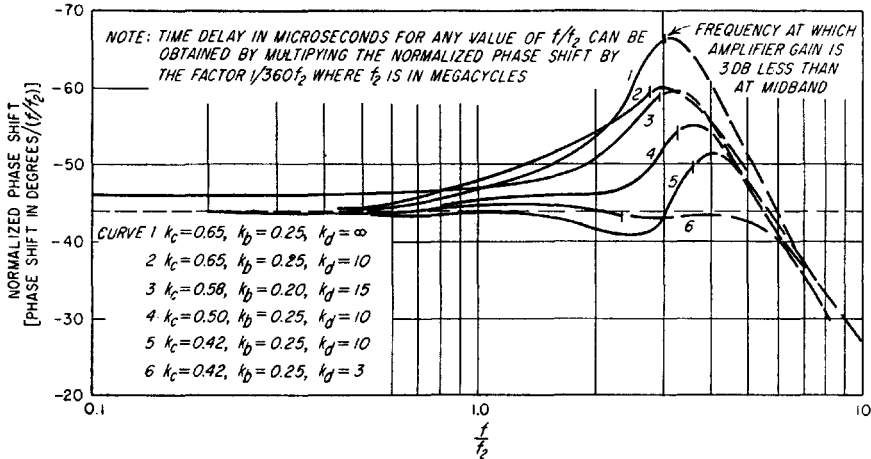
FIG. 3.40. Normalized phase-shift characteristics of an output circuit of the type shown in Fig. 3.38 when $a = 1$.

whose values increase in the high-frequency region where the boost is desired. An example is a parallel R, L, and C network in series with the plate-load resistance. The resonant frequency for L and C determines the region in which the boost will be obtained, and the value of R, the inductance Q, the plate-load resistance, and the tube characteristics determine the amount of boost.

High-frequency boost can also be obtained by bypassing the cathode and screen-grid circuits with capacitors whose values are such that they are ineffective at low
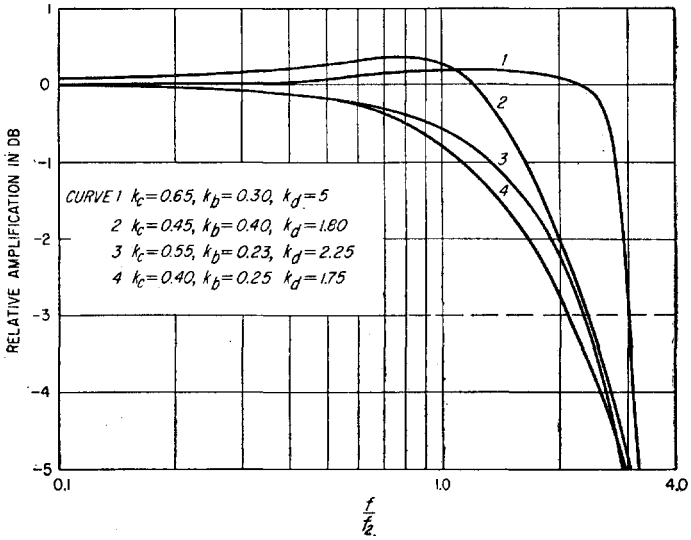
Fig. 3.41. Amplitude characteristics of an output circuit of the type shown in Fig. 3.38 when $a = 2$.
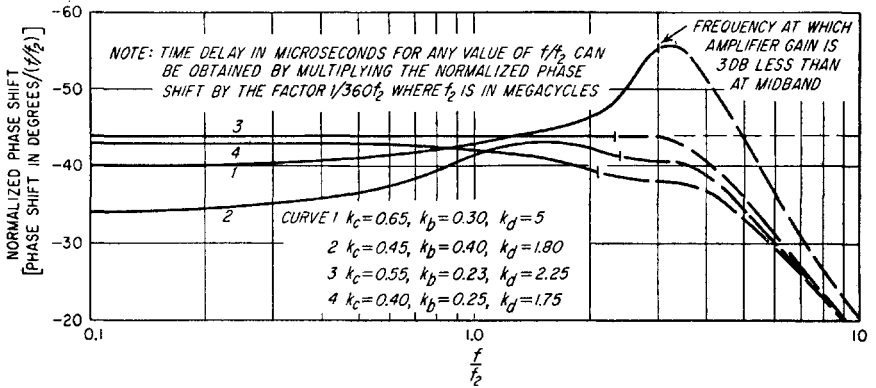
Fig. 3.42. Normalized phase-shift characteristics of an output circuit of the type shown in Fig. 3.38 when $a = 2$.

(a) PHASE DISTORTION CAUSED BY THE COMPENSATION GIVEN BY CURVE 3 OF FIGS. 3.39 AND 3.40

(b) DISTORTION CAUSED BY THE COMPENSATION GIVEN BY CURVE 6 OF FIGS. 3.39 AND 3.40 (DISTORTION IS DUE PRIMARILY TO THE NONUNIFORM AMPLITUDE RESPONSE AS A FUNCTION OF FREQUENCY)

Fig. 3.43. Examples of the distortion to a square wave at frequency f_2 in passing through two different output circuits. Small dip in (b) is due to reduced amplitude of 5th harmonic.

frequencies but which provide the desired increase in amplifier gain at high frequencies. Boost will be realized if either the product $R_k C_k$ or $R_s C_{sk}$ is greater than $r_b C_t$ (see Fig. 3.29). The term r_b is the a-c plate-load resistance, and C_t is the shunt capacitance across r_b.

Another common means of obtaining treble boost is to employ negative feedback in which the feedback loop has a transfer function which decreases with increasing
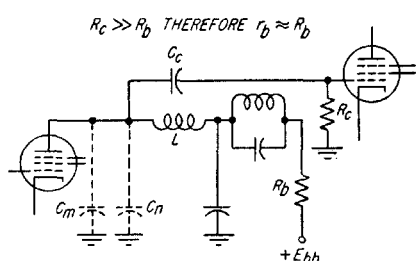
FIG. 3.44. Output circuit of an amplifier employing a low-pass filter for compensation. The shunt capacitance C_t is represented by C_m and C_n.
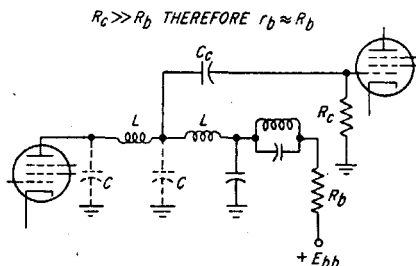
FIG. 3.45. Output circuit of an amplifier employing a low-pass filter for compensation.

frequency. The reduced feedback causes an associated increase in gain at high frequencies.

Example 3.12

Determine the value of L_b for a shunt-compensated amplifier employing a tube having a 1,500-ohm load and an output capacitance of 5 $\mu\mu$f. Let k_b be equal to 0.5. C_w is 8 $\mu\mu$f, and the following tube has an input capacitance of 13 $\mu\mu$f. Therefore

$$C_t = c_o + c_i + C_w$$
$$= 5 + 13 + 8$$
$$= 26 \ \mu\mu f$$

and from Eq. (3.62)

$$L_b = 0.5 \times 1,500^2 \times 26 \times 10^{-12}$$
$$= 29.3 \times 10^{-6} \text{ henry, or } 29.3 \ \mu h$$

Example 3.13

For the amplifier in Example 3.12, determine (1) f_2, the −3-db frequency without compensation and (2) f_2', the −3-db frequency with compensation.

From Eq. (3.61)

$$f_2 = \frac{1}{2 \times 3.14 \times 1,500 \times 26 \times 10^{-12}}$$
$$= 4.08 \text{ Mc}$$

FIG. 3.46. Normalized phase shift in output circuits shown in Figs. 3.44 and 3.45.

The gain-bandwidth improvement factor is determined from **Fig. 3.31** and is found to be equal to 1.8. Therefore
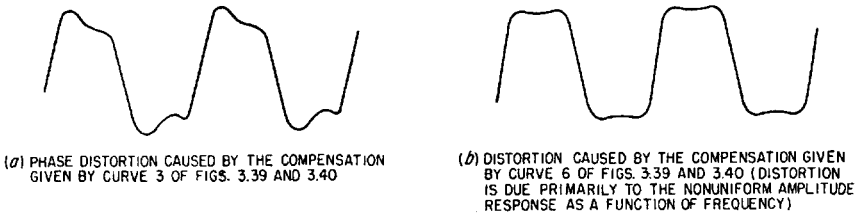
$$f_2' = 1.8 \times 4.08$$
$$= 7.34 \text{ Mc}$$

3.12. Low-frequency Compensation. The low-frequency response of an amplifier is usually established by one or more RC network combinations such as a cathode resistor and its bypassing capacitor or an RC interstage coupling network. In addition to the frequency-response considerations, caution should be exercised in using

RC networks in plate, screen, cathode, or interstage circuits if the signal being amplified has a d-c component. A unipolar signal is an example of a signal having a d-c component since the average value of the signal is not equal to zero. If a video amplifier having a plate-circuit decoupling network is used to amplify large positive unipolar input signals of long duration, the average plate current will increase considerably above that value caused by smaller signals of shorter duration. For this reason, the voltage drop across the decoupling network will not be constant; consequently, the effective supply voltage for the tube will not be constant and will result in a shift of the operating point and may cause a change in the amplifier gain. This same type of d-c shift will occur in cathode circuit self-biasing networks and screen-circuit decoupling netw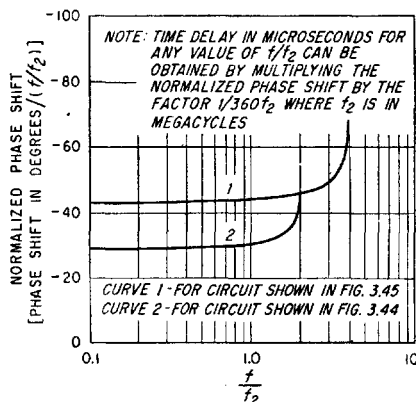orks. In video amplifiers which must be decoupled from the supply voltages and which are used to amplify nonsymmetrical signals having varying amplitudes, the recommended procedure is to minimize the size of the screen and plate decoupling resistors and maximize the size of the decoupling capacitors. In many instances it is necessary to incorporate special regulated supplies for the plate and screen circuits to achieve both the desired isolation and the desired stability in the supply. If possible, the bias should be obtained by returning the grid circuit to a biasing network. If self-biasing must be used, avoid bypassing the cathode resistor if the change in average tube current caused by a typical change in signal level will shift the operating point and noticeably affect the gain or limit the size of the signal being amplified. If an RC interstage coupling network is used, it is possible to employ a clamping network to avoid an appreciable shift in the operating point in the following stage provided only unipolar signals are being amplified. In the absence of a clamping network, the shift in the operating point will be equal to the d-c component in the signal. Typical interstage networks for clamping the signal to a given level are shown in Fig. 12.13.
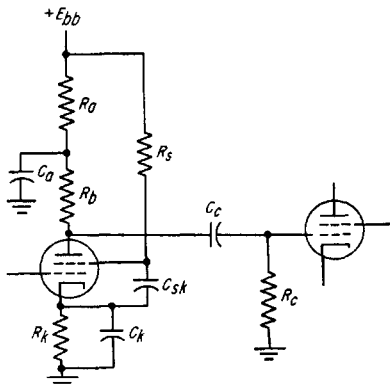
FIG. 3.47. Output circuit of an amplifier with the low-frequency compensating network consisting of R_a and C_a.

The design information contained in Secs. 3.12a to 3.12e is intended for class A amplifiers with sine-wave input signals; therefore the problems associated with changes in the level of nonsymmetrical signals do not exist.

3.12a. Low-frequency Compensating Network for an Amplifier. The low-frequency amplitude and phase characteristics of an amplifier as modified by the components R_a and C_a shown in Fig. 3.47 can be determined with the aid of Figs. 3.27 and 3.28 provided that the dynamic plate resistance r_p is much larger than $R_b + R_a$. The decibel scale on Fig. 3.27 should be read as decibel gain rather than loss, and the phase angle as determined from Fig. 3.28 should be read as lagging instead of leading. The individual curves are then for different values of A_a/A'_a, and the abscissas are plotted in terms of f/f_a. The terms f_a and A_a/A'_a can be determined from Eqs. (3.66) and (3.67), respectively.

$$f_a = \frac{1}{2\pi R_a C_a} \qquad (3.66)$$

where R_a is in ohms, C_a is in farads, and f_a is the frequency in cycles per second at which the reactance of C_a is equal to R_a. If r_p is much larger than $R_b + R_a$, the value

of A_a/A_a' to be used when referring to Figs. 3.27 and 3.28 can be determined from Eq. (3.67).

$$\frac{A_a}{A_a'} \simeq 1 + \frac{R_a}{R_b} \tag{3.67}$$

For all other cases, it is not possible to employ Figs. 3.27 and 3.28, and the amplifier gain and phase characteristics as modified by R_a and C_a must be determined experimentally or from Eq. (3.68).

$$A(f) = \frac{\mu[R_b + R_a/(1 + jf/f_a)]}{r_p + R_b + R_a/(1 + jf/f_a)} \tag{3.68}$$

3.12b. Low-frequency Compensation of a Bypassed Cathode and/or a Bypassed Screen-grid Resistor. If r_p is much larger than $R_a + R_b$, it is possible, with the proper choice of values for R_a and C_a, to exactly compensate for both the loss in gain and the leading phase shift caused by a bypassed cathode resistor and/or a bypassed screen-grid resistor. Equations (3.69) and (3.70) express the relationships required for exact compensation of both the amplitude and phase distortion caused by a bypassed cathode resistor.

$$R_a = g_m R_b R_k \tag{3.69}$$
$$R_a C_a = R_k C_k \tag{3.70}$$

Exact compensation for both amplitude and phase distortion caused by bypassing a screen-grid resistor to the cathode can be obtained by satisfying Eqs. (3.71) and (3.72).

$$R_a = \frac{R_b R_s}{r_{g2}} \tag{3.71}$$
$$R_a C_a = R_s C_{sk} \tag{3.72}$$

To compensate simultaneously for both a bypassed cathode resistor and a screen-grid resistor bypassed to the cathode, it is necessary that f_a, f_k, and f_s have a specific relationship to one another. These relationships are obtained by satisfying either Eqs. (3.70), (3.73), and (3.75) or Eqs. (3.72), (3.74), and (3.75).

$$f_s = \frac{A_k f_k}{A_k'} \tag{3.73}$$

$$f_k = \frac{A_s f_s}{A_s'} \tag{3.74}$$

$$\frac{A_a}{A_a'} = \frac{A_k}{A_k'} + \frac{A_s}{A_s'} \tag{3.75}$$

3.12c. Low-frequency Compensation for the Effects of Interstage RC Coupling. If the value of R_a were infinite and if r_p were large compared to R_b and R_c, it would be possible to provide perfect low-frequency compensation for the effects of RC interstage coupling. In practice, where this type of compensation is desired, R_a is made as large as the amplifier stage will permit and the value of C_a is determined from Eq. (3.76).

$$C_a = \frac{R_c C_c}{R_b} \tag{3.76}$$

where R_c and C_c = interstage coupling elements

If r_p is not large compared to R_b and R_c, the value of C_a is best determined experimentally.

3.12d. Low-frequency Compensation for the Combined Effects of Bypassing, Coupling, etc. In an amplifier, there are usually several sources which contribute to the loss in

gain and the leading phase characteristics at low frequencies. It is desirable to compensate each stage individually; however, it is not possible to choose values for R_a and C_a which will exactly compensate for the three possible sources which contribute to the inadequate low-frequency characteristics. Consequently, if good low-frequency response is required, it frequently is necessary either to use very large bypass and coupling capacitors or to avoid cathode and screen-grid bypassing and to use direct coupling between stages whenever possible. In those stages which have no frequency-sensitive networks, it is then possible to employ values of R_a and C_a which will contribute to the over-all amplifier compensation.

It is very difficult to determine the effectiveness of the over-all compensation by using a sine-wave signal generator and attempting to measure the amplifier amplitude and phase characteristics. The most effective means is to employ a square-wave generator and observe the output signal. The procedure is to adjust experimentally the values of the compensating networks until the output signal has the optimum characteristics.

3.12e. Low-frequency Boost in Audio Amplifiers. As discussed in Sec. 3.11f, there is not the severe requirement for obtaining as exact amplitude and phase compensation in an audio amplifier as in a video amplifier. Actually, low-frequency boost in an audio amplifier is frequently desirable even though it is obtained at the expense of phase distortion.

Low-frequency boost can satisfactorily be obtained by employing special equalizer networks of the types discussed in Sec. 17.2 or by employing impedance elements as plate loads whose values increase in the low-frequency region where the boost is desired. A typical example is a parallel R, L, and C network in series with the plate-load resistance. The resonant frequency for L and C determines the region in which the boost will be obtained, and the value of R, the inductance Q, the plate-load resistance, and the tube characteristics determine the amount of boost. Another example might be a network consisting of R_a and C_a as shown in Fig. 3.47. The value of R_a is chosen so that the value of A_a/A_a' determined from Eq. (3.67) is sufficient to produce the desired boost. The next step is to determine from Fig. 3.27 a value for f_a which will cause the boost to be realized at the desired frequencies. The value of C_a can then be determined from Eq. (3.66). It should be remembered that when using Figs. 3.27 and 3.28 for determining the effects of R_a and C_a, the decibel values given in the figure are gain values, phase angles are lagging, and the abscissa scale is in terms of f/f_a.

Another means of obtaining low-frequency boost is to employ a series capacitor in the feedback loop of a negative feedback amplifier. At low frequencies, the reactance of the capacitor increases and the amount of feedback decreases with an associated increase in amplifier gain. An example of a circuit of this type is shown in Fig. 3.48.
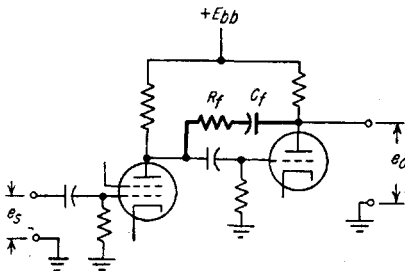
FIG. 3.48. An amplifier with feedback for bass boost.

3.13. Gain and Phase Characteristics of Multiple-stage Amplifiers. The over-all gain of a multistage amplifier can be expressed as the product of the individual stage gains. However, if the individual stage gains are given in decibels, the over-all gain in decibels will be equal to the sum of the individual stage gains in decibels. The over-all phase shift in a multistage amplifier is equal to the sum of the phase shifts in the individual stages.

3.13a. Gain of an Amplifier with n Stages. A multistage amplifier composed of n

stages will have an over-all power, voltage, or current gain A_o equal to the product of all the individual stage gains.

$$A_o = A_1 A_2 A_3 \ldots A_n$$

The value of A_o will vary as a function of frequency since the individual stage gains vary with frequency. The values of the individual stage gains must be determined when operating into the load imposed by the succeeding stages and with the output load connected. If all the stages are identical, the over-all gain A_o will be equal to A^n where n is the number of stages and A is the gain of each stage.

The power gain of the individual stages of an amplifier can also be expressed in decibels.

$$\text{Power gain, db} = 10 \log_{10} \frac{\text{power output from stage}}{\text{power input to stage}}$$

If the power gains of the individual stages of a multistage amplifier are expressed in decibels, the over-all power gain A_{op} in decibels can be expressed by the sum of the individual gains

$$\begin{aligned} A_{op}(\text{db}) &= A_1(\text{db}) + A_2(\text{db}) + \cdots + A_n(\text{db}) \\ &= 10 \log_{10} \frac{\text{power output from final stage}}{\text{power input to first stage}} \end{aligned} \tag{3.77}$$

Voltage or current gain can also be expressed in decibels. In general, most textbooks state that the input and output voltages or currents must be referred to the same impedance level in determining the voltage or current gain in decibels. In practice, however, the over-all voltage gain A_{oe} in decibels ordinarily refers to the actual ratio of the output voltage to the input voltage, and the over-all current gain A_{oi} in decibels refers to the actual ratio of the output current to the input current.

$$A_{oe}(\text{db}) = 20 \log_{10} \frac{\text{output voltage from final stage}}{\text{input voltage to first stage}} \tag{3.78}$$

$$A_{oi}(\text{db}) = 20 \log_{10} \frac{\text{output current from final stage}}{\text{input current to first stage}}$$

The incremental difference in gain $\Delta A_o(\text{db})$ of an amplifier at two different frequencies is often expressed in decibels. The decibel change in gain can be determined by the ratio of the output voltages at the two frequencies with a constant value of input signal. In general, the incremental gain is expressed as a plus or minus number of decibels with respect to the gain at the frequency chosen as a reference, e.g., the midband frequency.

$$\Delta A_{op}(\text{db}) = \Delta A_{oe}(\text{db}) = \Delta A_{oi}(\text{db}) = 20 \log_{10} \frac{e_o \text{ at evaluating frequency}}{e_o \text{ at midband}}$$

3.13b. Determination of the Gain of a Multistage Amplifier by Graphical Methods. Graphical methods are frequently used to determine the over-all gain of an amplifier over a wide frequency range if the individual stage gains are known. A logarithmic plot of the gains of the individual amplifier stages as a function of frequency, as shown in Fig. 3.49, permits the determination of the gain of the over-all amplifier. If the curves A_1 and A_2 represent the numerical gains of two amplifier stages, the sum of the heights of the respective curves above the unity gain base line, at any frequency, will be equal to the total gain of the amplifier at that frequency. The over-all amplifier gain can easily be plotted by performing the curve addition with a pair of dividers.

When the curves of gain versus frequency for the individual stages are plotted on a linear decibel gain scale, the over-all gain curve can also be determined by successive summation.

3.13c. Phase Response of an Amplifier Having n Stages. The expression for the gain of a single-stage plate-loaded amplifier including the phase response can be written in the form $A/180 + \theta$. The 180° component is the normal relationship between the plate circuit output signal and the input signal for an amplifier in which the plate, screen, and cathode circuits are resistive and electron transit time is negligible. The θ component denotes the departure from the 180° relationship due to circuit reactances. For a steady-state sine-wave input signal, the phase angle θ represents an equivalent time delay t between the input and output terminals.

$$t = \frac{\theta}{360f}$$

where t = time delay, sec
  f = frequency, cps
  θ = phase shift, deg

At any given frequency, the total phase shift is the sum of the phase angles associated with the individual stages.
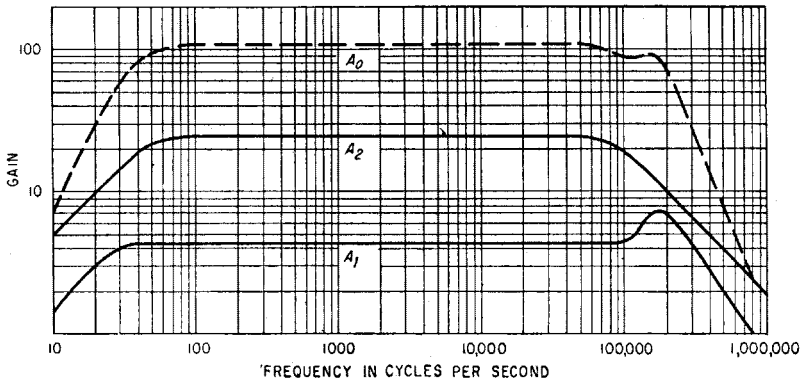
Fig. 3.49. Graphical computation of the gain as a function of frequency for a two-stage amplifier.

If sinusoidal or audio signals are to be amplified, the over-all phase shift is usually unimportant or at least noncritical. The satisfactory reproduction of complex signal waveshapes, however, can be achieved only with amplifiers having relatively uniform gain and linear phase characteristics for the major frequency components which make up the complex signal waveshape. A linear phase characteristic, i.e., the characteristic of introducing the same time delay to all frequency components, ensures that each frequency component in the output will have the same time relationship to all other frequency components as existed in the input signal.

3.14. Cathode Followers. A cathode follower is an example of a feedback amplifier which has 100 per cent negative voltage feedback. This amount of feedback accounts for the lack of gain and the low output impedance. Typical cathode-follower circuit configurations are shown in Fig. 3.50.

3.14a. Gain, Input Capacitance, and Input and Output Resistances. The gain A of a cathode follower is given by either Eq. (3.79) or (3.80). Note that the gain can never exceed $\mu/(\mu + 1)$.

$$A = \frac{\mu r_k}{r_p + r_k(\mu + 1)} \qquad (3.79)$$

and

$$\simeq \frac{g_m r_k}{1 + g_m r_k} \qquad (3.80)$$

where r_k = a-c circuit resistance between the cathode and ground, ohms

r_p = dynamic plate resistance, ohms

g_m = tube transconductance, mhos

The input capacitance c_i for a tetrode or pentode with the screen grid bypassed to the cathode is given by Eq. (3.81).

$$c_i = c_{gp} + (c_{g1g2} + c_{gk})(1 - A) \quad (3.81)$$

If the tube is a triode, c_{g1g2} is equal to zero.

If the bottom end of the grid resistor is returned either to ground, or to a bias supply, or is bypassed to ground as shown in Figs. 3.50a and 3.50b, the output resistance R_o is determined from Eq. (3.82) or (3.83) and the input resistance R_i is determined from Eq. (3.84). Equation (3.84) is based on the electron transit time and the cathode lead inductance being equal to zero.

$$R_o = \frac{1}{1/r_k + (\mu + 1)/r_p} \quad (3.82)$$

$$\simeq \frac{1}{1/r_k + g_m} \quad (3.83)$$

$$R_i = R_c \quad (3.84)$$

If the bottom end of the grid resistor is returned to a tap on the cathode resistor, as shown in Fig. 3.50c, the output resistance is determined from Eq. (3.85) or (3.86) and the input resistance is determined from Eq. (3.87). Equation (3.87) is valid only when neglecting such effects as transit time, cathode lead inductance, etc., and assuming that $R_c/(1 - A)$ is large compared to R_2.

(a) SELF OR EXTERNAL BIASING
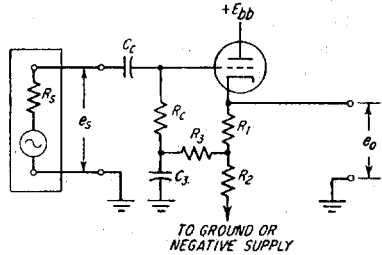
(b) CONTROL GRID DECOUPLED FROM CATHODE CIRCUIT

(c) CONTROL GRID WITH COUPLING TO THE CATHODE CIRCUIT

FIG. 3.50. Typical cathode-follower circuit configurations. Note: If tube is a tetrode or pentode, screen grid should be bypassed to the cathode.

$$R_o = \frac{r_p r_k}{r_p + r_k(\mu + 1) - \mu R_2 R_s r_k/[(R_1 + R_2)(R_c + R_s) + R_1 R_2]} \quad (3.85)$$

$$\simeq \frac{1}{1/r_k + g_m - g_m R_2 R_s/[(R_1 + R_2)(R_c + R_s) + R_1 R_2]} \quad (3.86)$$

$$R_i = \frac{R_c}{1 - A R_2/(R_1 + R_2)} \quad (3.87)$$

3.14b. Quiescent Operating Point, Signal Handling Capabilities, and Gain. In the case of a triode, the quiescent operating point is determined graphically by the intersection of the d-c plate-load line and the bias line as described in Sec. 3.3a.

If the tube is a pentode, the bias line must be drawn on the d-c transfer characteristic to determine the quiescent operating point. The d-c transfer characteristic is constructed from the d-c plate-load line as described in Sec. 3.4b.

For a triode, the slope of the d-c plate-load line is equal to the negative reciprocal of the d-c resistance R_k in the cathode circuit and is drawn from E_{bb} on the abscissa. In establishing the d-c resistance R_k of the circuit, the effect of any external d-c load

should also be included. If the tube is a pentode, the value of the cathode circuit resistance used to determine the d-c plate-load line must be modified as described in Sec. 3.4a because of the presence of the screen current. In those instances where the cathode resistor is returned to a negative supply, the d-c plate-load line must be drawn from a value on the abscissa which is equal to E_{bb} plus the absolute value of the negative supply.

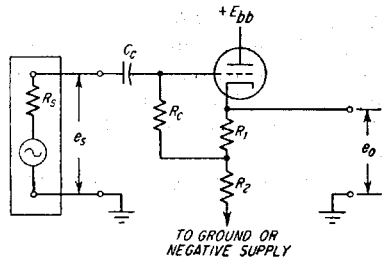The bias line for the circuit shown in Fig. 3.50a is constructed as detailed in Secs. 3.3a and 3.4c. If the circuit is of the type shown in Figs. 3.50b and 3.50c, the bias line is established in a similar manner. However, the effective resistance to be used in the construction of the bias line is $R_1R_k/(R_1 + R_2)$, where R_k is the net d-c resistance in the cathode circuit and includes R_1, R_2, and any d-c load resistance.

Having determined the quiescent operating point, it is necessary to locate the a-c plate-load line. If the cathode circuit a-c resistance is the same as the d-c resistance, the a-c load line is identical with the d-c load line. However, if the cathode circuit a-c resistance differs from the d-c resistance, it is necessary to calculate this new resistance and construct the a-c load line with its proper slope through the quiescent operating point.[1] From the a-c load line it is possible to directly determine the amplifier signal-handling capabilities and the cathode-follower gain.

Assuming no shunting capacitance across the cathode circuit impedance, it is quite simple to analyze the performance of a cathode follower. For a given output signal which is measured along the abscissa, the required input signal is equal to this output signal plus the absolute value of the change in tube bias associated with the assumed output signal.

3.14c. Maximum Sine-wave Output-signal Amplitude as Limited by Plate Current Cutoff. The maximum negative output-signal amplitude is determined by the quiescent plate current and the cathode circuit impedance. Due to capacitance shunting the cathode circuit a-c resistance, the cathode circuit impedance decreases with increasing frequency. It follows that the maximum output-signal amplitude also varies inversely with frequency. The maximum rms value of a sine-wave output signal, as established by plate current cutoff, can be determined from Eq. (3.88).

$$e_{o(max)} = \frac{0.707 I_o r_k}{\sqrt{1 + (f/f_k)^2}} \tag{3.88}$$

where $e_{o(max)}$ = maximum rms voltage in volts of a sine-wave output signal
$\qquad f$ = frequency in cycles per second at which $e_{o(max)}$ is to be determined
$\qquad f_k = 1/2\pi r_k C_k$
$\qquad I_o$ = quiescent plate current, amp
$\qquad r_k$ = cathode circuit a-c resistance, ohms

3.14d. High-frequency Response. The -3-db upper frequency limit f_2, which is determined by the tube and circuit parameters, can be calculated from Eq. (3.89) provided the output-signal amplitude is equal to or less than the value determined by Eq. (3.90). If the output-signal amplitude is larger than the value established from Eq. (3.90), distortion will result because of plate current cutoff.

$$f_2 = f_k \left[1 + \frac{r_k(\mu + 1)}{r_p} \right] \tag{3.89}$$
$$\simeq f_k(1 + g_m r_k)$$
$$e_o = \frac{0.707 I_o r_k}{\sqrt{1 + [1 + r_k(\mu + 1)/r_p]^2}}$$
$$\simeq \frac{0.707 I_o r_k}{\sqrt{1 + (1 + g_m r_k)^2}} \tag{3.90}$$

[1] To be exact, this a-c load line does not pass through the quiescent operating point because of harmonic distortion. See Sec. 3.3b.

where g_m is in mhos, r_k is in ohms, and r_p is in ohms. The high-frequency operating limits of a cathode follower as established by excessively large signals and the output shunting capacity are shown graphically in Fig. 3.51.

3.14e. *Transient Response.* The transient response of a cathode follower is dependent on the quiescent plate current, cathode impedance, and the amplitude and shape of the input signal. For positive pulses, the requirements for

(a) POSITIVE PULSE AND ASSOCIATED CURRENT WAVESHAPE

(b) NEGATIVE PULSE AND ASSOCIATED CURRENT WAVESHAPE

$A = 0.707\, I_0 r_k$

$B = 0.500\, I_0 r_k$

$C = \dfrac{0.707\, I_0 r_k}{\sqrt{1 + (1 + g_m r_k)^2}}$

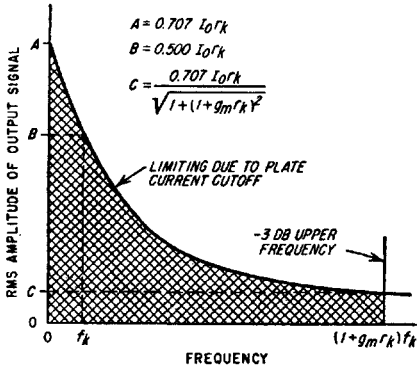LIMITING DUE TO PLATE CURRENT CUTOFF

-3 DB UPPER FREQUENCY

FIG. 3.51. High-frequency characteristics of a cathode follower. Output signal amplitude and frequency must fall in shaded area in order to avoid distortion due to plate current cutoff and/or losses greater than 3 db.

FIG. 3.52. Plate current waveshapes required to produce sample positive and negative output pulses.

satisfactory pulse reproduction are given by Eqs. (3.91), (3.92), and (3.93). Refer to Fig. 3.52 for typical current waveshapes.

$$I_o \geq \frac{0.8 C_k e}{T_2} \tag{3.91}$$

$$I_m = I_o + e \left(\frac{0.8 C_k}{T_1} + \frac{1}{r_k} \right) \tag{3.92}$$

and for best results,

$$R_o C_k \leq \frac{T_1}{5} \text{ or } \frac{T_2}{5} \quad \text{whichever is smaller} \tag{3.93}$$

where I_o = quiescent plate current, amp

C_k = cathode shunt capacitance, farads

T_1, T_2 = transition periods in seconds of leading and trailing slopes of the signal between 10 and 90 per cent amplitude points (see Fig. 3.52)

r_k = cathode circuit a-c resistance, ohms

R_o = output resistance of cathode follower, ohms

e = peak amplitude of cathode pulse, volts

I_m = maximum plate current, amp (must be obtained without grid being driven positive unless grid driving source has impedance which is low as compared to resistance of positively driven grid)
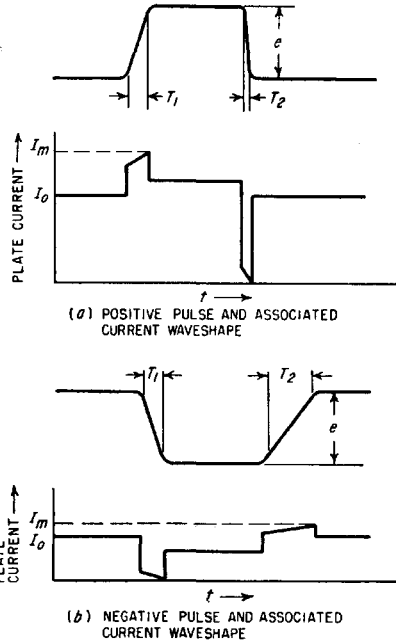
The requirements for transmitting negative pulses are contained in Eqs. (3.93), (3.94), and (3.95). Refer to Fig. 3.52 for typical current waveshapes.

$$I_o \geq e \left(\frac{0.8C_k}{T_1} + \frac{1}{r_k} \right) \tag{3.94}$$

$$I_m = I_o + \frac{0.8C_k e}{T_2} \tag{3.95}$$

Example 3.14

For the cathode follower shown in Fig. 3.53 and assuming the tube characteristics to be those given in Fig. 3.54, determine the quiescent operating point and the low-frequency gain, i.e., the gain when neglecting the effects of C_k.
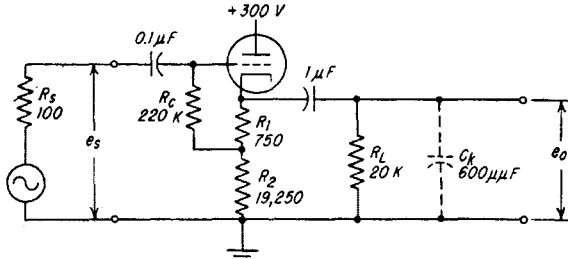
FIG. 3.53. Cathode-follower circuit for Examples 3.14 and 3.15.
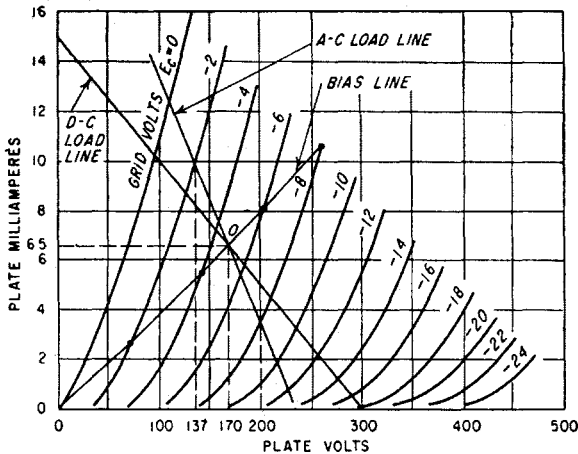
FIG. 3.54. Graphical solution for Example 3.14.

The d-c plate load line is drawn with a slope of $-1/(750 + 19,250)$ and from the 300-volt point on the abscissa as shown in Fig. 3.54. The bias line is drawn by connecting the specific values of plate current on the bias curves which are required through R_1 to obtain the associated values of bias. The point of intersection of the bias and load lines determines the quiescent operating point. The quiescent plate voltage and current are 170 volts and 6.5 ma.

The a-c load line is drawn with a slope of $-1/10,000$ through the quiescent operating point.

A low-frequency input signal which causes the instantaneous grid to cathode voltage to vary between -2 and -8 volts will develop a 63-volt output signal (137 to 200 volts as measured on the abscissa). The amplifier gain under these conditions is therefore equal to $63/(63 + 6)$, that is, (output signal)/(output signal + absolute change in bias), which is equal to 0.91. If Eq. (3.79) is used and if r_p and μ are taken as 7,700 and 20, respectively, A is calculated to be equal to 0.92.

Example 3.15

For the cathode follower shown in Fig. 3.53, determine (1) the largest 1-Mc sine-wave output signal which can be obtained and (2) the best rise time possible for a negative 18-volt output pulse.

From Eq. (3.88)

$$f_k = \frac{1}{2\pi \times 10,000 \times 600 \times 10^{-12}}$$
$$= 26.54 \text{ kc}$$

and

$$e_{o(max)} = \frac{0.707 \times 6.5 \times 10^{-3} \times 10,000}{\sqrt{1 + \left(\dfrac{10^6}{2.654 \times 10^4}\right)^2}}$$
$$= 1.22 \text{ volts (rms)}$$

From Eq. (3.94)

$$6.5 \times 10^{-3} = 18 \left(\frac{0.8 \times 600 \times 10^{-12}}{T_1} + \frac{1}{10,000}\right)$$
$$T_1 = 1.84 \ \mu\text{sec}$$

NOTE: For best performance $R_o C_k$ should be equal to or less than $T_1/5$ as specified by Eq. (3.93). In this example $R_o C_k$ is equal to 0.212 μsec; therefore the calculated rise time can be achieved.
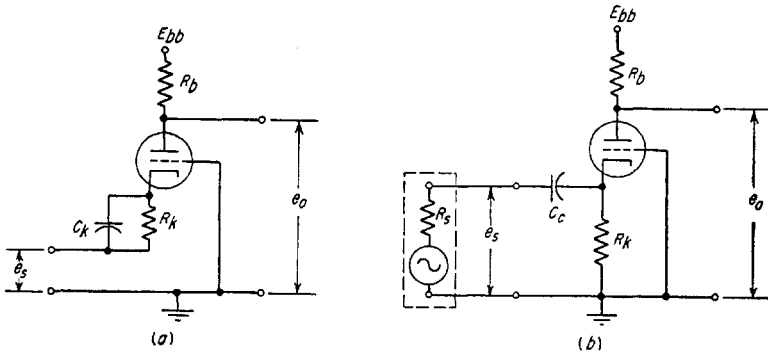
FIG. 3.55. Grounded-grid amplifiers.

3.15. Grounded-grid Amplifiers. A grounded-grid amplifier is nonphase-inverting and provides good isolation between the input and output circuits due to the low internal cathode-to-plate capacitance. At high frequencies it is more stable than a grounded-cathode stage, and, for this reason, they are frequently used in i-f and r-f amplifiers as described in Sec. 7.4i. The inherent characteristics of grounded-grid amplifiers permit the use of triodes at frequencies where grounded-cathode triodes are unstable because of their high input-to-output capacitance. The input resistance to grounded-grid amplifiers is relatively low; consequently they are often used to match low impedance lines.

The grounded-grid amplifier is unconventional in that the input signal e_s is applied to the cathode and the grid is grounded (see Fig. 3.55). The input signal is in series with the amplifier, and the circuit therefore acts as though the tube had an amplification factor of $(\mu + 1)$.

The gain of the amplifier shown in Fig. 3.55 is given by Eq. (3.96).

$$A = \frac{e_o}{e_s} = \frac{(\mu + 1)R_b}{r_p + R_b} \tag{3.96}$$

The input resistance R_i is given by Eqs. (3.97) and (3.98).

$$R_i = \frac{r_p + R_b}{\mu + 1} \quad \text{(for Fig. 3.55a)} \tag{3.97}$$

$$R_i = \frac{R_k(r_p + R_b)}{R_k(\mu + 1) + r_p + R_b} \quad \text{(for Fig. 3.55b)} \tag{3.98}$$

The output impedance Z_o is given by Eq. (3.99).

$$Z_o = \frac{R_b r_p + R_b Z_k(\mu + 1)}{R_b + r_p + Z_k(\mu + 1)} \tag{3.99}$$

where Z_k = impedance in cathode circuit at the frequency of interest.

The grid of this circuit can be used as an independent element for the application of positive or negative feedback in a multistage amplifier. The feedback signal is usually applied across a resistor from grid to ground. Such a resistance should not be large in a wideband amplifier as degeneration may result because of the cathode driving the grid through the cathode to grid capacitance.
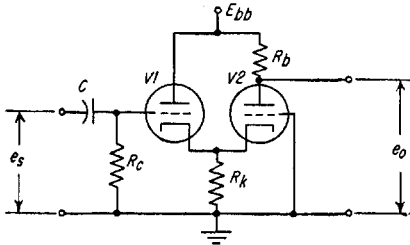

FIG. 3.56. Cathode-coupled amplifier.

Example 3.16

A generator with an open-circuit voltage of 4 volts and an internal resistance of 500 ohms is to drive the amplifier of Fig. 3.55b. Determine the gain A, the input resistance R_i, the amplifier input signal e_s, the output voltage e_o, and the output impedance Z_o. Assume that

$$\mu = 30 \qquad R_b = 27,000 \text{ ohms}$$
$$r_p = 9,000 \text{ ohms} \qquad R_k = 1,000 \text{ ohms}$$

From Eq. (3.96)

$$A = \frac{(30 + 1)27 \times 10^3}{(9 + 27)10^3} = 23.2$$

From Eq. (3.98)

$$R_i = \frac{10^3(9 + 27)10^3}{10^3(30 + 1) + (9 + 27)10^3} = 537 \text{ ohms}$$

$$e_s = e_{\text{gen}} \frac{R_i}{R_s + R_i} = 4 \times \frac{537}{500 + 537} = 2.07 \text{ volts}$$

$$e_o = A e_s = 23.2 \times 2.07 = 48.0 \text{ volts}$$

For computation of Z_o it is necessary to evaluate Z_k.

$$Z_k = \frac{R_s R_k}{R_s + R_k} = \frac{500 \times 1,000}{1,500} = 333 \text{ ohms}$$

From Eq. (3.99)

$$Z_o = \frac{(27 \times 10^3 \times 9 \times 10^3) + (27 \times 10^3 \times 333 \times 31)}{27,000 + 9,000 + 333(31)}$$

$$= 11,300 \text{ ohms}$$

3.16. Cathode-coupled Amplifiers. A cathode-coupled amplifier provides a high degree of isolation between the input and output circuits. The amplifier is nonphase-inverting and has a high input impedance.

The cathode-coupled amplifier, shown in Fig. 3.56, is essentially a cathode follower driving a grounded-grid amplifier. The common cathode resistor R_k is the coupling impedance for signal transfer from $V1$ to $V2$. The signal developed across R_k and

the output signal are in phase with the input signal at low frequencies. At high frequencies, the reactances of the shunt capacitances across R_k and R_b cause a phase angle to exist between the input and output signals.

This type of amplifier is exceptionally free from changes in input admittance due to changes in the load. Because of this property, it frequently is used as a buffer, e.g., between an oscillator and a varying load.

The analysis of the amplifier can be simplified by separating it into a cathode follower and a grounded-grid amplifier as shown in Fig. 3.57. The unloaded gain

FIG. 3.57. Circuit for the analysis of a cathode-coupled amplifier.

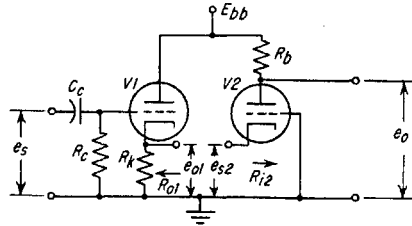of the cathode follower A_{cf} and its output resistance R_{o1} can be computed from Eqs. (3.100) and (3.101).

$$A_{cf} = \frac{e_{o1}}{e_s} = \frac{\mu_1 R_k}{r_{p1} + R_k(\mu_1 + 1)} \tag{3.100}$$

$$\simeq \frac{g_{m1}R_k}{1 + g_{m1}R_k}$$

$$R_{o1} = \frac{1}{1/R_k + (\mu_1 + 1)/r_{p1}} \tag{3.101}$$

$$\simeq \frac{1}{1/R_k + g_{m1}}$$

The input resistance R_{i2} to $V2$ can be computed with Eq. (3.102).

$$R_{i2} = \frac{r_{p2} + R_b}{\mu_2 + 1} \tag{3.102}$$

$$\simeq \frac{1}{g_{m2}} \qquad \text{provided } r_{p2} \gg R_b$$

The gain of the loaded cathode follower A'_{cf} is computed from Eq. (3.103).

$$A'_{cf} = A_{cf} \times \frac{R_{i2}}{R_{o1} + R_{i2}} \tag{3.103}$$

The gain A_2 of the grounded-grid stage is given by Eq. (3.104). The product of this value and the loaded cathode follower gain A'_{cf} determines the over-all gain A.

$$A_2 = \frac{e_o}{e_{s2}} = \frac{(\mu_2 + 1)R_b}{r_{p2} + R_b} \tag{3.104}$$

$$\simeq g_{m2}R_b \qquad \text{If } r_{p2} \gg R_b$$

$$A = \frac{e_o}{e_s} = A'_{cf}A_2 \tag{3.105}$$

The output resistance R_o of the over-all amplifier is equal to the output resistance for the single stage $V2$ when functioning with the same plate load R_b and a cathode resistor whose value is equal to the output resistance R_{o1} of $V1$.

$$R_o = \frac{R_b[r_{p2} + R_{o1}(\mu_2 + 1)]}{R_b + r_{p2} + R_{o1}(\mu_2 + 1)} \tag{3.106}$$

To determine the values of the dynamic plate resistances and the amplification factors, the operating condition of each tube must be known. The common cathode

resistor makes it necessary to determine the quiescent operating points by means of a simultaneous graphical solution. This procedure is illustrated in the following example.

Example 3.17

Determine the quiescent operating points and the gain of a dual-triode cathode-coupled amplifier in which

$$R_k = 4,000 \text{ ohms}$$
$$R_b = 30,000 \text{ ohms}$$
$$E_{bb} = 300 \text{ volts (for cathode-follower section)}$$
$$E_{bb} = 400 \text{ volts (for grounded-grid amplifier)}$$

Assume that the tube characteristics shown in Fig. 3.58 apply to each half of the dual triode.

1. Determine the quiescent operating points.
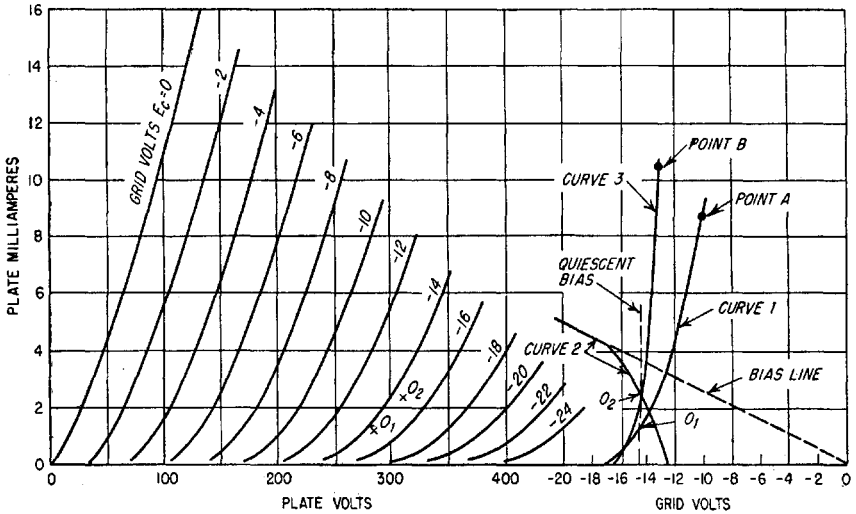A. Plot plate current versus bias for the cathode follower section.

FIG. 3.58. Graphical analysis for Example 3.17.

Assume values of bias between cutoff and zero bias and plot the values of plate current associated with the assumed values of bias. For example, if the bias is assumed to be −10 volts, the voltage across $V1$ is $300 - 10$, or 290 volts. The plate current for a bias of −10 volts and a plate voltage of 290 volts is 8.7 ma (see tube curves) and is shown as point A in Fig. 3.58. Curve 1 represents a series of points similarly located.

B. Plot the bias line.

The bias line is constructed on the same coordinates as the transfer characteristic drawn in the preceding step. The slope is $-1/R_k$, that is, $-1/4000$. The resulting line defines the bias as a function of the total cathode current.

C. Construct the curve that defines the values of I_2 which satisfy the requirement that the sum of the quiescent values of I_1 and I_2 must equal the value of current determined by the bias line.

This is accomplished by plotting those values of current obtained by subtracting curve 1, that is, I_1 versus bias, from the bias line, that is, $I_1 + I_2$ versus bias. The new curve is curve 2.

D. Calculate the plate voltage which would exist across $V2$ for the values of bias and plate current determined by curve 2. For each value of bias and the calculated value of plate voltage, determine from the tube curves the associated value of plate current. Plot these values of plate current as a function of bias.

If the bias is assumed to be −13 volts, the value of I_2 must be 0.6 ma (from curve 2). For a bias of −13 volts and a voltage drop across $V2$ of 369 volts, that is, $E_{bb} - |\text{bias}| - \text{drop}$ across R_b, or $400 - 13 - (30,000 \times 0.6 \times 10^{-3})$, the plate current through $V2$ would be

approximately 10.5 ma.   This value is shown as point B.   Curve 3 is a plot of similarly located points.

E. Determine the quiescent operating points for $V1$ and $V2$.

The intersection of curves 2 and 3 determines the quiescent bias for both $V1$ and $V2$ and the quiescent plate current through $V2$.   The quiescent plate current through $V1$ is read from curve 1.

$$I_{o2} = 2.4 \text{ ma}$$
$$I_{o1} = 1.2 \text{ ma}$$
$$E_c = -14.4 \text{ volts}$$

F. Determine the dynamic plate resistances and amplification factors for both $V1$ and $V2$ From tube curves,

$$\mu_1 = 17.5$$
$$r_{p1} = 25,000$$
$$\mu_2 = 17.5$$
$$r_{p2} = 17,000 \text{ ohms}$$

2. Compute A_{cf} from Eq. (3.100).

$$A_{cf} = \frac{17.5 \times 4,000}{25,000 + 4,000(17.5 + 1)}$$
$$= 0.707$$

3. Compute R_{o1} from Eq. (3.101).

$$R_{o1} = \frac{1}{\dfrac{1}{4,000} + \dfrac{17.5 + 1}{25,000}}$$
$$= 1,010 \text{ ohms}$$

4. Compute R_{i2} from Eq. (3.102).

$$R_{i2} = \frac{17,000 + 30,000}{17.5 + 1}$$
$$= 2,540 \text{ ohms}$$

5. Compute A_{cf}' from Eq. (3.103).

$$A_{cf}' = 0.707 \times \frac{2,540}{1,010 + 2,540}$$
$$= 0.506$$

6. Compute A_2 and A from Eqs. (3.104) and (3.105).

$$A_2 = \frac{(17.5 + 1)30,000}{17,000 + 30,000}$$
$$= 11.8$$
$$A = 0.506 \times 11.8$$
$$= 5.97$$

7. Compute R_o from Eq. (3.106).

$$R_o = \frac{30,000[17,000 + 1,010(17.5 + 1)]}{30,000 + 17,000 + 1,010(17.5 + 1)}$$
$$= 16,300 \text{ ohms}$$

3.17. Differential Amplifiers.   An ideal differential amplifier is an amplifier which will amplify only the instantaneous difference between two input signals.   Practical differential amplifiers, however, are not perfect and, consequently, amplify with a small amount of gain those signals which are common to the two input channels.

The differential amplifiers shown in Fig. 3.59 are essentially identical to cathode-

FIG. 3.59. Differential-amplifier configurations.

coupled phase inverters and differ only in the application and design considerations (see Sec. 3.18c). These circuits can be designed to have much lower gain for identical input signals than for differential signals. It is important to note that this discrimination will occur even when two identical signals are components of larger signals having differential values, e.g., power-supply ripple voltage which is in phase on the two sides of a balanced line. For balanced signals, i.e., push-pull signals, this circuit is simply a balanced amplifier. The gain of the differential amplifier for balanced input signals is given by Eq. (3.107).

$$A = \frac{e_{o1}}{e_{s1}} = \frac{e_{o2}}{e_{s2}} = \frac{-\mu R_b}{r_p + R_b} \qquad (\text{for } e_{s1} = -e_{s2}) \qquad (3.107)$$

When a single-ended input signal is applied, as when $e_{s2} = 0$, the gain to each plate is approximately equal to $A/2$ provided μR_k is large compared to $r_p + R_b$. The value of e_{o1} will be slightly larger than e_{o2}. For example, if e_{s1} is a positive input signal and e_{s2} is zero, e_{o1} will be a negative output signal whose amplitude is slightly larger than $Ae_{s1}/2$ and e_{o2} will be a positive output signal whose amplitude is slightly less than $Ae_{s1}/2$. The plate-to-plate signal will be Ae_{s1}. An exact expression for the gain of $V1$ when e_{s2} is equal to zero is given by Eq. (3.108).

$$A_o = \frac{e_{o1}}{e_{s1}} = A\frac{r_p + R_b + R_k(\mu + 1)}{r_p + R_b + 2R_k(\mu + 1)} \qquad \text{for } e_{s2} = 0 \qquad (3.108)$$

The gain A' for identical, i.e., common mode, input signals is given by Eq. (3.109).

$$A' = \frac{e_{o1}}{e_{s1}} = \frac{e_{o2}}{e_{s2}} = \frac{-\mu R_b}{r_p + R_b + 2R_k(\mu + 1)} \qquad \text{for } e_{s1} = e_{s2} \qquad (3.109)$$

A significant design parameter for differential amplifiers is the ratio of the gain for a push-pull input signal to the gain for identical input signals. This is often referred to as the *common mode rejection ratio* R_{cm}. With reference to the circuits shown in Fig. 3.59, the common mode rejection ratio as measured at the plate of either $V1$ or $V2$ can be determined from Eq. (3.110).

$$R_{cm} = \frac{A}{A'} = 1 + \frac{2R_k(\mu + 1)}{r_p + R_b} \qquad (3.110)$$

If a plate-to-plate load is employed, the common mode rejection ratio will be equal to infinity provided the system is balanced.

If the plate load resistor of the tube not being loaded is deleted, both the gain and the common mode rejection ratio can be increased somewhat beyond that given by Eq. (3.107) and Eq. (3.110). In this case, the common mode rejection ratio R_{cm} as measured at the single-ended output is given by Eq. (3.110a), the gain A_1 from the grid of the tube being loaded to the single-ended output is given by Eq. (3.110b), and the gain A_2 from the grid of the other tube to the single-ended output can be determined from Eq. (3.110c).

$$R_{cm} = 1 + \frac{2R_k(\mu + 1)}{r_p} \qquad (3.110a)$$

$$A_1 = -\frac{\mu R_b[r_p + R_k(\mu + 1)]}{r_p[r_p + R_b + R_k(\mu + 1)B]} \qquad (3.110b)$$

$$A_2 = \frac{\mu R_b R_k(\mu + 1)}{r_p[r_p + R_b + R_k(\mu + 1)B]} \qquad (3.110c)$$

where $B = 2 + R_b/r_p$

The importance of using large values of R_k can be seen from Eqs. (3.110) and (3.110a). To avoid an excessively large bias produced by a large value of R_k, the cathode can be operated from a negative supply with the grids returned to ground,

the grids can be returned to a tap on R_k, or R_k can be grounded and the grids returned to a positive voltage.

When extremely high rejection of identical input signals is desired, a tube can be used in place of R_k. Such a circuit is shown in Fig. 3.59c. If $V3$ has an unbypassed cathode resistor, R_6, and constants μ_3 and r_{p3}, the value to use in the preceding equations for R_k is given by Eq. (3.111).

$$R_k = r_{p3} + R_6(\mu_3 + 1) \qquad (3.111)$$

where R_k = equivalent cathode resistance due to $V3$ and R_6

The design of the circuits shown in Fig. 3.59 is facilitated by assigning voltage drops to R_k, the tube, and R_b such that the ratio R_k/R_b can be made as large as desired [see Eq. (3.110)]. The sum of these voltages must equal the applied d-c potential. With an assigned value of plate-to-cathode voltage E_b, it is possible to establish a suitable operating current I_b for each tube and the associated bias E_c from the tube curves. Each plate-load resistor R_b will have a value determined by the assigned voltage drop across the resistor divided by I_b. The value of the cathode resistor is equal to the assigned voltage drop across the resistor divided by $2I_b$. For the cathode circuit shown in Fig. 3.59c, it is first necessary to select a tube type for $V3$ and a value for R_6 which will maximize the equivalent value of R_k as established by Eq. (3.111). The value of E_{c3}, which is consistent with the desired plate voltage E_{b3}, voltage drop across R_6, and current $2I_b$, can be determined from the E_b-I_b curves for $V3$. The divider network consisting of R_7 and R_8 can then be designed to place the desired bias on $V3$.

The bias for $V1$ and $V2$ required for the assigned tube current I_b and the plate voltage E_b can be obtained by several methods as shown in Fig. 3.59. For the circuit shown in Fig. 3.59a, the voltage drop assigned to the cathode resistor must be such that the amplitude of the cathode voltage above ground is equal to the required amplitude of E_c. For the circuit shown in Fig. 3.59b, the value of R_1 is made equal to $E_c/2I_b$. The amplitude of the quiescent voltage on the grids of $V1$ and $V2$ in Fig. 3.59c, due to the divider network composed of R_4 and R_5, must be less than the quiescent cathode voltage of $V1$ and $V2$ with reference to ground by an amount equal to the absolute grid bias voltage on the tubes. The method of grid biasing shown in Fig. 3.59c is satisfactory if the power supply is free from objectionable amounts of ripple. If this is not the case, decoupling should be employed. Whenever the cathodes of the differential amplifier tubes are at an elevated potential with respect to ground, the maximum cathode to heater potential rating should not be exceeded. This may require a separate elevated heater supply.

The quiescent operating points for $V1$ and $V2$ can easily be determined graphically. Since each of the two tubes is operated under identical conditions, it is necessary only to consider one tube. The cathode resistor, however, must be treated as though it had a value equal to $2R_k$. In this manner, the effects of the other tube are included.

Example 3.18

If a dual triode is operated under the conditions where $E_c = -2$ volts, $\mu = 70$, $r_p = 44,000$ ohms, $I_b = 2.3$ ma (each half section), and $E_b = 250$ volts:

1. Design a differential amplifier to operate from a negative supply of -150 volts and a positive supply of 300 volts and to have both grid resistors returned to ground as shown in Fig. 3.59a.

With the grid resistors returned to ground, the cathode voltage E_k is equal to $-E_c$ or $+2$ volts.

$$R_k = \frac{|E_{\text{neg}}| + |E_c|}{2I_b} = \frac{150 + 2}{0.0046} = 33,000 \text{ ohms}$$

$$R_b = \frac{E_{bb} - (E_b + E_k)}{I_b} = \frac{300 - (250 + 2)}{0.0023}$$

$$= 20,900 \text{ ohms}$$

2. Assuming the output is to be taken from either the plate of $V1$ or $V2$, determine the gain for differential and identical input signals and determine the value of A/A'.

From Eq. (3.107)

$$A = \frac{-70 \times 20,900}{44,000 + 20,900} = -22.5$$

From Eq. (3.109)

$$A' = \frac{-70 \times 20,900}{44,000 + 20,900 + (2 \times 33,000 \times 71)} = -0.308$$

$$\frac{A}{A'} = \frac{22.5}{0.308} = 73.1$$

3.18. Phase Inverters. A phase inverter is used when it is necessary to convert a single-ended signal, i.e., unbalanced to ground, to a push-pull or balanced signal. Although this result can be achieved with a transformer having a center-tapped secondary or an inductor having a center tap, this discussion is limited to tube methods.

The desirable properties of a phase inverter are equal output signals, relative independence from effects due to power-supply variations, and long-time stability. In some cases bandwidth, output resistance, or number of components required will be of importance.

The most common types of phase inverters are:

1. Side-branch phase inverter
2. Split-load phase inverter
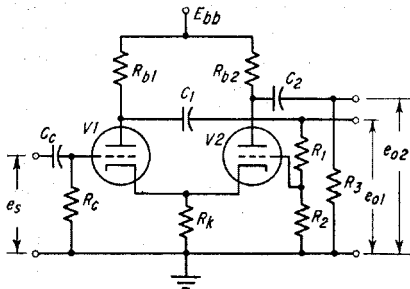3. Cathode-coupled phase inverter
4. Floating paraphase inverter

3.18a. Side-branch Phase Inverter. Ideally, if the plate circuits of the phase inverter shown in Fig. 3.60 are symmetrical and linear, a common unbypassed cathode resistor can be used without causing degeneration. In practice, where the tube characteristics are not linear or symmetrical, the balance of the outputs will be improved if the cathode is unbypassed, since the resulting degeneration will tend to equalize the currents through $V1$ and $V2$. The voltage divider consisting of R_1 and R_2 must satisfy Eq. (3.112).

FIG. 3.60. Side-branch phase inverter.

$$-\frac{R_1 + R_2}{R_2} = A_2 = \frac{-\mu_2 r_{b2}}{r_{p2} + r_{b2}} \tag{3.112}$$

To preserve plate circuit symmetry, R_3 should be made equal to the sum of R_1 and R_2. The output resistance R_o at either plate is given by Eq. (3.113).

$$R_o = \frac{r_p r_b}{r_p + r_b} \tag{3.113}$$

Disadvantages of this circuit are that changes in either the circuit elements or the characteristics of $V2$ will cause an unbalance in the two outputs. The circuit simplicity and small number of components are advantageous.

Example 3.19

Determine the values of the circuit components for a phase inverter using tubes having the constants listed below:

$$\mu = 20 \qquad R_{b1} = R_{b2} = 33,000 \text{ ohms}$$
$$r_p = 12,000 \text{ ohms} \qquad R_3 = 470,000 \text{ ohms}$$

1. Determine the sum of R_1 and R_2.
To preserve plate circuit symmetry:

$$R_1 + R_2 = R_3 = 470,000$$

2. Determine the a-c plate-load resistance r_{b2} for $V2$.

$$r_{b2} = \frac{R_{b2}R_3}{R_{b2} + R_3} = \frac{33,000 \times 470,000}{503,000} = 30,800 \text{ ohms}$$

3. Determine A_2.
From Eq. (3.112)

$$A_2 = \frac{-20 \times 30,800}{12,000 + 30,800} = -14.4$$

4. Determine the values of R_1 and R_2.
From Eq. (3.112)

$$\frac{R_1 + R_2}{R_2} = 14.4$$

$$R_2 = \frac{470,000}{14.4} = 32,600 \text{ ohms}$$

$$R_1 = 470,000 - 32,600 = 437,400 \text{ ohms}$$

In practice a potentiometer is usually used to provide the correct values of R_1 and R_2.
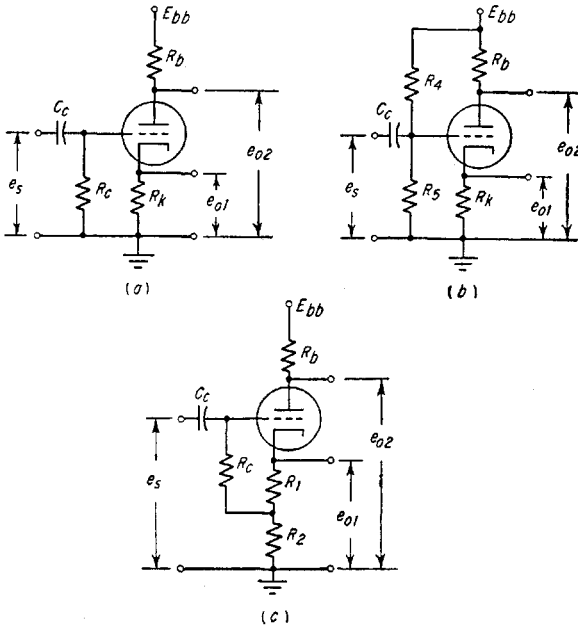
FIG. 3.61. Split-load phase inverter.

3.18b. Split-load Phase Inverter. An amplifier with an equal load in the plate and cathode circuits comprises a very simple phase inverter as shown in Fig. 3.61a. In a triode circuit, the plate current flowing through the two resistors results in two equal voltages of opposite phase. Pentodes or other multielement tubes are not as satisfactory, since the cathode and plate currents are not identical and do not necessarily retain the same ratio from tube to tube.

The gain of the split-load phase inverter to either output is less than unity.   However, the following stage is driven by a signal which is the sum of e_{o1} plus e_{o2}, indicating that the total stage gain is double the gain to either output.   The gain from the input to either output for the circuits shown in Fig. 3.61 is given by Eq. (3.114).

$$A = \frac{e_{o1}}{e_s} = \frac{-e_{o2}}{e_s} = \frac{\mu R_k}{r_p + R_b + R_k(\mu + 1)} \qquad (3.114)$$

where $R_k = R_b$

Equation (3.114) indicates that with large values of R_k the gain can approach $\mu/(\mu + 2)$ as a limit.   In the circuit shown in Fig. 3.61a, any attempt to make R_k extremely large will not be practicable since the bias on the tube would be increased to cutoff.   The circuits shown in Figs. 3.61b and 3.61c make it possible to obtain any reasonable value of bias with any given value of R_k and thus permit the gain to become very nearly equal to $\mu/(\mu + 2)$.   Another big advantage of these modified circuits
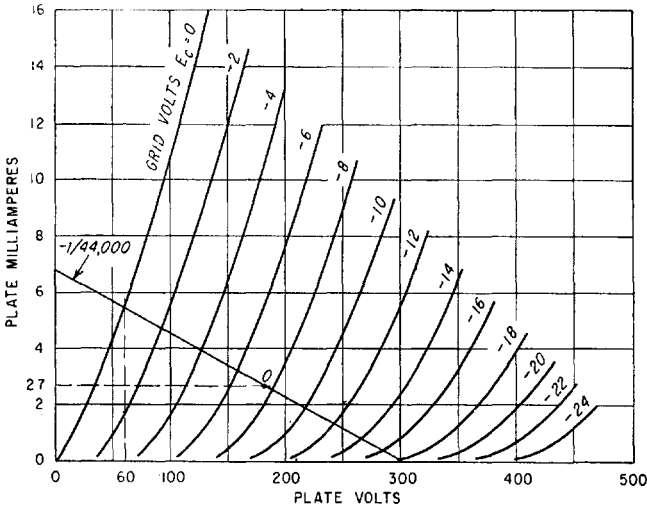

FIG. 3.62. Graphical analysis for Example 3.20.

is the ability to handle larger signals than is possible with the simple circuit.   The configuration shown in Fig. 3.61a can never deliver a sinusoidal output signal with a peak value greater than the bias voltage.

In the split-load phase inverter the two outputs have different source resistances R_o.

$$R_{o1} = \frac{R_k(r_p + R_b)}{r_p + R_b + R_k(\mu + 1)} \qquad (3.115)$$

$$R_{o2} = \frac{R_b[r_p + R_k(\mu + 1)]}{r_p + R_b + R_k(\mu + 1)} \qquad (3.116)$$

Disadvantages of this type of circuit are the low gain and the ability to handle only approximately one-half of the peak signal level delivered by the other types of inverters.   This fact is apparent since the sum of the two peak-to-peak output signals plus the minimum tube-plate voltage can never exceed the supply voltage E_{bb}.

Example 3.20

Design a split-load phase inverter for maximum signal handling ability (see Fig. 3.61b or Fig. 3.61c) assuming that the tube to be used has the characteristics shown in Fig. 3.62. Use a 300-volt power supply and 22,000-ohm loads for R_k and R_b.   Determine the maximum peak-to-peak output voltage, the gain, and the output resistances.

1. Determine the quiescent operating point and the maximum peak-to-peak output voltage.

It is necessary to draw a load line to establish the optimum operating point. In Fig. 3.62 a load line is drawn for 44,000 ohms $(R_k + R_b)$. The maximum plate current occurs at zero bias and is equal to 5.4 ma. The plate current can therefore swing between 0 and 5.4 ma. If the midpoint, i.e., 2.7 ma, is selected as the quiescent operating point, the static voltage at the cathode will be equal to 22,000 × 0.0027, or 59.4 volts. The maximum peak-to-peak output voltage swing will therefore be between 0 and 2 × 59.4, or 119 volts.

Along the load line, the bias E_c corresponding to 2.7 ma is −7.5 volts. The other constants taken at this operating point are:

$$\mu = 20$$
$$r_p = 14,000 \text{ ohms}$$

2. Determine the circuit values.

If the circuit of Fig. 3.61b is used, the grid divider composed of R_4 and R_5 must be selected to apply 59.4 − 7.5, or 51.9 volts, to the grid. For the circuit of Fig. 3.61c, the resistor R_1
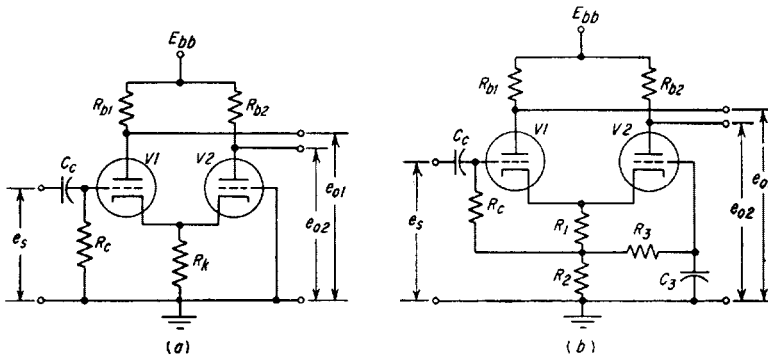

FIG. 3.63. Cathode-coupled phase inverter.

provides the bias. The required value of R_1 would be 7.5/0.0027, or 2,780 ohms. The resistor R_2 is made equal to 22,000 − 2,780, or 19,220 ohms.

3. Determine the gain to each output and the two output resistances.

From Eq. (3.114)

$$A = \frac{20 \times 22,000}{14,000 + 22,000 + 22,000(21)} = 0.88$$

From Eq. (3.115)

$$R_{o1} = \frac{22,000(14,000 + 22,000)}{14,000 + 22,000 + 22,000(21)} = 1,590 \text{ ohms}$$

From Eq. (3.116)

$$R_{o2} = \frac{22,000[14,000 + 22,000(21)]}{14,000 + 22,000 + 22,000(21)} = 21,000 \text{ ohms}$$

3.18c. Cathode-coupled Phase Inverter. In Fig. 3.63a, the signal current i_1 through $V1$ produces a voltage across the cathode resistor R_k which in turn causes an out-of-phase current i_2 to flow through $V2$. The value of i_1 will always be greater than i_2; consequently, for a balanced output R_{b2} must theoretically be larger than R_{b1}. The required value of R_{b2} for a balanced output is given by Eq. (3.117).

$$R_{b2} = \frac{R_{b1}[r_{p2} + (\mu_2 + 1)R_k]}{(\mu_2 + 1)R_k - R_{b1}} \tag{3.117}$$

It can be seen that if the value of $(\mu_2 + 1)R_k$ is very large compared to r_{p2} and R_{b1}, R_{b2} can be made equal to R_{b1}. The circuit shown in Fig. 3.63b permits a large value of R_k since the desired bias can be obtained by selecting the proper value for R_1.

Having selected the values of R_{b1} and R_{b2} for a balanced output, the gain can be determined from Eq. (3.118).

$$A = -\frac{e_{o1}}{e_s} = \frac{e_{o2}}{e_s} = \frac{-\mu_1 R_{b1}}{r_{p1} + R_{b1} + R_k(\mu_1 + 1)\left(1 - \dfrac{R_{b1}}{R_{b2}}\right)} \qquad (3.118)$$

The cathode-coupled phase inverter offers the advantages of a simple circuit and a voltage gain. The disadvantages are that changes in μ_2, r_{p2}, R_k, R_{b1}, and R_{b2} will
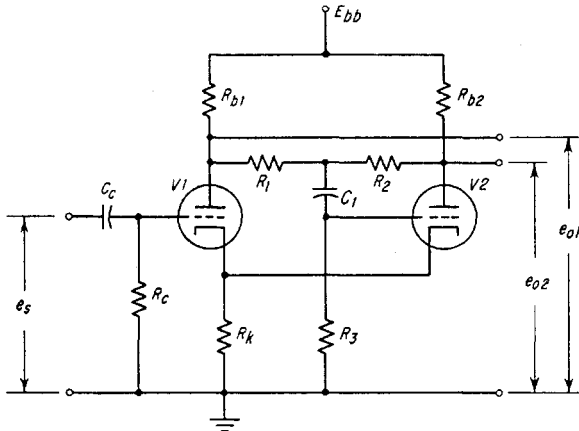

FIG. 3.64. Floating paraphase inverter.

affect the degree of balance. The larger the value of R_k, the less the balance will be affected for a given percentage change in R_k.

Example 3.21

Determine the value of R_{b2} for balanced output and the gain for the cathode-coupled phase inverter in which $V1 = V2$, $\mu = 20$, $r_p = 14,000$ ohms, $R_k = 1,500$ ohms, and $R_{b1} = 10,000$ ohms.

1. Determine the required value of R_{b2} from Eq. (3.117).

$$R_{b2} = \frac{10,000[14,000 + (20 + 1)1,500]}{(20 + 1)1,500 - 10,000}$$
$$= 21,200 \text{ ohms}$$

2. Determine the gain to the plate of $V1$ from Eq. (3.118).

$$A = \frac{-20 \times 10,000}{14,000 + 10,000 + 1,500(20 + 1)\left(1 - \dfrac{10,000}{21,200}\right)}$$
$$= -4.92$$

3.18d. Floating Paraphase Inverter. The circuit illustrated in Fig. 3.64 is characterized by the addition of the two output signals e_{o1} and e_{o2} to develop the input signal for $V2$. The feedback around $V2$ tends to stabilize the performance of this part of the circuit. The cathode resistor should be bypassed unless the plate loads for $V1$ and $V2$ are identical. In the case of identical plate loads, an unbypassed cathode resistor will tend to improve a faulty signal balance and will introduce no degeneration when the output is balanced.

To produce equal output voltages it is necessary to satisfy Eq. (3.119).

$$R_1 = \frac{R_2 R_3(|A_2| - 1)}{R_2 + R_3(|A_2| + 1)} \qquad (3.119)$$

where

$$A_2 = \frac{-R_{b2}(\mu_2 R_2 - r_{p2})}{R_2(r_{p2} + R_{b2}) + r_{p2}R_{b2}} \qquad (3.120)$$

This circuit offers relatively good balance and stability.

Example 3.22

Determine the required value of R_1 for a balanced output and the resulting gain if $V1 = V2, \mu = 20, r_p = 10,000$ ohms, $R_2 = 220,000$ ohms, $R_3 = 1$ megohm, and

$$R_{b2} = 40,000 \text{ ohms.}$$

1. Determine A_2 from Eq. (3.120).

$$A_2 = \frac{-40,000(20 \times 220,000 - 10,000)}{220,000(10,000 + 40,000) + 10,000 \times 40,000}$$
$$= -15.4$$

2. Determine R_1 from Eq. (3.119).

$$R_1 = \frac{220,000 \times 10^6(15.4 - 1)}{220,000 + 10^6(15.4 + 1)}$$
$$= 191,000 \text{ ohms}$$

3.19. D-C Amplifiers. D-C amplifiers are used primarily in instruments, servo-mechanisms, analog computers, and other devices in which the amplification of either a d-c voltage or a slowly varying voltage is required. The basic types of d-c amplifiers are direct-coupled amplifiers, carrier amplifiers, and amplifiers employing the combination of a direct-coupled amplifier and a carrier amplifier. This latter type of amplifier ordinarily employs a particular type of carrier amplifier known as a chopper amplifier. The combination of a direct-coupled amplifier and a chopper amplifier is usually referred to as a chopper-stabilized amplifier.

3.19a. Direct-coupled D-C Amplifiers. These amplifiers use direct coupling between stages, and, in general, the basic design is either limited or complicated by design considerations which fall in one or more of the following categories:

1. Interstage coupling techniques
2. Variable attenuators
3. Level adjustment
4. Stability

1. Interstage Direct-coupling Techniques. Conductive coupling is required from input to output terminals. The use of reactive elements, e.g., capacitors and peaking coils, must be restricted to those cases where the resulting frequency response is acceptable or desirable.

Since the plate voltage on a tube must be higher than its grid voltage, cascaded power sources are required for a multistage amplifier if the plate of each tube is connected to the grid of the following stage. In Fig. 3.65a, a cascaded power supply
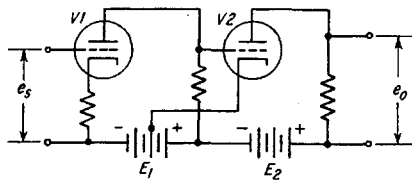
(a) D-C AMPLIFIER WITH CASCADED POWER SUPPLIES
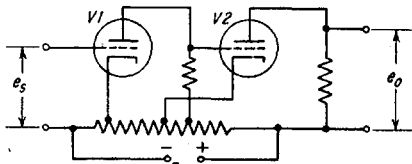
(b) D-C AMPLIFIER WITH SINGLE POWER SUPPLY AND DIVIDER

FIG. 3.65. Methods of supplying power to multistage d-c amplifiers.

using batteries is illustrated. A single power supply may be used with a voltage divider supplying the specific potentials as shown in Fig. 3.65b. For amplifiers with more than two stages, higher voltages are required and other interstage coupling methods usually become more desirable. Inherent disadvantages of the type of coupling shown in Fig. 3.65 are the high cathode operating potentials in the latter stages, the fact that the output is superimposed on a high d-c voltage, and undesirable degeneration or regeneration due to common power source impedances. One advantage, however, is that there is no coupling attenuation between stages.

A form of interstage coupling that avoids the problem of cascaded power sources is one which utilizes a coupling divider which is returned to a negative supply as shown in Fig. 3.66a. This type of coupling results in reduced gain since the signal is attenuated between the plate and the following grid by the ratio $R_2/(R_1 + R_2)$. As the voltage of the negative supply becomes larger, the value of this fraction increases, reducing the attenuation. This type of coupling will have impaired high-frequency response because of the resistor R_1, through which signal currents must flow, and the input capacitance c_i of the following tube. This condition can be corrected
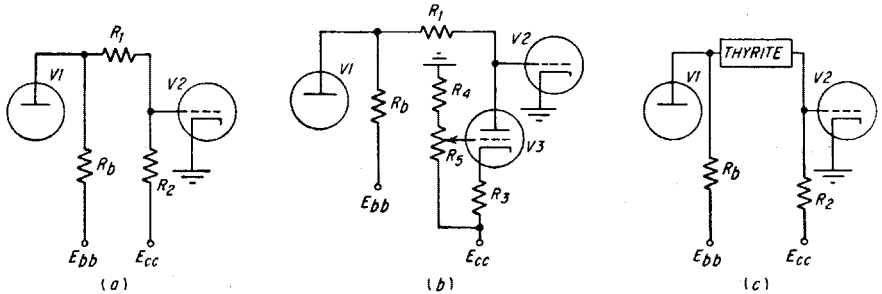

FIG. 3.66. Interstage coupling networks utilizing a negative supply.

by paralleling R_1 with a capacitor C_1. The value of C_1 to form a frequency-compensated voltage divider is given by Eq. (3.121).

$$C_1 = \frac{R_2 c_i}{R_1} \tag{3.121}$$

In Fig. 3.66b, a tube having a combination of fixed and self-bias has been substituted for R_2. The circuit can be adjusted to have the desired d-c resistance to place the required bias on $V2$ and will have an a-c resistance equal to $r_{p3} + R_3(\mu_3 + 1)$. If the a-c resistance is sufficiently large, the signal loss incurred by the divider will be negligible.

A thyrite element is a passive nonlinear device having an a-c resistance which is smaller than its d-c resistance for any value of d-c current through the element. For this reason, thyrites are sometimes used in d-c interstage coupling networks as shown in Fig. 3.66c, since they provide less attenuation to a signal than does a resistor which has a value equal to the d-c resistance of the thyrite element.

The type of d-c coupling illustrated in Fig. 3.67a employs a gas tube, usually a voltage reference type, to introduce a nearly constant voltage drop between the plate and the succeeding grid. Gas tubes introduce several problems which limit the possible application of this circuit. The current through a gas tube must be equal to or greater than some minimum value below which the regulation is poor. This stipulates that the resistors R_b and R_1 must be made small enough to ensure that more than this minimum current flows. By limiting the value of the resistors that constitute the plate load for the amplifier, the available gain is also limited. Other

problems associated with gas tubes are random noise and changes in d-c resistance as a function of temperature and ambient light level. Gas tubes generate wideband noise currents and random d-c voltage level changes as great as one volt. The bandwidth of the noise current can be reduced by a capacitor shunting the gas tube. The maximum value of this capacitor is specified by the manufacturer. In general, the noise and random d-c level changes prohibit the use of this type of coupling in sections of an amplifier where the signal levels are low. The gain can be computed in the normal manner where the plate load is assumed to be equal to R_b in parallel with R_1. This is based on the assumption that the a-c resistance of the gas tube is much less
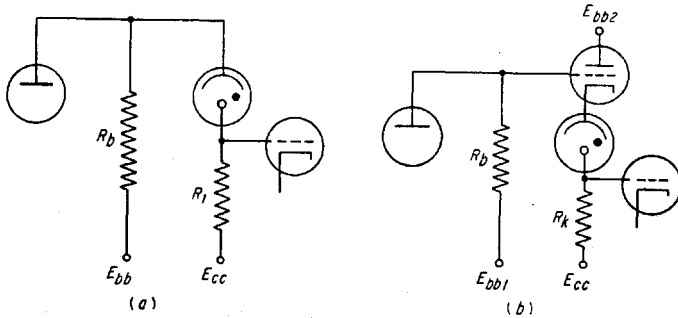

FIG. 3.67. D-C amplifier coupling utilizing gas tubes.

(a) D-C COUPLING WITH SCREEN   (b) HIGH-FREQUENCY COMPENSATION FOR
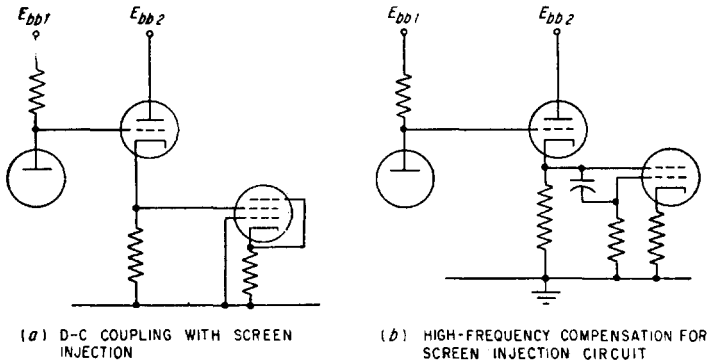   INJECTION   SCREEN INJECTION CIRCUIT

FIG. 3.68. D-C coupling to screen grid.

than the value of R_1. If this assumption is valid, the signal at the grid of the following tube is approximately equal to the signal existing at the plate of the first tube. The resistor R_1 may be returned to ground or to a negative supply.

Figure 3.67b illustrates an improved circuit for use of the gas tube, but the improvement is obtained at the expense of adding a cathode follower to supply the current for the gas tube which permits R_b to be determined by amplifier considerations only. The gain of this composite stage will be the amplifier gain multiplied by the cathode-follower gain provided R_k is much larger than the a-c resistance of the gas tube. As in the other gas-tube circuit, the noise can be reduced by shunting the gas tube with a capacitor.

The circuit in Fig. 3.68a makes it unnecessary to drop the d-c level between the output of one tube and the input to the next. This is possible because of the ability of the pentode to operate with the screen potential equal to or higher than that of the plate. The low input resistance of the screen grid makes it desirable to drive the

screen from a cathode follower to avoid loading the preceding amplifier. The gain of a pentode with signal injection in the screen grid is low compared to that possible with grid injection. The gain of this type of amplifier can be evaluated in the conventional manner if the screen-grid constants μ_{o2} or g_{o2p} are substituted for the control-grid constants. The control grid of the second stage can be used to extend the high-frequency response far beyond the normal upper frequency limit. The circuit of Fig. 3.68b shows a method of a-c coupling the signal to the control grid. This must not be done within the normally flat response range of the amplifier since it will cause a pronounced hump in the gain-frequency characteristic. If, however, the coupling of the signal into the control grid becomes appreciable only at those frequencies where the plate-load impedance is dropping because of shunt capacitance, the region of essentially flat response will be increased.
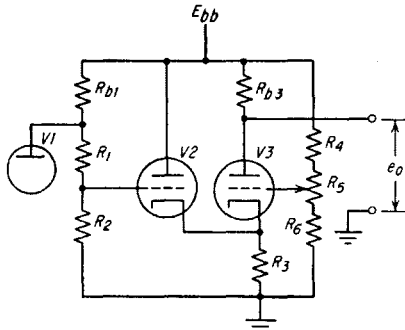

Fig. 3.69. D-C coupling using a cathode follower and cathode coupling.

Another type of interstage coupling which does not require a negative supply is that shown in Fig. 3.69. The quiescent cathode voltage on $V2$ and $V3$ can be made relatively high if the value of R_3 is made sufficiently large. In this manner, the
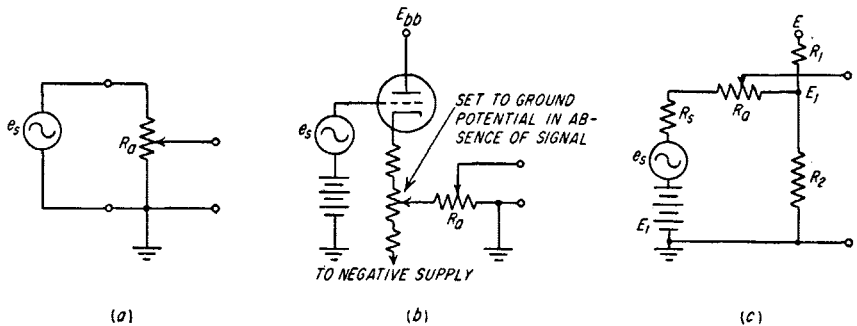

Fig. 3.70. Variable attenuators for d-c amplifiers.

quiescent grid voltage on $V2$ can also be made relatively high, thereby permitting a small signal loss in the divider composed of R_1 and R_2. If signal inversion is desired, the grids of $V2$ and $V3$ can be interchanged.

2. Variable Attenuators for Direct-coupled Amplifiers. In general, the operation of a variable attenuator in a direct-coupled d-c amplifier must not produce any change in the quiescent output voltage.

A basic principle to be followed in the location of an attenuator within an amplifier is that the two points joined by the potentiometer or step switch must have the same steady-state potential with no input signal. The simplest variable attenuator is illustrated in Fig. 3.70a. This type is satisfactory where one terminal of the generator is grounded and the zero signal level is ground potential. Another attenuator of this type is illustrated in Fig. 3.70b where a signal on a d-c base is applied to a cathode follower operating between a positive and a negative power supply. The attenuator has one of its reference points taken as ground; hence, the other end of the attenuator must be connected to a point on the cathode-follower resistor so as to be

at ground potential in the absence of an input signal.   A reference voltage other than ground potential can be used by employing a voltage regulator for the stable reference point.   If a voltage divider is used to provide the reference potential as shown in Fig. 3.70c, complete attenuation of the signal cannot be realized. With the potentiometer R_a, the achievable ratio K of the maximum output to the minimum output is given by Eq. (3.122).

$$K = 1 + \frac{R_a(R_1 + R_2)}{R_1 R_2} \quad (3.122)$$

In Fig. 3.71, $V1$ and $V2$ constitute a cathode-coupled amplifier whose gain can be varied by the potentiometer R_a.   If the bias on $V2$ is initially adjusted so that the cathode voltage on $V2$ is equal to the cathode voltage on $V1$ in the absence of an input signal, different settings of R_a will introduce different values of attenuation to the signal without disturbing the quiescent output level.   For this circuit, the ratio K achievable with R_a is given by Eq. (3.123).
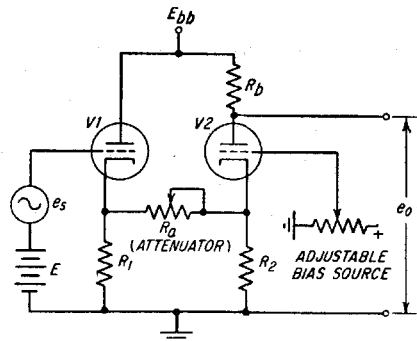

FIG. 3.71. Method of introducing attenuation in a cathode-coupled amplifier. If the bias on $V2$ is adjusted so that the cathode voltage on $V2$ is equal to the cathode voltage on $V1$ in the absence of a signal, different settings of R_a will not cause d-c level shifts in the output.

$$K = 1 + \frac{R_a}{r_{p1}R_1/[r_{p1} + R_1(\mu_1 + 1)] + R_2(r_{p2} + R_b)/[r_{p2} + R_b + R_2(\mu_2 + 1)]} \quad (3.123)$$

In a push-pull d-c amplifier, the method of introducing attenuation shown in Fig. 3.72 is frequently used.   The available attenuation is infinite, and if the maximum value of R_a is much greater than $R_{b1} + R_{b2}$, essentially all of the potential gain of the stage can be realized.

3. Level Adjustment of a Direct-coupled Amplifier. Because of the fact that there may be changes in the output voltage as a result of tube replacements, variations in heater and supply voltages, temperature changes and aging effects in components, it is necessary to provide a means of adjusting the quiescent output voltage level in a high-gain direct-coupled amplifier. In general, the procedure is to place the amplifier input terminals at the zero signal input level and adjust the level control to obtain the desired quiescent output voltage.   This is usually referred to as "balancing" the amplifier.   The balancing control must be in a low-level stage; otherwise excessive drift in the input stage might cause either the second or third stage of the amplifier to be cut off

FIG. 3.72. Attenuator for a balanced d-c amplifier.

or driven into the positive grid region.   In many applications, it is important that the gain not be changed by an adjustment of the balance control.

An amplifier having a push-pull or differential input can be balanced by the addition of a potentiometer as shown in Fig. 3.73.   Balancing of a differential input stage can also be achieved by connecting one end of a small potentiometer to the cathode of

one tube and the other end to the cathode of the other tube.   The cathode resistor should be connected to the potentiometer arm.   With either circuit, a minor adjustment of the balance potentiometer causes no significant change in gain.

Other methods of providing output level adjustments without causing objectionable gain changes utilize the coupling networks shown in Figs. 3.66b and 3.69.   Balancing is accomplished by an adjustment of R_5.

With reference to Fig. 3.67a or b, the signal level is very nearly the same at both the cathode and anode of the gas tube.   Therefore, if a potentiometer is shunted across the gas tube, the signal voltage at the arm will remain very nearly the same as the potentiometer is adjusted.   The d-c level, however, will change appreciably. If the grid of the following stage is connected to the arm of the potentiometer, the potentiometer will serve as a satisfactory balancing control.
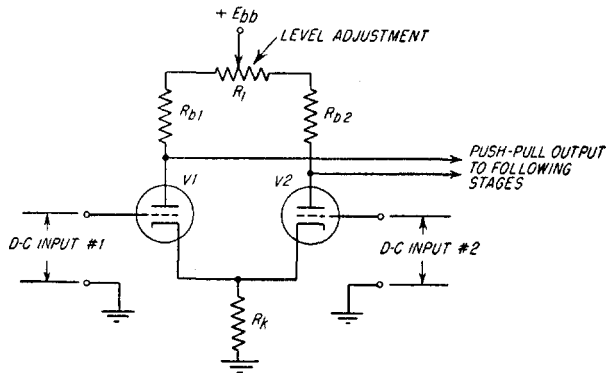

FIG. 3.73. Method of adjusting the output level of a d-c amplifier having a differential input stage.

4. Stability of a Direct-coupled Amplifier.   In a high-gain direct-coupled amplifier, a very small drift in the input stage can result in a significant error in the d-c output voltage.   Drift in the input stage can ordinarily be attributed to one or more of the following:

1. Variations in grid current
2. Variations in heater voltage
3. Variations in the d-c supply voltages
4. Variations in the input tube and other circuit components

Effect of Grid Current Variations.   A constant value of grid current in the input stage of a high-gain direct-coupled amplifier is usually not detrimental since the balance control can be used to compensate the bias component introduced by the grid current flowing through the grid circuit impedance.   However, a small change in the grid current through the grid circuit impedance can produce a sufficiently large bias change to introduce an undesirable change in the d-c output voltage.   Replacement of tubes of the same type having considerably different values of grid current will also necessitate a readjustment of the balance control.   If the balance control is in a low-level stage, this should not present a problem.   If the balance control is not in one of the low-level stages, variations in grid current in the input stage may require that the balance-control voltage range be exceptionally large.

Changes in grid current are introduced by heater voltage changes, aging of tubes, and variations in the grid, screen grid, and plate voltages.   Grid current can be either positive or negative even though the grid-to-cathode voltage is negative (see Sec.

3.8a). A typical plot of grid current as a function of bias voltage for a constant value of plate voltage is shown in Fig. 3.74. At the larger values of bias, the grid current is negative and has the minimum rate of change as a function of bias voltage. Another important consideration is that the changes in grid current resulting from heater voltage changes will be considerably less for larger values of grid bias since a change in heater voltage essentially shifts the grid current curve horizontally. In general, the transition between negative and positive grid current will occur at a bias voltage between -1 and -2 volts for most receiving-type tubes.

In the design of a low-level input stage, the input tube should be operated in the negative grid current region so as to minimize grid current changes. Also, a type of tube which has low microphonics and which has small tube-to-tube variations in both grid current and other characteristics is ordinarily selected in preference to a tube type which has high gain only. Therefore it is suggested that experimental test

FIG. 3.74. Typical plot of grid current as a function of grid bias for a constant value of plate voltage.

data be obtained for the tube types being considered for the input stage of a high-gain direct-coupled amplifier. The bias region in which the grid current remains fairly constant as a function of grid voltage should be established, and a comparison of many tubes of the same type should be made, in order to determine the expected variations from tube to tube. Differences in grid currents among tubes of the same type appear to be significantly larger for subminiature tubes than for miniature and octal receiving-tube types. A further consideration which often is of importance is that the plate current associated with the operating conditions required to optimize the grid current characteristics must be sufficiently large to permit operation in the regions covered by the published tube characteristics. Otherwise, the performance of the tube cannot be predicted and, in some cases, may vary appreciably from tube to tube.

Effect of Heater Voltage Variations. The effect of heater voltage changes is most pronounced in the input stage of a direct-coupled amplifier. Variations in the heater voltage cause changes in both the grid current and the voltage required between the cathode and the other electrodes to maintain a given value of plate current. Assuming a low value of plate current, approximately a 100-mv increase in grid bias is required to maintain a constant plate current if the heater voltage of a tube having an oxide-coated unipotential cathode is increased by 10 per cent. This relationship can also be expressed by stating that there is an equivalent grid bias change of 100-mv as the result of a 10 per cent heater voltage change. As the plate current increases, the effect becomes somewhat greater and more erratic.

Although heater voltage changes can be satisfactorily minimized by regulating the line voltage or by applying a regulated d-c heater voltage to the input tube, more economical methods are ordinarily used. As an example, a fraction of the output voltage of an unregulated power supply is sometimes used to vary the control grid, screen grid, or cathode voltage of the input tube so as to compensate for the heater voltage changes. However, more typical methods of compensating for heater voltage variations include the use of an additional tube. The plate current variations in the added tube due to changes in heater voltage causes the voltage across a common cathode resistor to vary. This change in cathode voltage introduces a change in the grid-to-cathode potential of the input tube of the correct polarity so as to tend to hold the plate current of the input tube constant.
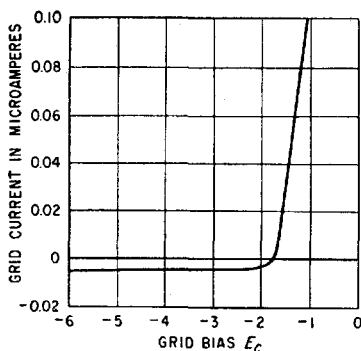
A compensating circuit of the above type is shown in Fig. 3.75a. The cathode follower $V2$ and the amplifier input stage $V1$ have a common cathode resistor R_k. The value of R_2 should be approximately equal to r_{p2}/μ_2 where r_{p2} is the dynamic plate resistance of $V2$ under the bias condition established by the value of R_1. In practice, it is usually found that the value of R_2 should be in the order of one to three times as large as R_1. In order to operate $V1$ and $V2$ at approximately the same values of plate current, it is usually necessary to make the value of the plate supply voltage for $V2$ much lower than the plate supply voltage for $V1$. In practice, the values of R_1 and R_2 are usually established experimentally, which accounts for the use of the
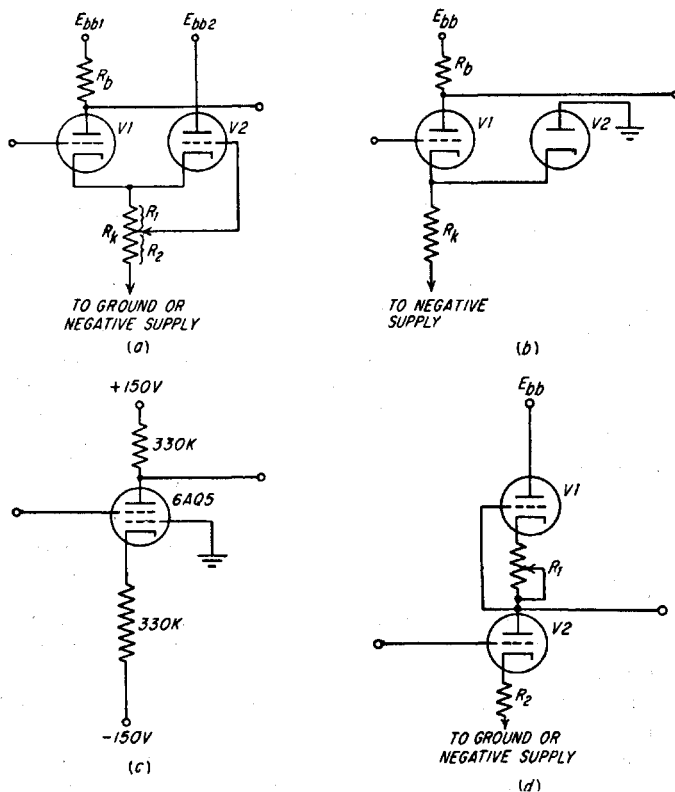

Fig. 3.75. Circuits which compensate for variations in heater voltage.

potentiometer. For best results, the tubes should be aged and the heater voltage cycled numerous times before the adjustment is made. A correct choice of values for R_1 and R_2 can be expected to reduce the effects of the heater voltage changes by a factor of 15 to 20. In the absence of the compensating tube, the equivalent grid voltage drift due to a 10 per cent heater voltage change will usually be in the order of 100 mv. Therefore, with proper compensation, the equivalent drift at the grid for a 10 per cent heater voltage change can be expected to be as small as 5 mv. A disadvantage in using this circuit is the significant reduction in the gain of $V1$ and the fact that the optimum adjustment of R_k changes with tube aging and tube replacements.

In Fig. 3.75b, a diode is used as the compensating tube. In order to maximize

the compensating voltage at the cathode of $V1$, the value of R_k should be large compared to the resistance changes in V_2 caused by heater voltage changes. For given supply voltages and bias on $V1$, the value of R_k must not be too large, otherwise $V2$ will be cut off. Larger values of R_k can be used as the value of the negative supply becomes more and more negative. The gain of the input stage will not be appreciably affected by using a larger value of R_k since the cathode circuit resistance is equal to R_k in parallel with the a-c resistance of the diode. With this particular circuit, the equivalent drift at the grid of $V1$ for a 10 per cent heater voltage change can be expected to be as low as 10 mv. The gain of $V1$ with respect to the uncompensated circuit will usually not be reduced as much as when the circuit shown in Fig. 3.75a is used. In an effort to achieve similar cathode characteristics, one-half of a dual triode is often used as the input tube and the other half is used as the diode. Best performance can be expected if $V1$ and $V2$ have the same cathode as in the diode-triode and diode-pentode tubes.

The operation of the circuit shown in Fig. 3.75c is similar to that of Fig. 3.75b since the control grid serves as a diode. For this application, a pentode with a relatively high screen grid to plate amplification factor is desirable. If the circuit values are correctly chosen, a 6V6 or a 6AQ5 will provide a stage gain comparable to that of a medium-μ triode with relatively good stabilization against the effects of heater voltage variations. For the circuit shown in Fig. 3.75c the gain is approximately 25, and the equivalent drift at the screen grid is approximately 5 mv for a 10 per cent change in heater voltage.

The circuit shown in Fig. 3.75d can be adjusted to provide complete cancellation of the equivalent grid voltage drift due to heater voltage changes. If $V1$ and $V2$ are identical, complete compensation and a gain equal to $\mu/2$ results when R_1 is equal to R_2. Since the characteristics of $V1$ and $V2$ will very likely be noticeably different, R_1 should have a value between 50 and 150 per cent of R_2 in order that it can be adjusted to provide complete compensation. As in the case of the circuit shown in Fig. 3.75a, the tubes should be aged and the heaters cycled numerous times before the final adjustment is made.

Differential amplifiers and bridge circuits (see Figs. 3.77 and 3.78) provide very high rejection to both filament voltage and power-supply voltage variations and are treated in conjunction with the circuits providing compensation for power-supply voltage variations.

Effect of Power-supply Variations. The requirement that the power supplies for a direct-coupled amplifier be well regulated becomes more stringent as the gain of the amplifier is increased. The actual degree of regulation required can be evaluated for a specific application only. For example, if a value is assigned to the permissible deviation of the amplifier.output voltage due to power-supply voltage changes, the allowable variation at the plate of the input stage can be established by dividing the assigned tolerance by the gain from the plate of the input stage to the output stage. This value can then be converted into the permissible power-supply voltage variation based on the characteristics of the input stage. The seriousness of power-supply variations can be appreciated by considering as an example a direct-coupled amplifier which has a gain of 200 from the plate of the input tube to the amplifier output. If a 1-volt variation in the power supply for the input stage should cause a 250-mv change in the plate voltage of the input tube, the amplifier output voltage change resulting from the 1-volt power-supply change would be equal to 200 × 0.25, or 50 volts.

The input circuit frequently consists of one of the circuits shown in either Fig. 3.75, 3.76, 3.77, or 3.78. The input circuits shown in Fig. 3.75 have previously been described and offer no special discrimination against power-supply voltage changes. The circuit shown in Fig. 3.76 is a differential amplifier with a single-ended output. $V2$ provides a moderate stabilizing effect against heater voltage changes and also

permits the use of a second input[1] without phase inversion.   This circuit has very
low gain for common mode (see Sec. 3.17) input signals on the grids.   The single-
ended output voltage, however, is sensitive to power-supply voltage changes.   The
common mode rejection ratio and the gain from the grids of $V1$ and $V2$ are given by

Eqs.   (3.110a),   (3.110b),   and   (3.110c),
respectively.   The gain A_{cm} for common
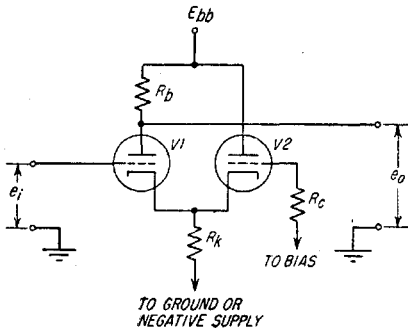mode   input   signals   can   be   determined
from Eq. (3.124).

$$A_{cm} = \frac{-\mu R_b}{r_p + R_b + BR_k(\mu + 1)} \quad (3.124)$$

where $B = 2 + R_b/r_p$

Input   circuits   which   provide   very   high
rejection   to   both   heater   voltage   changes
and   power-supply   voltage   changes   are
shown in Figs. 3.77 and 3.78.

In the input circuit shown in Fig. 3.77,
a   precision   voltage   divider   network   is   in
parallel   with   the   circuit   shown   in   Fig.
3.75d. R_3 in the divider is adjusted so

Fig.   3.76.   Differential   amplifier   with   a
single-ended output.

that the voltage at the tap on the voltage divider is equal to the quiescent voltage at
the plate of $V2$.   Increases or decreases in either the d-c supply voltage or the heater
voltage should not disturb this equality.   Since the two inputs of the differentia
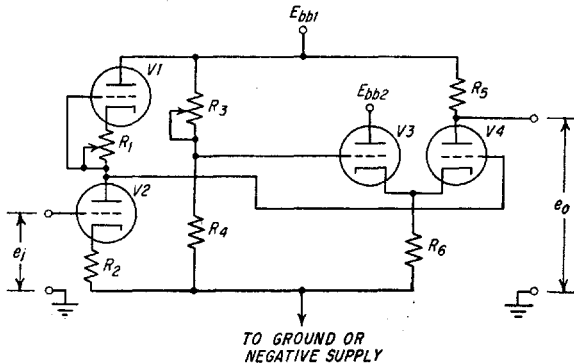

Fig.   3.77.   Input   circuit   which   is   stabilized   against   heater   and   supply   voltage   changes.
Circuit consists of a bridge circuit and a differential amplifier circuit having a single-ended
output.

amplifier are connected to the plate of $V2$ and to the tap on the voltage divider,
respectively, the identical voltage changes which appear at the inputs of the differen-
tial amplifier because of supply voltage changes are common mode signals.   As can
be seen from Eq. (3.124), the gain of the common mode input signals to this stage can
usually be considered to be insignificant.

A differential amplifier with a balanced output is used in the input stage of the
circuit shown in Fig. 3.78.   If the two halves of the input tube are identical, the
incremental voltage changes at the plates of the tube due to heater and supply voltage
changes will be equal and appear as common mode input signals to the following

[1] The second input is often used in conjunction with balancing (see the input stage of
Fig. 3.78) and, in a chopper stabilized amplifier, as the point at which the chopper amplifier
output is injected into the common amplifier (see Sec. 3.19c).

differential amplifier stage. Since the differential amplifier stage has very low gain for common mode input signals, the effects of the typical variations in regulated power supplies are usually insignificant. The same considerations apply to grid current changes provided the resistances in each grid circuit of the input tube are identical. Many of the premium twin-triode tubes are well balanced and, consequently, have particular value when employed in differential amplifiers. The *plate-to-plate* gain A of a balanced differential amplifier for the case where an input signal is applied to only one of the grids is given by Eq. (3.125).

$$A = \frac{\mu R_b}{r_p + R_b} \qquad (3.125)$$

where R_b = plate-load resistance

The gain to each plate is approximately equal to $A/2$ provided μR_k is large compared to $r_p + R_b$.

Effect of Changes in Components. With particular reference to the low-level stages, variations in tube characteristics and changes in resistor values are sources of undesired changes in the d-c output voltage. In critical applications, (1) the tubes should
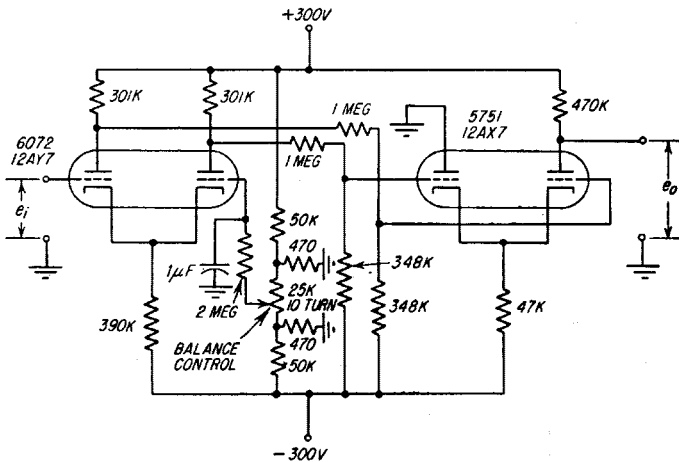

FIG. 3.78. Input circuit consisting of cascaded differential amplifiers to provide stabilization against heater and supply voltage changes.

be aged, (2) resistors having low temperature coefficients should be used, (3) the amplifier ambient temperature should be stabilized, and (4) the power dissipated in each resistor should not be more than approximately one-eighth to one-fourth the wattage rating of the resistor.

3.19b. Carrier D-C Amplifiers. In this type of amplifier, the input signal is filtered in a low-pass filter and then converted into an a-c signal, viz., a carrier modulated by the input signal, which is amplified in an a-c coupled amplifier. The amplified signal is rectified and then filtered in a low-pass filter to produce the output signal. If the bandwidth of the input low-pass filter is equal to or greater than one-half the carrier frequency, there will be ambiguities in the relationships between the input signal frequencies and the frequencies at which the carrier is modulated. For example, an input signal frequency f_i and all other input signal frequencies which are displaced from the fundamental and the harmonics of the carrier frequency by a frequency f_i will cause the carrier to be modulated at a frequency equal to f_i. Consequently, the input bandwidth is usually limited to some frequency less than one-half the carrier

frequency.   In addition, the bandwidth of the output low-pass filter must be made less than the carrier frequency, otherwise the carrier will appear in the output.   Since the carrier type of d-c amplifier does not employ direct-coupled amplifiers, the problems of drift normally encountered with multistage direct-coupled amplifiers do not exist.

The circuit shown in Fig. 3.79a represents one type of carrier d-c amplifier.   The input voltage level establishes the magnitude of the a-c voltage, i.e., the carrier, which is fed into the amplifier.   The output is obtained by rectifying and filtering the amplified carrier.

A carrier amplifier which employs two choppers is shown in Fig. 3.79b.   The choppers are usually operated at the power-line frequency.   One of the choppers is
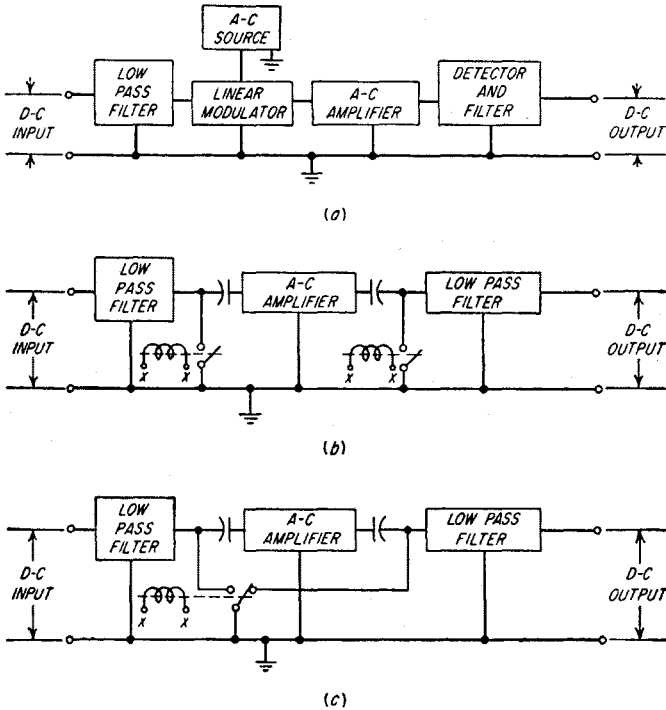

FIG. 3.79. Sample configurations of carrier type d-c amplifiers.

used to alternately switch the amplifier input between the d-c input signal level and ground potential.   In this manner, a square wave carrier is generated which has an amplitude and phase determined by the magnitude and polarity of the d-c input voltage.   The resulting square waves, after amplification, are synchronously rectified by switching the amplifier load circuit, viz., a low-pass filter, between the amplifier output and ground potential.   This is accomplished by means of the second chopper. For any given polarity of the d-c input signal, the polarity of the d-c output signal can be reversed by either adding another amplifier stage for phase reversal, by moving the amplifier output lead to the unused contact on the output chopper (see Fig. 3.79b), or by reversing the phase of the a-c driving voltage at the coil of one of the choppers. If an a-c voltage at the line frequency should be introduced on the input lead to the chopper contact, there will be a d-c component introduced in the output since the chopper is synchronous with the a-c voltage pickup.   For this reason, the chopper

coil and its leads are often shielded in order to minimize pickup at the chopper contact terminals.

The circuit in Fig. 3.79c is the more common type of chopper amplifier and requires a single-pole double-throw switching arm. It is necessary to use the proper number of amplifier stages to achieve the desired phase relation between the input and output signals. When high-gain a-c amplifiers are used, the chopper contacts should be make-before-break to minimize the switching transients introduced into the amplifier input which might otherwise introduce oscillations. If break-before-make contacts are used, it is usually necessary to use separate input and output choppers as shown in Fig. 3.79b to isolate the input and output stages.
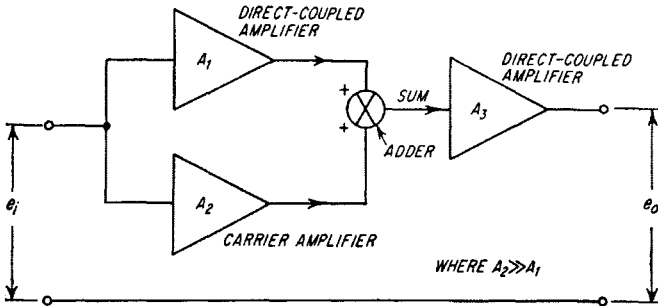

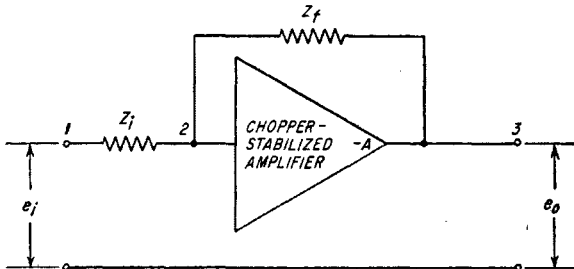FIG. 3.80. Block diagram of a typical chopper-stabilized amplifier. '

FIG. 3.81. Circuit configuration of an operational amplifier employing a chopper-stabilized amplifier.

3.19c. Chopper-stabilized Amplifiers. A chopper-stabilized amplifier is an amplifier which contains both a direct-coupled amplifier and a chopper amplifier. A block diagram of a typical configuration is shown in Fig. 3.80 where a chopper amplifier is placed in parallel with a direct-coupled amplifier which has considerably less gain. The d-c outputs of the two amplifiers are mixed and amplified in a common direct-coupled amplifier. The reason for the incorporation of the chopper amplifier is to provide increased stability against drift in the direct-coupled amplifier when over-all negative feedback is employed.

Chopper-stabilized amplifiers are often used as shown in Fig. 3.81. By the correct choice of components for Z_i and Z_f, specific mathematical operations, e.g., integration, differentiation, summation, etc., can be performed. Chopper-stabilized amplifiers with the associated input and output impedances so employed are referred to as operational amplifiers and are used in analog computers. In many operational amplifier applications utilizing chopper-stabilized amplifiers, there is a gain requirement from direct current to several kilocycles. Since reliable choppers have not, as yet, been made to operate at such high frequencies, it is not possible to use a chopper

amplifier only. The chopper amplifier, therefore, provides gain at low frequencies only, and the higher-frequency requirements are satisfied by the direct-coupled amplifier.

Drift in a Chopper-stabilized Amplifier. The particular function which an operational amplifier performs is dependent on the characteristics of the input impedance Z_i and the feedback impedance Z_f of Fig. 3.81. The accuracy and stability are dependent on the stability and precision of Z_i and Z_f and the drift in the d-c amplifier.

Consider a direct-coupled amplifier having a gain of $-1,000$ which has been balanced so that the output voltage is equal to zero when the input terminal to the amplifier is grounded. Assume, after balancing, that the heater voltage is changed sufficiently to cause the output voltage to change from 0 to -10 volts. The drift caused by the heater voltage change can, therefore, be expressed as an equivalent input voltage change Δe_e equal to $-10/-1,000$, or 0.01 volt. If d-c negative feedback is applied as shown in Fig. 3.82, the feedback will cause the input terminal voltage at point 2 to have a value Δe_i very nearly equal in magnitude to the equivalent input voltage change Δe_e except that it will have the opposite sign. It should be noted that negative feedback in a direct-coupled amplifier *does not* tend to reduce the input voltage to zero if the drift is due to any of the referenced sources. Instead, d-c negative feedback in a direct-coupled amplifier causes the input voltage to be
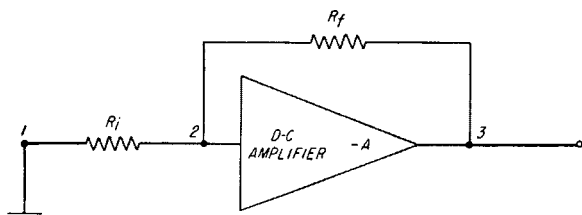

Fig. 3.82. D-C amplifier with negative feedback.

approximately equal to $-\Delta e_e$. Also, it follows that d-c feedback does not tend to reduce the output voltage offset caused by the referenced sources of drift to zero but to a value equal to $(R_i + R_f)\,\Delta e_i/R_i$. In contrast, a chopper amplifier with d-c negative feedback in a configuration as in Fig. 3.82 *tends* to reduce both the input terminal voltage (point 2) and the output voltage to zero; hence the term *automatic balancing* is often applied to chopper-stabilized amplifiers. Therefore, in operational amplifiers, a chopper amplifier is usually used in parallel with a direct-coupled amplifier as shown in Fig. 3.80. With reference to Fig. 3.80, the factor by which the chopper amplifier A_2 reduces the input and output voltages resulting from drift in the direct-coupled amplifier A_1A_3 is approximately equal to $A_1/(A_1 + A_2)$, where A_1 is the gain of the direct-coupled amplifier which parallels the chopper amplifier. For this reason, A_2 should be high compared to A_1.

Another major source of drift is grid current. With the proper choice of tubes and operating conditions, i.e., usually -2 to -3 volts bias, low-plate current, and low plate, screen, and filament voltages, the grid current can be fairly small, e.g., -10^{-4} μamp or smaller. Grid current flowing through the grid circuit impedance develops a d-c input voltage which is a source of error in an operational amplifier. D-C negative feedback around either a direct-coupled amplifier or a chopper amplifier, however, tends to reduce this input voltage to zero. The output voltage will be driven to a value approximately equal to the product of the total grid current and the feedback resistance.

To eliminate the problems associated with grid current flowing through the impedances Z_i and Z_f, the input to the direct-coupled amplifier A_1 is sometimes capacitively

coupled (see Fig. 3.86). This is especially desirable in the case of an integrator or a holding amplifier. If the chopper amplifier input and output low-pass filter time constants are established by Eqs. (3.126) and (3.127), respectively, the time constant of the added input capacitor C and grid resistor R should be $10/f_c$ or larger, where f_c is the chopper frequency. This time constant will place the low-frequency break point (Fig. 3.87) for the amplifier consisting of A_1 and A_3 below f_x. Precautions must be exercised in the use of capacitive coupling since the input capacitor can acquire an appreciable charge when the operational amplifier is driven to saturation. The result is that the input stage may be cut off for a relatively long period after the input voltage has been reduced to zero or reversed in polarity. A satisfactory solution consists of reducing the input voltage to zero and shorting out the input capacitor with a relay whenever it is necessary to reset the amplifier after an overload.
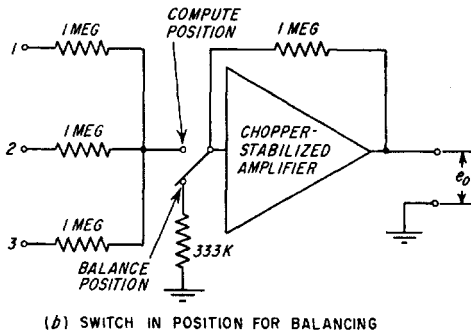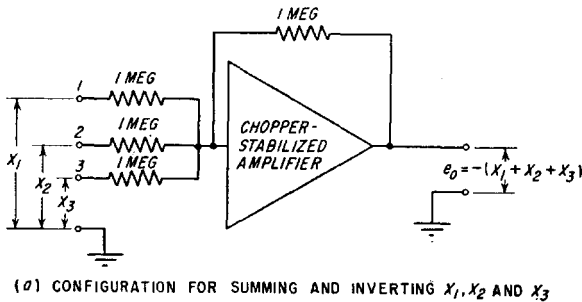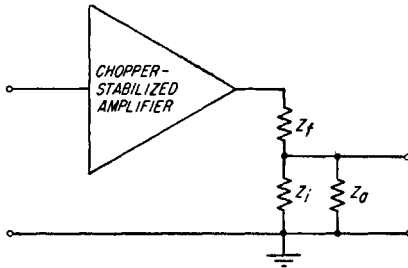

(a) CONFIGURATION FOR SUMMING AND INVERTING X_1, X_2 AND X_3

(b) SWITCH IN POSITION FOR BALANCING

FIG. 3.83. Summing and inverting operational amplifier.

Balancing an Operational Amplifier. The balance control is usually adjusted so that the amplifier output e_o is equal to zero when the input terminal, i.e., point 1 of Fig. 3.81, is grounded. Balancing is usually done just prior to the use of the amplifier so as to minimize the interval during which drift can occur. During the balancing adjustment, it is important that the grid circuit resistance have a value identical to that which it has when used in the computing operation so as to ensure that the amplifier will be properly balanced when in the compute condition. As an example, the summing amplifier shown in Fig. 3.83a can be balanced when points 1 through 3 are grounded, or, if more convenient, the grid circuit can be switched as shown in Fig. 3.83b for balancing.

In the process of balancing, a d-c offset voltage is introduced at some point in the direct-coupled amplifier so as to make the d-c output voltage equal to zero. The chopper section tends to maintain the output voltage equal to zero except for the drift resulting from grid current changes.

Stability Considerations. Since chopper-stabilized amplifiers are usually employed with feedback, they should be designed to minimize the external circuitry required to prevent oscillations or undesired transient behavior. Inverting, summing, integrating, and holding operational amplifiers present no particular problems as far as attaining over-all stability. However, if the *chopper-stabilized amplifier* has a phase shift in the order of $-90°$ in the frequency range where the *operational amplifier open-loop* (see Fig. 3.84) *gain* is equal to unity, it is not possible to design a stable differentiator utilizing a single capacitor for Z_i and a single resistor for Z_f (see Fig. 3.81). A differentiator of this type would oscillate since the feedback resistor and capacitor would introduce an additional phase lag of 90°. Therefore, in the case of a differentiator, it is necessary to include a lead network in the feedback path to reduce the open-loop phase shift to a value significantly less than 180° at the frequency where the open-loop gain is unity. The required phase lead can usually be obtained by either placing a resistor in series with the differentiator

WHERE Z_Q IS THE INPUT IMPEDANCE TO THE CHOPPER-STABILIZED AMPLIFIER

FIG. 3.84. Open-loop configuration for the stability analysis of the operational amplifier shown in Fig. 3.81.

capacitor or a capacitor in parallel with the feedback resistor as shown in Fig. 3.85. The considerations in the design of a chopper-stabilized amplifier having a maximum phase lag of approximately 90° in the frequency range where the amplifier gain is unity are discussed in the remainder of this section.
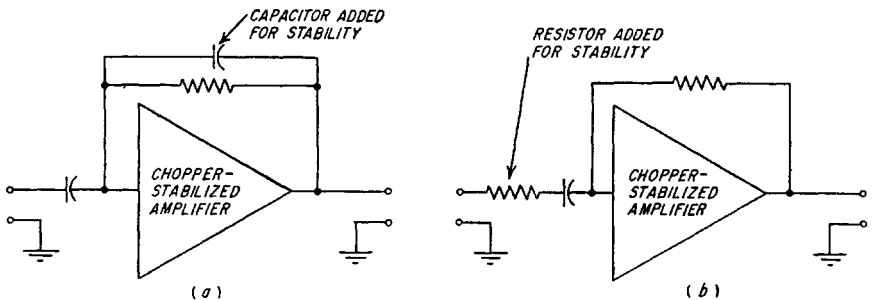

FIG. 3.85. Typical differentiating circuits with stabilizing components.

As shown in Fig. 3.80, the chopper amplifier is in parallel with a direct-coupled amplifier. The over-all gain through the chopper-stabilized amplifier is equal to $A_3(A_1 + A_2)$. In a typical amplifier, A_2 is much larger than A_1 at direct-current and at frequencies small compared to the chopper frequency. At higher frequencies, A_1 is larger than A_2. A partial wiring diagram of a typical amplifier is shown in Fig. 3.86, and a Bode plot (see Sec. 18.5b) in Fig. 3.87. The Bode plot is typical of the desired frequency response for a chopper-stabilized amplifier. This type of response can be achieved, in part, by establishing the chopper input and output time constants R_1C_1 and R_2C_2 (see Fig. 3.86) from Eqs. (3.126) and (3.127).

$$R_1C_1 \simeq \frac{1.6}{f_c} \qquad (3.126)$$

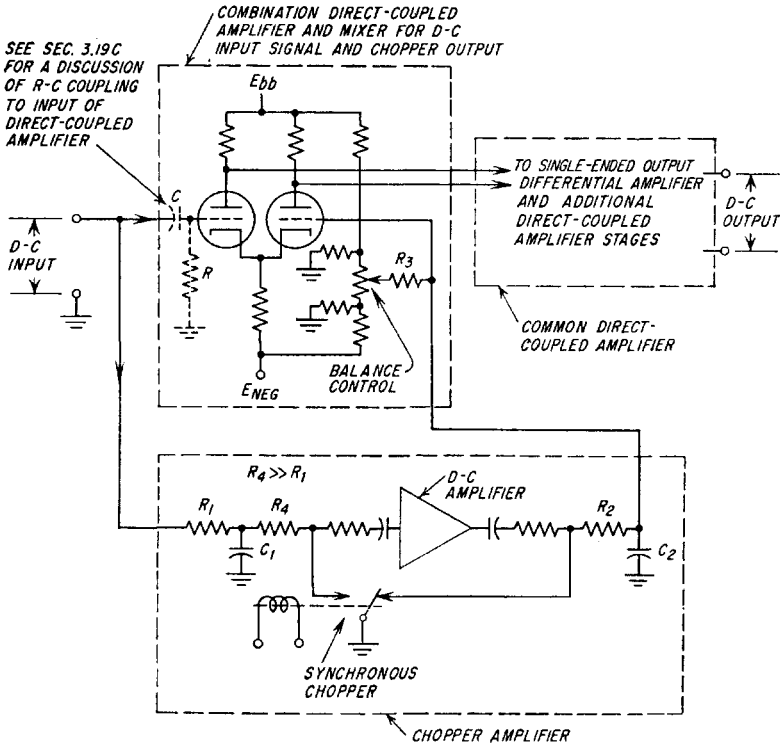$$R_2C_2 \simeq \frac{6.4A_2}{f_cA_1} \qquad (3.127)$$

FIG. 3.86. Schematic of a typical chopper-stabilized amplifier.

$$f_w \approx \frac{0.025\, f_c A_1}{A_2}$$

$$f_x \approx 0.025\, f_c$$

$$= \frac{A_2 f_w}{A_1}$$

$$f_y \approx 0.1\, f_c$$
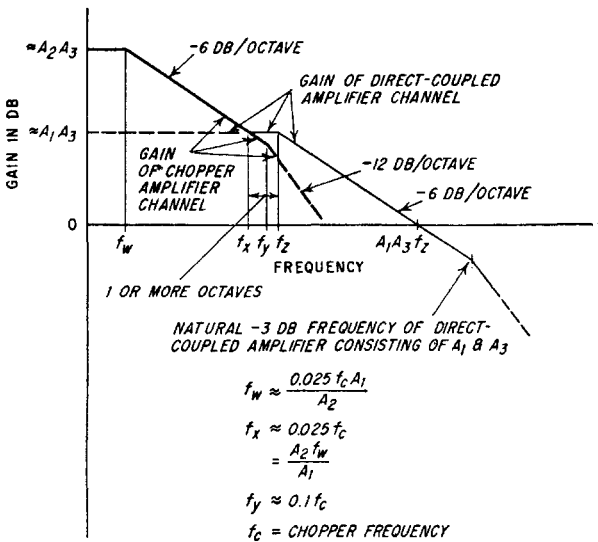
$$f_c = CHOPPER\ FREQUENCY$$

FIG. 3.87. Bode plot of a typical chopper-stabilized amplifier.

where A_1 = d-c gain of the direct-coupled amplifier paralleling the chopper amplifier

A_2 = d-c gain of the chopper amplifier

f_c = chopper frequency

The time constant R_2C_2 establishes the break frequency f_w in Fig. 3.87.   The break point at f_x is determined by making the time constant R_1C_1 less than $6.4/f_c$. Since Eq. (3.126) establishes R_1C_1 to be approximately equal to $A_1R_2C_2/4A_2$, the break point at f_x is assured.   If the recommended time constants for R_1C_1 and R_2C_2 are used, it can be assumed that the over-all amplifier consists only of the chopper amplifier A_2 and the direct-coupled amplifier A_3 for d-c input signals and input signals having frequencies below f_x.   However, for frequencies greater than f_x, the gain A_1A_3 will be greater than A_2A_3, and, consequently, the chopper section can then be disregarded.   The break frequency f_y is determined by the time constant R_1C_1 and, as established by Eqs. (3.126) and (3.127), is two octaves above f_x.   The break frequency f_z is dependent on the desired bandwidth of the amplifier.   In typical applications, this break frequency is made one or more octaves greater than f_x and is accomplished by introducing a lag circuit, e.g., a shunt capacitor across one of the plate-load resistors, in the common amplifier A_3.   With reference to the Bode plot shown in Fig. 3.87, the lag circuit in the common amplifier A_3 has a time constant equal to $0.8/f_c$.   The natural -3-db upper cutoff frequency for the amplifier consisting of A_1 and A_3 in series, viz., the -3-db upper frequency before the introduction of the lag circuit in the amplifier A_3, should be considerably higher than the frequency at which the over-all amplifier response curve crosses the 0-db gain line.   If this is not the case, a special lead-lag network will be required to minimize the phase shift in the frequency region where the amplifier gain is near unity.   The 0-db gain crossover frequency will be equal to $A_1A_3f_z$ provided the amplifier response curve has a slope of -6 db per octave from the frequency f_z to the 0-db gain frequency.

Power Amplifiers

4.1. Class A1 Single-ended Transformer-coupled Power Amplifiers. Class A1 transformer-coupled power amplifiers are used primarily in audio and video applications. In a class A1 amplifier, the input signal never causes the grid-to-cathode voltage to become positive. Consequently, there is no grid input power requirement. Transformer coupling to the load is ordinarily used in a power amplifier since it provides d-c isolation between load and tube and, by utilizing a transformer having the correct turns ratio, the optimum a-c plate-load resistance can be presented to the tube.

Typical operating conditions for power tubes used as class A1 single-ended transformer-coupled power amplifiers are published by the tube manufacturers. The data

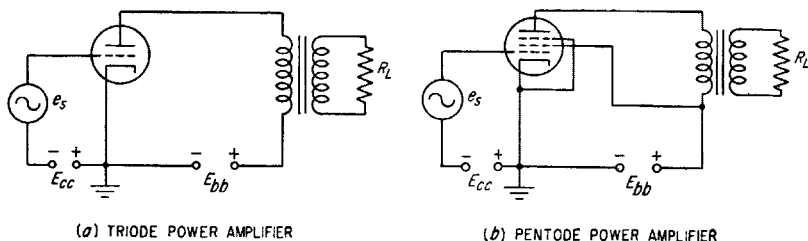(a) TRIODE POWER AMPLIFIER   (b) PENTODE POWER AMPLIFIER

FIG. 4.1. Basic power amplifier circuits.

usually include typical tube quiescent operating points and the associated signal voltages and currents, power outputs, and distortion percentages for different a-c load resistances. If the actual operating conditions do not correspond with the published typical conditions, a complete graphical analysis of the amplifier must be made to establish its performance.

The output stage of an amplifier is usually a power amplifier. It ordinarily requires a relatively large amount of d-c power and usually distorts the signal more than all the preceding stages combined. This distortion occurs because the plate current swing is so large that the tube is operated in nonlinear regions. Ordinarily, the a-c plate-load resistance and the quiescent operating point are selected to minimize the harmonic distortion and to maximize the output signal power. In determining the maximum power output for any given load line and operating point, the grid input signal is assumed to be adequate to drive the grid to either zero bias or to the value of bias which produces the maximum allowable distortion.

4.1a. Fundamental Power Relationships in the Plate Circuit of a Class A1 Power Amplifier. In Fig. 4.1 are triode and pentode single-ended transformer-coupled amplifiers having fixed bias. For each amplifier, the quiescent operating point is established by the intersection of the d-c load line and the grid-bias curve which has a value equal to the bias supply voltage[1] E_{cc}. The d-c load line is drawn from the plate supply voltage E_{bb} on the abscissa and has a slope of $-1/R_p$ where R_p is the resistance of the transformer primary winding. Neglecting transformer core losses, the a-c plate-

---

[1] If self-bias is employed, the quiescent point for a triode amplifier is established by the intersection of the bias line and the d-c load line (see Sec. 3.3a), and, for a tetrode or pentode amplifier, the quiescent point is established by the intersection of the bias line and the d-c transfer characteristic (see Sec. 3.4).

4–2

load resistance r_b is given by Eq. (4.1). If self-bias is used, it is assumed that the cathode resistor is adequately bypassed at signal frequencies.

$$r_b = R_p + \left(\frac{N_p}{N_s}\right)^2 (R_s + R_L) \qquad (4.1)$$

where R_s = secondary winding resistance
R_L = secondary load resistance
N_p = number of primary turns
N_s = number of secondary turns

The a-c load line is drawn with a slope equal to $-1/r_b$ and passes through the quiescent operating point only if the tube characteristics are perfectly linear. If nonlinearities exist, the amount of the displacement of the a-c load line from the quiescent operating point is proportional to the signal amplitude, the degree of nonlinearity, and the d-c load resistance (primary winding resistance and cathode resistance if self-bias is employed). The procedure for plotting the a-c load line (Sec. 3.3b) is identical for triodes, tetrodes, and pentodes. From the d-c and a-c load lines, it is possible to determine (1) the amplitudes of the signal harmonics, (2) the required d-c input power to the plate circuit, (3) the signal output power, (4) the plate circuit efficiency, and (5) the plate power dissipation with and without a signal. The plate circuit efficiency is defined as the ratio of the output power at the fundamental frequency to the d-c power supplied to the plate circuit expressed as a percentage.

Determination of Harmonic Distortion. To make a harmonic analysis, it is necessary to accurately construct the a-c load line and the dynamic transfer characteristic. For the 2nd, 3rd, and 4th harmonics, the method described in Sec. 3.3b permits the determination of the peak value of each harmonic and the percentage that each harmonic is of the fundamental.

Plate Circuit D-C Input Power P_{dc}. The d-c input power P_{dc} to the plate circuit can be determined from Eq. (4.2).

$$P_{dc} = E_{bb}I_{ba} \qquad (4.2)$$

where E_{bb} = d-c plate supply voltage
I_{ba} = average value of plate current and can be calculated from (Eq. 3.7) (I_{ba} = quiescent plate current I_{bo} if there is no input signal or if there is no distortion)

Signal Output Power P_o. Assuming no distortion, the power output P_o of a single-ended class A1 amplifier having a sinusoidal input signal can be determined from Eq. (4.3).

$$P_o = 0.125[e_{b(max)} - e_{b(min)}][i_{b(max)} - i_{b(min)}] \qquad (4.3)$$

where $e_{b(max)}$ = maximum plate voltage with signal
$e_{b(min)}$ = minimum plate voltage with signal
$i_{b(max)}$ = maximum plate current with signal
$i_{b(min)}$ = minimum plate current with signal

If distortion exists, the power output due to any harmonic component n can be determined by evaluating Eq. (4.4).

$$P_{on} = I_{Hn}^2 r_b \qquad (4.4)$$

where I_{Hn} = rms value of the nth harmonic
P_{on} = output power due to the nth harmonic

The rms values of the 2nd, 3rd, and 4th harmonics can be determined by multiplying the peak values given by Eqs. (3.8), (3.10), and (3.12) by 0.707. The rms value of the fundamental I_{H1} is given by Eq. (4.5).

$$I_{H1(\text{rms})} = 0.236(i_{b(\text{max})} + i_x - i_y - i_{b(\text{min})}) \tag{4.5}$$

where $i_{b(\text{max})}$, i_x, i_y, and $i_{b(\text{min})}$ are defined in Sec. 3.3b

The total output power is the sum of the power output terms due to the fundamental and each of the harmonics.

Plate Circuit Efficiency η. The plate circuit efficiency η is given by Eq. (4.6).

$$\eta = \frac{P_{o1} \times 100}{P_{dc}} \tag{4.6}$$

where P_{o1} = signal output power due to fundamental component of plate current
P_{dc} = d-c input power to plate circuit with signal present

If there is no distortion, P_{o1} can be determined from Eq. (4.3) and P_{dc} will be equal to $E_{bb}I_{bo}$. Under these conditions the theoretical maximum value of η is equal to 50 per cent and is obtained when $e_{b(\text{min})}$ and $i_{b(\text{min})}$ are equal to zero and $e_{b(\text{max})}$ and $i_{b(\text{max})}$ are equal to twice E_{bb} and I_{bo}, respectively. In practice, the efficiencies of class A1 power amplifiers are between 15 and 35 per cent. Greatest efficiencies are obtained with tetrodes and pentodes.

In some instances the over-all efficiency of a power amplifier stage is desired. In this case, the input power to the screen-grid circuit and the filament power must be included with the d-c input power to the plate circuit.

Plate Power Dissipation. The quiescent plate power dissipation P_{bo} is given by Eq. (4.7).

$$P_{bo} = E_{bb}I_{bo} - I_{bo}{}^2(R_p + R_k) \tag{4.7}$$

where R_k = cathode resistor (if used)
I_{bo} = quiescent plate current
R_p = resistance of the transformer primary winding

If a signal is present, the plate power dissipation P_{bs} can be determined from Eq. (4.8).

$$P_{bs} = E_{bb}I_{ba} - P_o - I_{ba}{}^2(R_p + R_k) \tag{4.8}$$

where I_{ba} = average plate current with the signal present
P_o = total signal output power, i.e., power due to fundamental and all harmonic components of plate current

Ordinarily, P_{bs} will be 15 to 35 per cent less than P_{bo}. In class A1 operation, the tube actually operates at a lower plate dissipation with the signal applied than it does in the absence of a signal.

4.1b. Triode Class A1 Single-ended Transformer-coupled Power Amplifiers. The harmonic distortion in class A1 single-ended triode amplifiers is almost entirely 2nd harmonic. If the distortion is assumed to be entirely due to the 2nd harmonic, the per cent distortion can be determined from Eq. (4.9).

$$\text{Per cent } H_2 = \frac{0.5(p/n - 1) \times 100}{p/n + 1} \tag{4.9}$$

where p = value of the positive plate current peak = $i_{b(\text{max})} - I_{bt}$
n = value of the negative plate current peak = $I_{bt} - i_{b(\text{min})}$
I_{bt} = plate current on a-c load line associated with bias which exists when a signal is present (see Sec. 3.3b)

Equation (4.9) can be derived from Eq. (3.9) provided i_x is halfway between $i_{b(max)}$ and I_{bt} and if i_y is halfway between I_{bt} and $i_{b(min)}$, as is the case when the 3rd harmonic is equal to zero. The ratio p/n can readily be established from the a-c load line.

A compromise between maximum power output and minimum harmonic distortion can be obtained by careful selection of the operating point and the a-c load resistance r_b. In general, the optimum a-c load resistance for a triode class Al single-ended transformer-coupled amplifier can be determined from Eq. (4.10).

$$r_b \simeq 2r_p \qquad (4.10)$$

The curves shown in Fig. 4.2 illustrate the manner in which the relative power output, per cent distortion, and ratio r_b/r_p are related in a typical triode class Al single-ended transformer coupled amplifier. They show that for a given value of distortion the maximum power output is obtained when the a-c load resistance r_b is approximately equal to $2r_p$. Associated with each value of the a-c load resistance, however, is also the selection of the optimum operating point.

FIG. 4.2. Curves showing power output and total distortion as a function of the ratio of plate-load resistance to dynamic plate resistance for a typical power triode.

The optimum operating conditions are obtained by a series of experimental graphical solutions. For a given supply voltage E_{bb}, equal to or less than the maximum rated
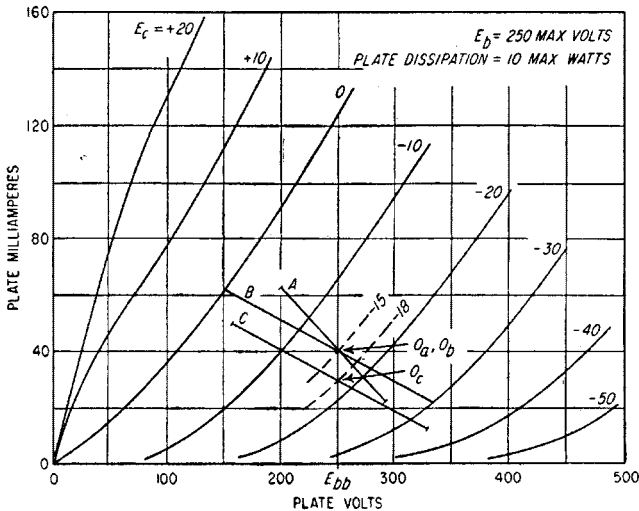

FIG. 4.3. Typical power triode characteristics with three different a-c load lines, each of which causes 5 per cent second harmonic distortion.

voltage for the tube, locate on the tube characteristics the plate current which causes the maximum rated plate dissipation. This procedure establishes point O_a in Fig. 4.3. At this point, the a-c plate resistance is equal to 2,250 ohms. To illustrate how the fundamental power output varies as a function of the ratio of the load resistance to the dynamic plate resistance, assuming a given value of distortion, three load lines

TABLE 4.1. CALCULATED VALUES OF POWER OUTPUT AND HARMONIC DISTORTION
FOR THE A-C LOAD LINES PLOTTED IN FIG. 4.3

Load line	A-C load resistance, ohms	Fundamental power output, watts	2nd harmonic, per cent	Efficiency, per cent
A	2,250	0.45	5	4.5
B	4,500	0.95	5	9.5
C	4,500	0.80	5	11.2

have been constructed. Load line *A* has been drawn through point O_a for an a-c
plate load of 2,250 ohms; therefore, $r_b/r_p = 1$. Load lines *B* and *C* have been drawn
for $r_b/r_p = 2$. The length of each load line indicates the plate current excursion
which will cause the 2nd harmonic to be equal to approximately 5 per cent of the
fundamental. From Table 4.1, it can be seen that for a total of 5 per cent 2nd har-
monic distortion, considerably more output power can be obtained when the a-c plate-
load resistance is twice as large as the dynamic plate resistance. The tabulated data
for load line *C* indicate that it may be desirable to consider operation at plate current
values less than the maximum rated value provided comparative values of output
power and improved efficiency can be obtained. If additional load lines and operating
points are evaluated, a family of curves such as those shown in Fig. 4.2 can be con-
structed, thereby establishing the optimum operating conditions.
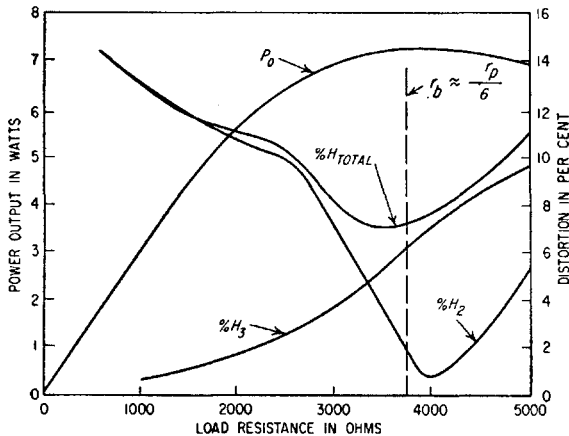

FIG. 4.4. Power output and distortion curves for a class A1 single-ended amplifier using a
6L6 beam power tube with $E_{bb} = E_{sg} = 250$ volts, $E_c = -14$ volts, and peak grid signal
voltage $= 14$ volts. At the quiescent point, $r_p = 22,500$ ohms.

*4.1c. Beam Tetrode and Pentode Class A1 Single-ended Transformer-coupled Power
Amplifiers.* The principal advantages in using beam tetrodes and pentodes instead
of triodes are that higher efficiencies can be obtained and less grid-drive voltage is
required per unit of output power. These tubes are more efficient than triodes since
the shapes of their E_b-I_b characteristics permit the path of operation to be extended
more nearly toward zero plate voltage when the grid is driven to zero bias. Typical
efficiencies of beam tetrode and pentode class A1 single-ended transformer-coupled
power amplifiers are between 35 and 45 per cent.

The power output and distortion curves of Fig. 4.4 are typical of single-ended

class A1 beam tetrode and pentode power amplifiers and differ appreciably from those obtained from triodes in similar applications. The power output curve in Fig. 4.4 indicates that the maximum power output for a 6L6 tube is realized when r_b is approximately equal to $r_p/6$. The total harmonic distortion percentage has a minimum value when r_b has a value slightly less than $r_p/6$. In this particular example, the optimum value of r_b is approximately equal to that value which minimizes the overall distortion since the power output under the minimum distortion conditions is approximately equal to the maximum possible power output. A general relationship for optimum r_b for beam tetrodes and pentodes is given by Eq. (4.11).

$$\frac{r_p}{12} \leq r_b \leq \frac{r_p}{6} \qquad (4.11)$$

where r_p = dynamic plate resistance of tube at quiescent operating point

In addition, the optimum r_b is very nearly established by the a-c load line which passes through the quiescent plate current, as determined by the maximum plate voltage and maximum plate dissipation ratings, and the knee of the zero bias curve.

The distortion in beam tetrode and pentode class A1 single-ended power amplifiers is appreciably greater than for triodes used in the same type of operation and is primarily due to the 3rd harmonic when operated in the region of maximum power output. Figure 4.4 shows that the 2nd harmonic reaches a minimum value when r_b is approximately equal to the value which produces maximum power output.

Precautions should be taken to avoid excessive screen dissipation. The average screen current can be determined by applying Eq. (3.7) to the dynamic screen current. The screen dissipation is equal to the product of the average screen current and the screen grid-to-cathode voltage.

Example 4.1

A 6L6 is to be operated as a class A1 single-ended power amplifier with fixed bias and with E_{sg} and E_{bb} both equal to 250 volts. Assuming that the d-c voltage drop in the output transformer is negligible, determine the quiescent and dynamic operating points, a-c plate-load resistance, power output, total harmonic distortion, and efficiency. The maximum allowable plate dissipation is 19 watts, and the transformer secondary load impedance is equal to 600 ohms.

1. Locate the quiescent operating point.

$$I_{bo} = \frac{P_{bo}}{E_{bo}} = \frac{19}{250} = 76 \text{ ma}$$

From Fig. 4.5, the fixed bias is found to be equal to -13.7 volts.

2. Construct a tentative a-c load line from the knee of the zero bias curve through I_{bo} to $-2E_c$, that is, -27.4 volts.

3. Determine the shift in the a-c load line from the quiescent value due to the presence of the signal (see Sec. 3.3b).

A tentative value for the average plate current I_{ba} is calculated based on the tentative a-c load line.

$$\begin{aligned} I_{ba} &= 0.167(i_{b(\text{max})} + 2i_x + 2i_y + i_{b(\text{min})}) \\ &= 0.167(155 + 236 + 82 + 15) \\ &= 81.5 \text{ ma} \end{aligned}$$

Assuming I_{ba}, as established from the tentative a-c load line, to be correct, relocate the a-c load line to pass through the calculated value of I_{ba}. Recalculating I_{ba} based on the shifted a-c load line establishes I_{ba} to be approximately 82 ma. Since the second calculated value of I_{ba} very nearly agrees with the value of I_{ba} through which the a-c load line has been redrawn, it can be assumed that the new position of the load line is very nearly correct. Since the bias is fixed, the bias does not change with a change in the average plate current. Consequently, the intersection of the shifted a-c load line and the -13.7-volt bias line establishes the time axis of the signal. At this point the plate current I_{bt} is approximately 76 ma.
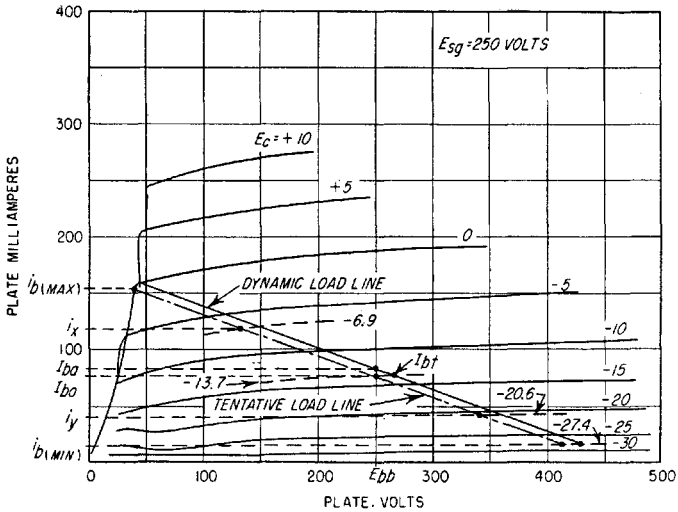
FIG. 4.5. Average plate characteristics for a 6L6 showing load lines for Class A power amplification.

4. Determine r_b from the slope of the a-c load line.
From Fig. 4.5, the slope is $-3.64 \times 10^{-4} = -1/r_b$.
The value of r_b is, therefore, 2,750 ohms.
5. Determine the power output in the fundamental using Eqs. (4.5) and (4.4).

$$I_{H1(rms)} = 0.236(158 + 120 - 42 - 15) = 0.0522 \text{ amp}$$
$$P_{o1} = (0.0522)^2 \times 2,750$$
$$= 7.49 \text{ watts}$$

6. Determine the total harmonic distortion.

$$\%H_T = \sqrt{(\%H_2)^2 + (\%H_3)^2 + (\%H_4)^2}$$

The values of $\%H_2$, $\%H_3$, and $\%H_4$ can be calculated from Eqs. (3.9), (3.11), and (3.13), respectively. Their values are as follows:

$$\%H_2 = 8.15$$
$$\%H_3 = 2.94$$
$$\%H_4 = 2.83$$

Therefore, $\%H_T = 9.13$.
7. Determine the efficiency, using Eq. (4.6).

$$\eta = \frac{P_{o1} \times 100}{P_{dc}}$$
$$= \frac{7.49 \times 100}{250 \times 0.082}$$
$$= 36.5\%$$

8. Determine the required transformer characteristics.
The transformer must be designed to have a 2,750-ohm primary winding and a 600-ohm secondary winding, to operate with approximately 82 ma direct current through the primary winding and to have a power rating of at least 7.5 watts. In addition, the desired frequency response and maximum distortion which the transformer may introduce should be specified.

4.2. Classes A, AB, and B Push-pull Transformer-coupled Power Amplifiers. Two typical push-pull transformer-coupled power amplifier circuits are shown in Fig. 4.6. In Fig. 4.6a, fixed bias is used, and the circuit shown in Fig. 4.6b has self-bias. The

input signal can be supplied from a transformer, a push-pull stage, or a phase inverter (see Figs. 3.60 et seq. for typical phase inverter circuits).

A push-pull transformer-coupled power amplifier has several advantages over a single-ended amplifier. The most important advantages are (1) elimination of d-c saturation in the output transformer, (2) no generation of even-order harmonics, (3) higher efficiency for the same per cent distortion, and (4) the cancellation of power-supply ripple in the output transformer, thereby minimizing the filtering required in the power supply.

1. Effect of the D-C Plate Current through the Primary Winding of the Output Trans-former. The magnetizing force due to the quiescent plate current through one-half of the primary winding directly opposes the magnetizing force due to the quiescent
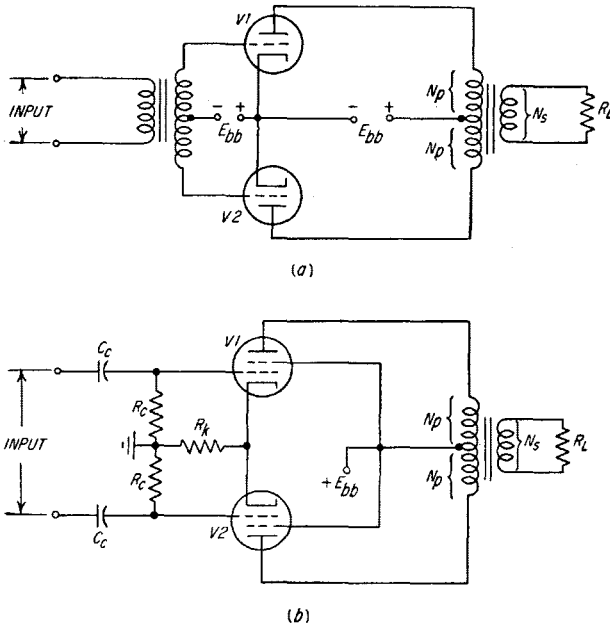
FIG. 4.6. Typical push-pull transformer-coupled power amplifier configurations.

plate current through the other half of the primary winding. If the tubes have equal quiescent currents and if the transformer has been properly center-tapped, good cancellation will be achieved. As a result, the output transformer for a push-pull amplifier can be made smaller than one designed to have the same power output and frequency response for use with a single-ended amplifier.

2. Harmonic Components. If the tubes are matched, a push-pull transformer-coupled amplifier will introduce no even harmonics. However, if the input is distorted and contains even harmonics, these harmonics, and any odd harmonics which may also exist, will be amplified and appear in the output. It is only the even harmonics caused by distortion in the output tubes which are canceled.

3. Efficiency. Since only odd harmonics are introduced by push-pull transformer-coupled amplifiers, it follows that the tubes can be driven harder than in single-ended amplifiers for the same total per cent distortion. Driving the tubes harder produces an increase in the output power and results in an increase in efficiency.

4. Power-supply Ripple-voltage Requirements. If the two tubes have identical dynamic plate resistances, the power-supply ripple voltage will cause equal a-c cur-

rents to flow in each tube. Because these currents flow in opposite directions through the transformer primary winding, no voltage is induced in the output winding. If the tubes are matched at quiescence, power supplies may be used which have considerably more ripple voltage than would be acceptable for single-ended amplifiers. The presence of a signal, however, causes the instantaneous plate resistance of one tube to increase and the instantaneous plate resistance of the other tube to decrease, thereby causing an induced voltage to appear in the output due to the power-supply ripple voltage. If the power-supply ripple voltage does not cause an objectionable signal to be induced in the output at quiescence, the ratio of the desired output signal to the undesired induced signal will usually be sufficiently large that the undesired signal will be negligible compared to the desired signal.

In Fig. 4.6, the midband plate-to-plate resistance R_{pp} presented to $V1$ and $V2$ by the terminated output transformer is given by Eq. (4.12) provided the transformer exciting current is neglected.

$$R_{pp} = R_{p1} + R_{p2} + 4(R_L + R_s) \left(\frac{N_p}{N_s} \right)^2 \qquad (4.12)$$

where N_p = number of turns in one-half primary winding
N_s = number of turns in secondary winding
R_{p1} = resistance of primary winding connected to $V1$
R_{p2} = resistance of primary winding connected to $V2$
R_L = terminating resistance
R_s = resistance of secondary winding

Neglecting the transformer winding resistances, Eq. (4.12) reduces to Eq. (4.13).

$$R_{pp} = 4R_L \left(\frac{N_p}{N_s} \right)^2 \qquad (4.13)$$

The analysis of a push-pull transformer-coupled power amplifier is complicated by the fact that the instantaneous plate voltage across each tube varies not only as a result of the variations in plate current through that tube but also because of the inductive coupling of the plate voltage variations existing at the plate of the other tube. However, general relationships in regard to the loads presented to the individual tubes can be stated. If $V1$ and $V2$ have equal instantaneous dynamic plate resistances, each tube will operate into an a-c load resistance equal to $R_{pp}/2$. If either $V1$ or $V2$ is at cutoff, the other tube will operate into a resistance equal to $R_{pp}/4$. Whenever the dynamic plate resistances of the two tubes are not the same, the tube having the smaller dynamic plate resistance will operate into an a-c load resistance having a value between $R_{pp}/2$ and $R_{pp}/4$ and the other tube will operate into an a-c load resistance having a value between $R_{pp}/2$ and infinity. Consequently, in a push-pull amplifier, neither tube operates into a fixed load resistance since the instantaneous values of tube dynamic plate resistance are a function of the instantaneous values of E_b and I_b which change with the instantaneous values of the signal. The actual analysis, however, can be accomplished in a straightforward manner by constructing the *composite characteristics* for the two tubes (see Sec. 4.2a). The composite characteristics are a composite E_b-I_b plot which takes into account the two tubes when used in push-pull operation, thereby permitting the construction of a single load line for the two tubes.

4.2a. Class A1 Push-pull Transformer-coupled Power Amplifiers Using Triodes. In class A1 push-pull operation, the common cathode resistor used for self-biasing is usually not bypassed. If the tubes are balanced, the total cathode current will be equal to the sum of the d-c and the even-order harmonic currents through both tubes. There will be no fundamental signal component across the biasing resistor. Con-

sequently, the signal that does exist across the cathode resistor consists primarily of a d-c and a 2nd harmonic term.

For a given distortion percentage, a triode class A1 push-pull transformer-coupled amplifier can be designed to have more than twice the power output obtainable from a single-ended amplifier employing the same tube type. The distortion in triode amplifiers is predominantly 2nd harmonic and will appear in the output of a single-ended amplifier but will be canceled in a push-pull transformer-coupled amplifier.
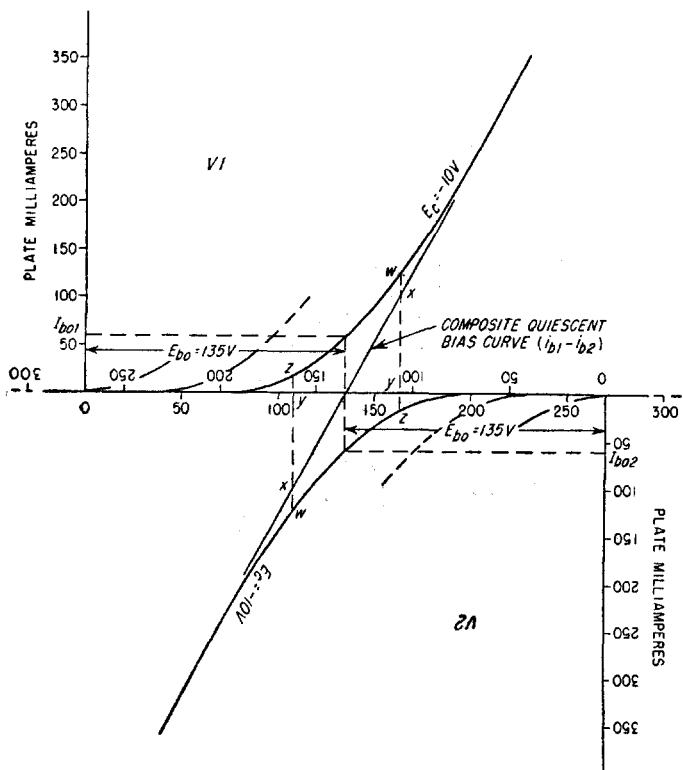
Fig. 4.7. Method of constructing the composite characteristics for a push-pull transformer coupled amplifier which utilizes triodes. Only the quiescent bias curves for $V1$ and $V2$ have been shown in order to simplify the figure.

The efficiency for a triode class A1 push-pull transformer-coupled amplifier is usually between 30 and 45 per cent.

Composite Characteristics. The construction of the composite characteristics for a push-pull transformer-coupled amplifier consists of properly positioning the individual E_b-I_b characteristics and plotting the combined E_b-I_b characteristics, thereby permitting the construction of a single load line for the two tubes. One of the tube characteristics is inverted and positioned relative to the other so that the individual quiescent plate voltage values are coincident. Consider the example in which the supply voltage is equal to 145 volts, the quiescent plate current through each tube is equal to 60 ma, and the resistance of each half of the primary winding is equal to 167 ohms. The quiescent IR drop in each half of the primary winding in this case is equal to 10 volts, and the quiescent plate voltage on each tube is 145 − 10, or 135 volts. It is, therefore, necessary to position the individual tube characteristics as shown in Fig. 4.7.

Only the quiescent bias curves for $V1$ and $V2$ have been shown to simplify the figure. If self-bias were employed, the quiescent plate voltage would be equal to 135 volts minus the bias voltage.

Having properly positioned the tube characteristics, the first step is to construct the composite quiescent bias line. This line represents the instantaneous differences in plate currents which would flow through $V1$ and $V2$ if the bias on each tube were maintained equal to the quiescent value and the plate voltages on the two tubes varied in equal incremental steps in opposite directions. The line is drawn so that it is
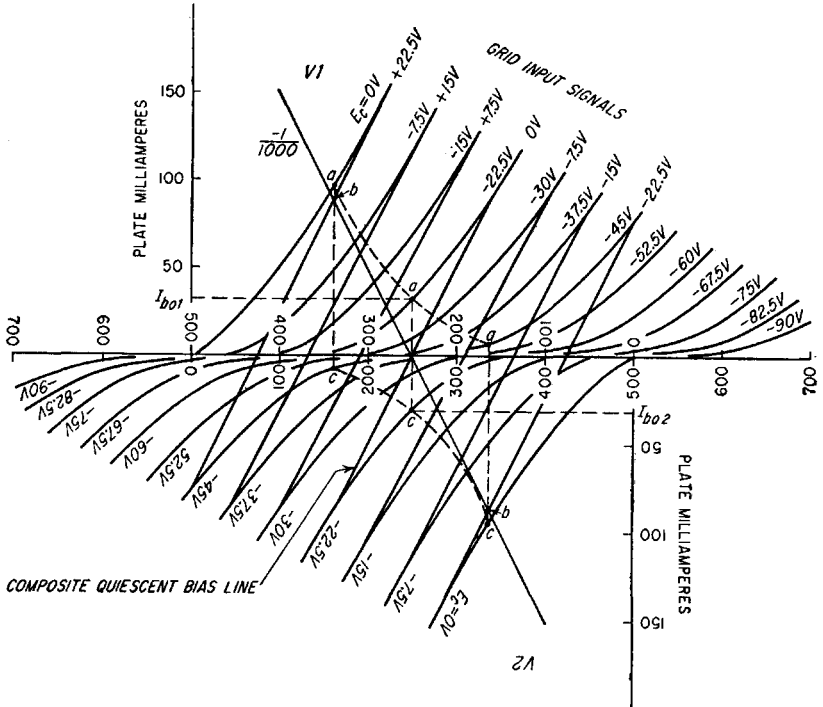
Fig. 4.8. Composite characteristics for push-pull 6L6's triode connected with $E_{bo} = 250$ volts, $E_c = -22.5$ volts, and $R_{pp} = 4,000$ ohms.

equal to $i_{b1} - i_{b2}$ for all the resulting values of plate voltage. The first point is at the quiescent plate voltage where i_{b1} is equal to I_{bo1} and i_{b2} is equal to I_{bo2}; hence, $i_{b1} - i_{b2}$ at this point is equal to zero. The other points are located so that the current yx is equal to $yw - yz$. The line connecting these points is the composite quiescent bias line.

The complete composite characteristics for two 6L6 triode-connected tubes operated in push-pull are shown in Fig. 4.8. The other composite lines have been constructed in a manner similar to the method used for constructing the composite quiescent bias line. For example, consider the composite line labeled +7.5 volts. It has been assumed that the bias on $V1$ has been decreased by 7.5 volts and the bias on $V2$ has been increased by 7.5 volts. If these bias voltages are maintained and the respective plate voltages varied in equal incremental steps in opposite directions, the instantaneous differences in plate currents will be defined by the line labeled +7.5 volts. It is

apparent that the composite characteristics are much straighter and more nearly equidistant than the individual tube characteristics.

Composite A-C Load Line. The composite a-c load line is constructed with a slope equal to $-4/R_{pp}$ [see Eqs. (4.12) and (4.13)] and passes through the composite quiescent bias line at the zero current point. In Fig. 4.8 the value of R_{pp} is equal to 4,000 ohms, and the composite a-c load line has been plotted with a slope equal to $-1/1,000$.

The composite a-c load line defines the composite path of operation for the two tubes, and the two curved dashed lines describe the paths of operation for the individual tubes. The curvature is due to the transformer coupling between the two tubes, and the slope at any given point indicates the instantaneous a-c load resistance into which the tube is operating. The individual paths of operation are determined by constructing vertical lines through the points at which the composite bias lines intersect the a-c load line, e.g., see points labeled *b* in Fig. 4.8. The dashed lines can then be constructed by connecting the points defined by the intersections of the vertical lines and the individual bias lines associated with the composite bias lines, e.g., see points labeled *a* in Fig. 4.8. For class A operation, the quiescent bias for each tube must not be sufficiently negative for the individual plate currents (dashed lines) to go to zero for any portion of the cycle. The quiescent bias is not critical, and relatively large changes in bias do not significantly alter the linearity of the composite bias lines. The considerations for establishing the quiescent bias are plate current cutoff during the cycle as one extreme (too much bias) and excessive plate power dissipation as the other extreme (insufficient bias).

The composite bias lines shown in Fig. 4.8 have been labeled in terms of input signal amplitudes. To determine the output signal peak amplitude for a given input signal, consider the example in which the input signal at the grid of each tube is a sine wave having a peak amplitude of 22.5 volts. Along the composite a-c load line, the composite path of operation will be between the points at which the load line intersects the composite bias lines labeled +22.5 and −22.5 volts. The plate-to-plate peak output signal amplitude is, therefore, equal to 343 − 157, or 186 volts for a 22.5-volt input signal since the plate voltage on one tube increases to 343 volts as the plate voltage on the other tube decreases to 157 volts.

The optimum composite load line for a triode push-pull transformer-coupled amplifier, i.e., the load which maximizes the power output and minimizes the harmonic distortion, has a slope approximately equal to the negative of the slope of the composite characteristics. Consequently, the optimum a-c plate-to-plate load resistance is approximately equal to four times the inverse of the slope of the composite characteristics. The approximate value of the optimum a-c plate-to-plate load resistance can also be determined from Eq. (4.14).

$$R_{pp} \simeq 2r_p \tag{4.14}$$

where r_p = dynamic plate resistance of each tube at quiescent operating point

Harmonic Distortion and Power Output. The distortion in a push-pull transformer-coupled amplifier is due primarily to the 3rd and 5th harmonics since the even harmonics are equal to zero. The rms value of the fundamental output current and the percentage values of the 3rd and 5th harmonic currents can be determined from Eqs. (4.15), (4.16), and (4.17), respectively. These equations are based on the instantaneous composite plate current values resulting from the instantaneous grid signal values shown in Fig. 4.9. These particular values of the grid-input signal have been chosen to maximize the accuracy in determining the value of the fundamental and the percentages of the odd harmonics.

$$I_{H1} = \frac{7i_1 + 6i_{0.8} + 10i_{0.5} - 6i_{0.3}}{15\sqrt{2}} \tag{4.15}$$

$$\% \; H_3 = \frac{0.33(i_1 - 2i_{0.5})}{\sqrt{2}\, I_{H1}} \times 100 \tag{4.16}$$

$$\% \; H_5 = \frac{0.20(i_1 - 2i_{0.8} + 2i_{0.3})}{\sqrt{2}\, I_{H1}} \times 100 \tag{4.17}$$

where I_{H1} = rms value of the fundamental output current

$\% \; H_3$ = I_{H3} expressed as a percentage of I_{H1}

$\% \; H_5$ = I_{H5} expressed as a percentage of I_{H1}

and i_1, $i_{0.8}$, $i_{0.5}$, and $i_{0.3}$ are determined as shown in Fig. 4.9.
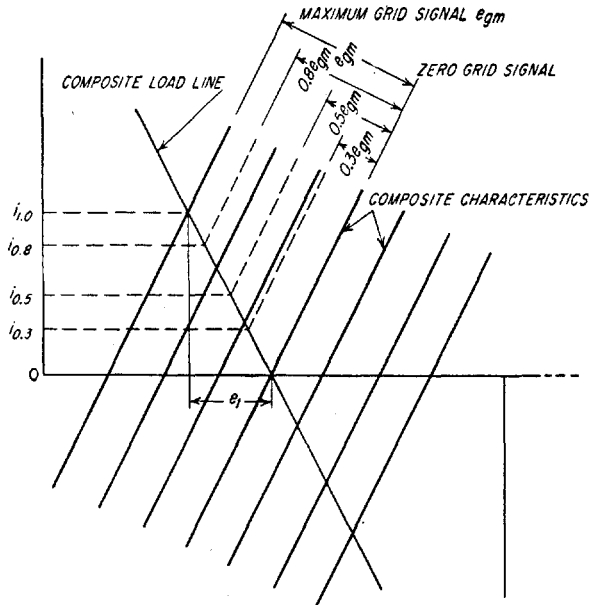
FIG. 4.9. Composite characteristics for push-pull triodes showing the current points for power output and distortion calculations.

The power output P_{o1} due to the fundamental can be determined from Eq. (4.18).

$$P_{o1} = \frac{I_{H1}^2 R_{pp}}{4} \tag{4.18}$$

where P_{o1} is in watts, I_{H1} is the rms value of the plate current in amperes as determined from Eq. (4.15), and R_{pp} is in ohms. If the 3rd and 5th harmonics are small, which is usually the case for push-pull triodes, the approximate power output P_o can be determined from either Eq. (4.19) or (4.20).

$$P_o = \frac{i_1^2 R_{pp}}{8} \tag{4.19}$$

$$P_o = \frac{e_1 i_1}{2} \tag{4.20}$$

where P_o is in watts, e_1 is in volts, and is determined from Fig. 4.9, i_1 is in amperes and is determined from Fig. 4.9, and R_{pp} is in ohms.

curve. The composite quiescent bias line goes through zero at the quiescent plate voltage since $yw = yz$ at this value of plate voltage.

The composite grid signal lines have been constructed similar to the method used to construct the composite bias line. As an example, consider the 5-volt grid signal line shown in Fig. 4.10. It has been assumed that the bias on $V1$ has been reduced by 5 volts and the bias on $V2$ increased by 5 volts. If these bias voltages are maintained and the respective plate voltages varied in equal incremental steps in opposite directions, the instantaneous differences in plate currents will be defined by the +5-volt grid-signal line. The difference current $y'x'$ is seen to be equal to $y'w' - y'z'$ at all points on the composite +5-volt grid line.
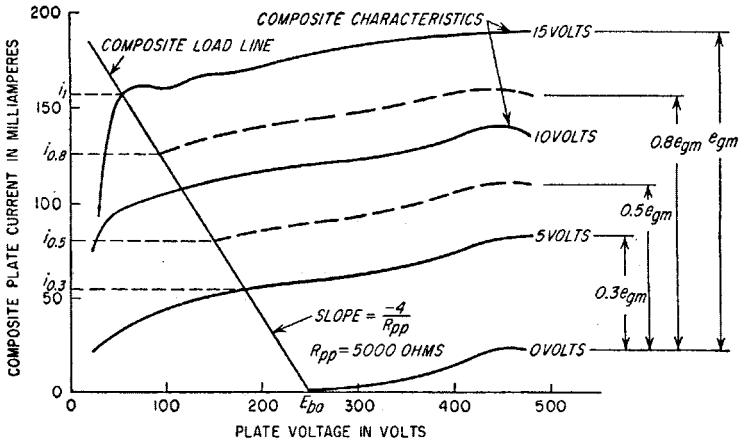

Fig. 4.11. Upper half of the composite characteristics for push-pull transformer–coupled 6L6's which have a quiescent plate voltage of 250 volts, a screen voltage of 250 volts, and a bias of −15 volts.

Figure 4.11 shows one-half of the complete family of composite characteristics for push-pull 6L6 tubes operating with quiescent plate voltages of 250 volts, screen voltages of 250 volts, and a fixed bias voltage of −15 volts. Note that the curvature of the composite lines is much the same as the individual characteristics.

The optimum composite load line for push-pull tetrodes and pentodes extends from the knee of the composite curve associated with the most positive grid signal to the point defined by zero plate current and the quiescent plate voltage (see Fig. 4.11). As in the case of a triode transformer-coupled amplifier, the composite load line slope is equal to $-4/R_{pp}$, where R_{pp} is the plate-to-plate a-c load resistance [see (Eq. 4.12)]. The optimum value of the plate-to-plate a-c load resistance is, therefore, equal to 4 times the reciprocal of the optimum composite load line slope. The paths of operation for the individual tubes can be found in the same manner as for push-pull triodes (Sec. 4.2a).

The power output can be calculated by the use of either Eq. (4.18), (4.19), or (4.20), whichever is applicable. The percentages of 3rd and 5th harmonics can be determined using Eqs. (4.16) and (4.17) and the composite characteristics as shown in Fig. 4.11.

Example 4.2

Calculate the maximum power output and the harmonic distortion for the 6L6's operating under the conditions delineated in Fig. 4.11. The current values taken from the figure are as follows:

$$i_1 = 157 \text{ ma}$$
$$i_{0.8} = 126 \text{ ma}$$
$$i_{0.5} = 80 \text{ ma}$$
$$i_{0.3} = 54 \text{ ma}$$

From Eq. (4.15)

$$I_{H1} = \frac{(7 \times 0.157) + (6 \times 0.126) + (10 \times 0.080) - (6 \times 0.054)}{15 \sqrt{2}}$$

$$= 0.11 \text{ amp rms}$$

From Eq. (4.18)

$$P_{o1} = \frac{(0.11)^2 \times 5,000}{4}$$

$$= 15.1 \text{ watts}$$

From Eq. (4.16)

$$\% H_3 = \frac{0.33[0.157 - (2 \times 0.080)]}{\sqrt{2} \, I_{H1}} \times 100$$

$$= -0.64\%$$

The negative sign indicates that the harmonic subtracts from the fundamental when the fundamental reaches its positive crest value.

From Eq. (4.17)

$$\% H_5 = \frac{0.20[0.157 - (2 \times 0.126) + (2 \times 0.054)]}{\sqrt{2} \, I_{H1}} \times 100$$

$$= 1.67\%$$

4.2c. Class AB1 and Class AB2 Push-pull Transformer-coupled Power Amplifiers.
Class AB push-pull amplifiers are used where higher power is required from a given set of tubes than is obtainable when class A operation is used. In both class AB1 and class AB2 amplifiers, each tube conducts for more than 180 but less than 360 electrical degrees of each cycle. The class AB2 amplifier differs from class AB1 only in the respect that grid current flows in class AB2 during a portion of the cycle. As a result of the grid current flow, additional consideration must be given to the design of the input circuit.

The limiting case for class A push-pull operation is realized if the bias voltage is so selected that the plate current of one tube reaches zero at the instant the grid signal and the plate current of the other tube are at a maximum. If the bias were to be increased, class A operation would no longer exist and class AB would begin since each tube would be driven beyond cutoff for a portion of each cycle. In class AB operation, the same tubes will yield a higher power output and an increase in plate efficiency. The harmonic distortion also increases because the plate current no longer flows for the full cycle in either tube. Because of the mutual coupling in the output transformer, however, the effect of interrupted plate current flow in the individual tubes is greatly reduced.

In class AB operation, the quiescent value of the plate current is made lower for a particular plate voltage than in class A operation because of the increased bias voltage. Therefore a higher plate supply voltage can be used without exceeding the quiescent plate dissipation rating of the tube. However, associated with the application of the grid signal is an increase in the average value of the plate current. Hence, the power supplied by the plate power supply increases with grid signal in the class AB case, whereas in the class A linear amplifier the input power remains constant. The plate dissipation may either increase or decrease depending upon the rate at which the power output and power input change relative to one another. In general, for high negative grid bias voltages, the plate dissipation increases over the quiescent value with grid excitation, and for bias voltages approaching class A operation, the plate

dissipation decreases. The increase in power output over a class A amplifier is due to the increased plate voltage and the extension of the path of operation into the nonlinear region of the tube characteristics.

The composite characteristics for push-pull class AB operation are constructed in the same manner as for class A amplifiers except that in the case of class AB2 it is necessary to include the positive grid bias curves for the individual tubes. The bias voltage is selected so that the plate current in each tube will flow for more than half but less than the full electrical cycle.

Fixed bias is recommended in preference to self-bias in class AB amplifiers, unless a constant amplitude signal is being amplified, since the magnitude of the self-bias is a function of the average plate current which is a function of the input signal level.

1. Push-pull Class AB1 Power Amplifiers. Class AB1 push-pull power amplifiers present no specific design considerations beyond those detailed in the preceding paragraphs. Typical efficiencies will usually be between 45 and 55 per cent.

2. Push-pull Class AB2 Power Amplifiers. Class AB2 amplifiers are used where higher power output is required than is possible with a given pair of tubes in class AB1 operation, and the distortion is to be less than is obtained with class B operation. Typical efficiencies for class AB2 push-pull amplifiers are usually between 50 and 60 per cent.

The phase-inverter method of grid excitation, which could be used with class A and class AB1 amplifiers, cannot be applied to class AB2 amplifiers since a low-impedance driving source is required to drive the grids in the positive region with a minimum of distortion. If the output impedance is sufficiently low, the variation in grid loading, due to driving the grid from the negative region into the positive region, will not cause the grid input waveform to be objectionably distorted. The most common low-impedance input system is a step-down transformer with a center-tapped secondary (see Sec. 4.2d for a discussion of driver stages). Since the driver stage must supply power to the grid circuit, it also is referred to as a power amplifier.

4.2d. Push-pull Class B Audio Power Amplifiers. Class B push-pull amplifiers are used in both audio and radio frequency applications. Class B r-f amplifiers are treated in Sec. 4.3.

Class B push-pull audio amplifiers are used where higher power is required from a given set of tubes than is obtainable from either a class AB1 or a class AB2 amplifier. Associated with the increase in power output is an increase in distortion, driving power, and efficiency. The theoretical maximum value of efficiency in a class B push-pull amplifier is 78.5 per cent. Practical values, however, are ordinarily between 50 and 65 per cent.

In class B operation, the bias voltage on the two tubes is increased until the individual tube currents are very nearly equal to zero at quiescence. The composite characteristics are constructed in the same manner as for class A, but, because of the high value of bias, the composite characteristics are essentially the individual characteristics as shown in Fig. 4.12. Only at small values of plate current do the composite characteristics deviate from the individual characteristics.

Class B amplifiers are almost always operated in class B2, where the grids are driven far into the positive region by the grid input voltage peaks. The optimum load line for class B push-pull triodes has a slope very nearly equal to the negative of the average slope of the individual characteristics. In general, the optimum load line for class B push-pull tetrodes or pentodes can be considered to be that load line which intersects the knee of the most positive grid-bias curve to which the tube is driven. For example, if the quiescent bias were equal to -30 volts, an input signal having a peak value of 50 volts would drive the grid to a potential of $+20$ volts. The optimum load line in this case should be drawn from the knee of the $+20$-volt grid-bias curve through the point defined by zero plate current and the quiescent plate voltage. For

either triodes, tetrodes, or pentodes, the optimum plate-to-plate load resistance R_{pp} is equal to four times the negative inverse of the optimum load line slope.

The distortion of the grid-input signal caused by driving the grids in the positive region for a portion of each cycle is often the largest source of distortion in a push-pull class B amplifier. By utilizing a sufficiently low impedance driver, however, this source of distortion can be disregarded. The use of a driver stage having a step-down output transformer or a driver stage consisting of two cathode-follower circuits
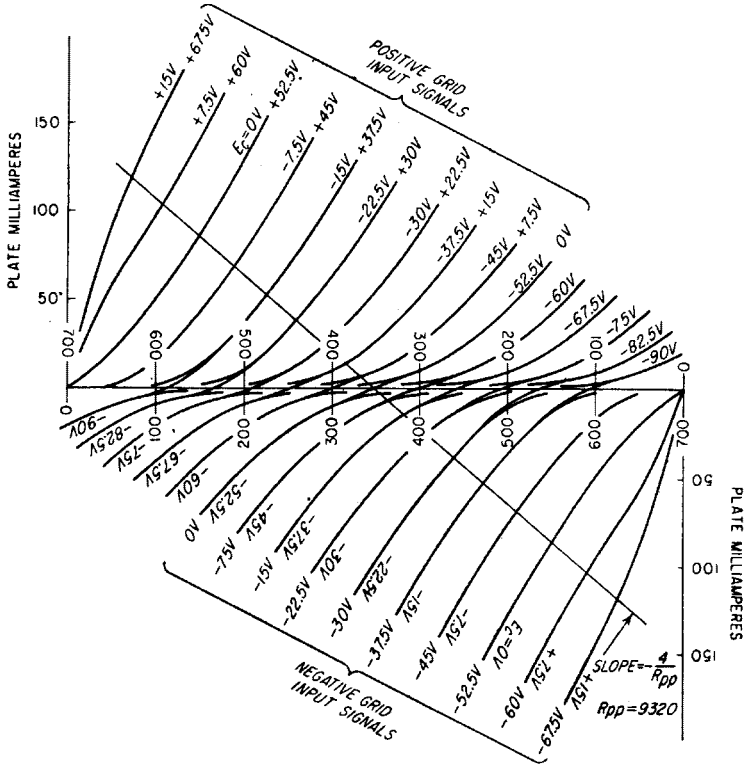
FIG. 4.12. Composite characteristics for push-pull class B triode connected 6L6 tubes. $E_{bo} = 350$ volts and $E_{cc} = 52.5$ volts.

operated in push-pull represents two of the most common means of obtaining a low-output impedance. If a transformer is used, it should have a low leakage inductance since this inductance is effectively in series with the grids. Another reason for minimizing the leakage inductance is that the abruptly changing grid-load impedance often causes transient ringing in the resonant circuit consisting of the leakage inductance and the capacitance shunting the secondary. The d-c resistance of the grid current path must also be very low or the flow of grid current will cause a change in bias. This criterion imposes the requirement of a bias source having a low internal resistance. In this respect, the push-pull cathode-follower driver has an advantage over the driver utilizing a step-down transformer since the cathode followers can be operated so that their quiescent cathode voltages are at the potential necessary to provide the proper bias on the output stage. Self-bias is never used in class B stages for the reasons given in Sec. 4.2c for class AB amplifiers.

Although the peak grid input power ordinarily imposes a more severe requirement on the design of the driver stage than does the average grid-input power, there often is a need to calculate the average grid input power. If the input signal is sinusoidal, the grid will conduct for only a small portion of each cycle. Since the grid signal is very nearly equal to its maximum value during the entire period of grid conduction, the average grid-input power P_g per tube can be determined from Eq. (4.21).

$$P_g = e_{gm}i_{ga} \qquad (4.21)$$

where e_{gm} = peak value of grid-input voltage
      i_{ga} = average value of grid current in each tube

The power-supply regulation for a class B stage should be good, otherwise the large change in average plate current between quiescence and full output power will cause a significant change in the operating point.

The power output and plate dissipation of a class B push-pull power amplifier can be determined approximately if the following simplifying assumptions are made:

1. The grid-input voltage waveform is sinusoidal.
2. The output transformer is an ideal transformer.
3. The plate current through each tube is a half sinusoid.

Based on the preceding assumptions, the power output can be determined from either Eq. (4.19) or (4.20) where i_1 is the maximum instantaneous value of the plate current per tube. The average value of the half-sinusoid plate current waveform per tube and the total power supplied by the plate circuit power supply can be determined from Eqs. (4.22) and (4.23), respectively.

$$I_b = \frac{i_1}{3.14} \qquad (4.22)$$

where I_b = average current through each tube
      i_1 = peak plate current through each tube

$$P_b = 2E_{bb}I_b \qquad (4.23)$$

where P_b = total power supplied by plate circuit power supply

In Fig. 4.12 are typical composite characteristics for a class B push-pull amplifier. The power output with a grid signal having a peak value of 67.5 volts is 16.1 watts with a total harmonic distortion of 11 per cent. If the 6L6's are operated as pentodes, the maximum push-pull power output with a plate voltage of 350 volts is approximately 50 watts.

4.3. Class B Radio-frequency Power Amplifiers. In a class B r-f power amplifier, the plate load is resonant and plate current flows for 180° of each electrical cycle. A class B r-f amplifier is used where the power level of a signal is to be increased, but with a linear relationship between the input and output voltages. It is frequently referred to as a "linear amplifier." A class A r-f amplifier is also a linear amplifier, but the efficiency is considerably lower. A class C r-f amplifier (see Sec. 4.4) is somewhat more efficient than a class B amplifier, but a linear relationship between the input and output voltages does not exist. The linearity of a class B r-f amplifier is important when it is used to amplify signals such as an amplitude-modulated carrier or a single sideband suppressed carrier.

A typical linear amplifier is shown in Fig. 4.13. The amplifier does not have to be push-pull in r-f applications since the plate circuit is resonant. In most r-f power amplifiers, the maximum power output possible from a given tube is usually desired. Therefore the input grids are ordinarily driven into the positive grid region during the peak of the excitation cycle, causing grid current to flow. This means that grid-leak bias cannot be used since a change in modulation would cause a change in the

average grid current and result in a change in bias. It follows that the bias source must be well regulated if nonlinearities between the input and output voltages are to be prevented.

In the design of a class B r-f power amplifier, the semigraphical method outlined in Sec. 4.4 for class C power amplifiers should be used with the following modifications:

1. Since the angle of plate current conduction is 180° for class B, the bias voltage E_{cc} can be selected directly from the constant current characteristics for the tube. This bias voltage is the value which just causes zero plate current to flow with zero grid excitation.

2. The minimum instantaneous plate-to-cathode voltage is limited to approximately twice the peak positive grid-to-cathode voltage during the excitation peak.

3. In the selection of a tube for a class B r-f power amplifier which is to amplify a 100 per cent amplitude-modulated carrier, the rated plate dissipation of the tube should be of the order of twice the unmodulated carrier power output. In Eq. (4.39), the value of k is equal to 2.0.
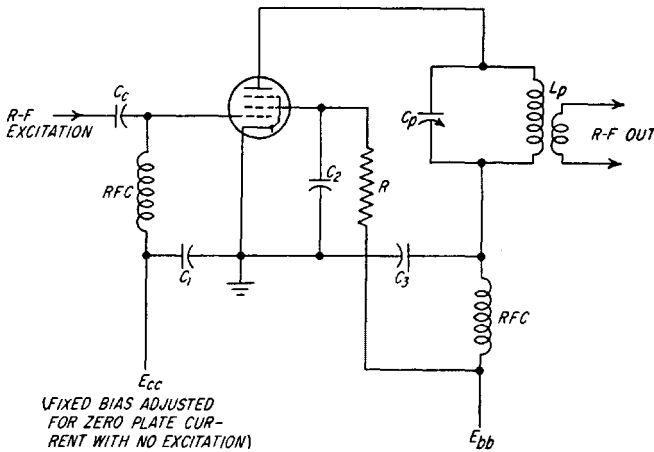
FIG. 4.13. Class B radio frequency power amplifier.

The remainder of the design considerations and calculations is the same as for a class C amplifier. The maximum theoretical plate efficiency for the linear amplifier is the same as for a class B audio amplifier, i.e., 78.5 per cent. Usual peak operating efficiencies are between 60 and 70 per cent.

When a carrier is 100 per cent amplitude-modulated by a sinusoidally varying function, the peak r-f voltage during modulation reaches twice the carrier level and the minimum voltage drops to zero. Therefore, if the modulated carrier is to pass through the amplifier without distortion, the unmodulated input carrier level must be adjusted so that the output voltage is one-half the maximum output voltage that the tube is capable of delivering. The fundamental component of plate current and the average plate current in a class B r-f power amplifier are directly proportional to the amplitude of the grid excitation voltage; therefore the efficiency of the amplifier is directly proportional to the grid voltage provided the amplifier is not overdriven. Since the efficiency of the linear amplifier is in the order of 60 to 70 per cent with maximum grid excitation, i.e., at the peak of a 100 per cent modulated carrier, the efficiency of the amplifier for an unmodulated carrier is in the order of 30 to 35 per cent. Since the class B r-f power amplifier is linear, the plate circuit d-c input power remains constant and there is a 50 per cent increase in the average output power when 100 per cent sinusoidal modulation is applied to the carrier. With 100 per cent

modulation, the average efficiency is in the order of 45 to 50 per cent. The fact that the plate current with modulation remains constant indicates that cathode bias can be used. The absence of a change in the input power as modulation is introduced does not ensure that the amplifier is linear; hence an oscilloscope should be used for modulation checks.

4.4. Class C Radio-frequency Power Amplifiers. Class C amplifiers provide high efficiencies and are used primarily in r-f applications where either an increase in power level or a multiplication of frequency is required and where there is no need for linearity between the input and output voltages.

Class C amplifiers can be amplitude-modulated by injecting the modulating voltage on the plate, screen grid, suppressor grid, control grid, or cathode. They can also be used to amplify a frequency-modulated carrier. A detailed analysis of modulation is given in Sec. 5.

Although plate circuit power efficiencies as high as 85 per cent are possible in class C amplifiers, most amplifiers are designed to operate at efficiencies in the order of 75 per cent since there must be a very significant increase in the grid driving power to increase the plate circuit efficiency from 75 to 85 per cent. Highest efficiencies are obtained with high plate voltages.

In a class C amplifier, the grid is biased below cutoff and plate current flows for less than 180° of each cycle of the grid excitation voltage. Consequently, class C amplifiers are not used in audio applications since the output voltage is not proportional to the input voltage and the short plate current pulse would cause excessive distortion even if a push-pull arrangement were used. However, a class C amplifier can be used as an r-f amplifier if the plate circuit is resonant at either the fundamental or some harmonic of the grid excitation frequency.[1] In an r-f amplifier, the fundamental component of the plate current pulses is filtered by the plate resonant circuit to provide an output which is very nearly sinusoidal. When the amplifier is used as a frequency multiplier, i.e., when the plate circuit is resonant at some harmonic of the excitation voltage, the power output is somewhat reduced. Class C amplifiers used for frequency doubling and tripling will ordinarily have efficiencies in the order of 40 to 50 and 15 to 20 per cent, respectively.

4.4a. Basic Circuit Configurations. Several typical circuit configurations of class C r-f amplifiers and associated neutralizing networks are shown in Fig. 4.14.

It should be noted that in those circuits having a balanced grid and/or plate circuit, the center tap of the inductance is not connected to the rotor of the split stator tuning capacitor. Connecting these points can result in a reduction in the plate circuit efficiency due to the increase in circulating currents in the tank circuit if either the inductor or the tuning capacitor is not exactly symmetrical. In the circuits of Fig. 4.14 the rotor of the tuning capacitor has been put at zero signal potential and the center tap of the inductance establishes its own signal level since it has been isolated by either a grid-leak biasing resistor or an r-f choke. In Fig. 4.14b, the rotor of C_3 has been placed at zero signal potential by the capacitor C_{11}. This permits C_3 to have a lower voltage breakdown rating since, with the addition of R_2, the d-c voltage drop across C_3 has been removed. R_2 is not needed if the amplifier is not to be plate-modulated and the rotor of C_3 can be connected directly to the high-voltage power supply.

Lumped L and C input and output circuits as shown in Fig. 4.14 are seldom used at frequencies higher than approximately 300 or 400 Mc. Instead, the input and output circuits usually consist of either coaxial lines or parallel lines. Also, grounded-grid amplifiers are usually used at these higher frequencies since they are inherently more stable than grounded cathode amplifiers.

[1] In push-pull operation, the plate circuit must be tuned to either the fundamental or an odd harmonic.
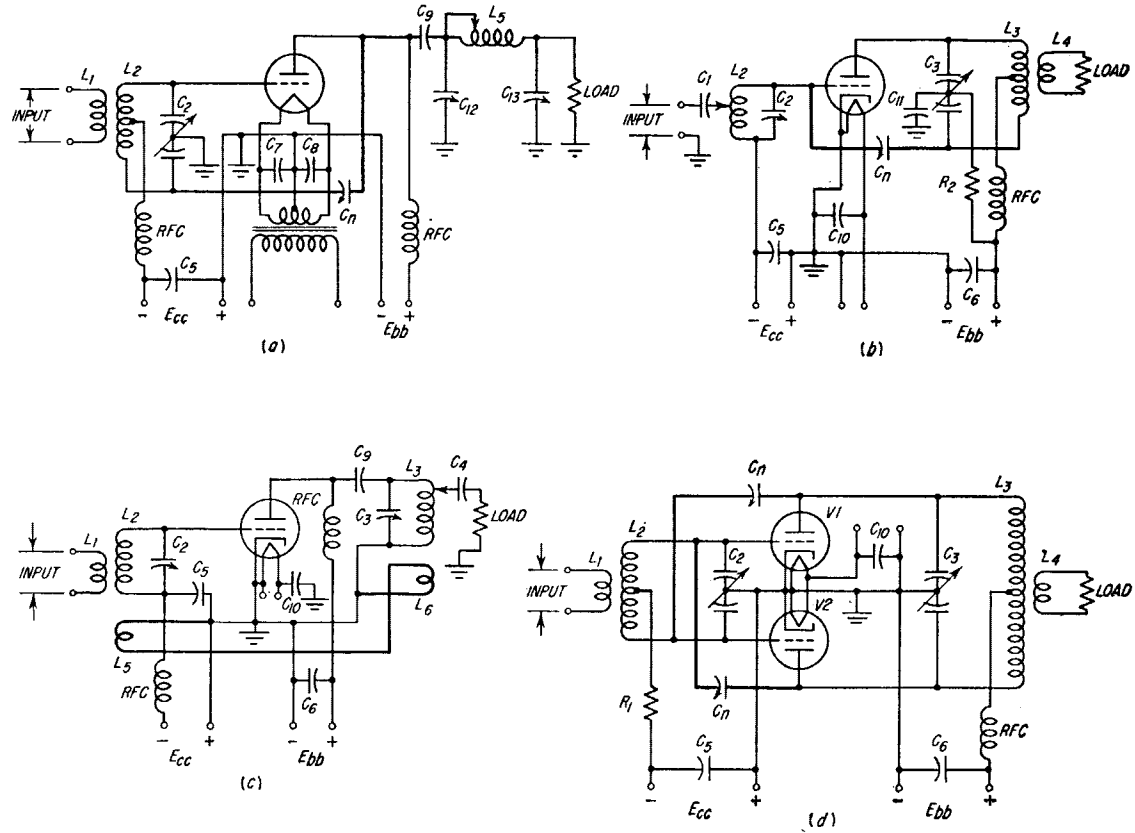
FIG. 4.14. Typical class C r-f power amplifier configurations.

Bias Considerations.   Cathode bias, fixed bias, and grid-leak bias can be used individually or in combinations to provide the bias for a class C amplifier.   Grid-leak bias is most commonly used and obviates the need of a bias supply.   Also, grid-leak bias is self-regulating, and, consequently, the bias automatically tends to adjust itself to the peak value of the grid excitation voltage, thereby eliminating the need for a critical adjustment of the excitation voltage.   However, if grid-leak bias is used, caution must be taken to ensure that excessive plate or screen current does not flow in the absence of grid excitation.   Frequently, this minimum bias requirement is provided in the form of either cathode bias or fixed bias (see Fig. 4.14*d*).   Protection can also be achieved by utilizing a current overload relay in series with the high voltage supply (this relay is usually placed in the ground lead to reduce the relay insulation requirements).   Loss of excitation causes an appreciable increase in plate current which operates the relay, thereby providing the desired protection.   Another means of obtaining protection when screen-grid tubes are employed is to use a screen-grid voltage dropping resistor and to shunt the screen grid with a small tube which uses the same bias source as the tube to be protected.   With the loss of excitation, the added tube conducts sufficiently so that the voltage drop across the screen-grid resistor causes the screen voltage of the power tube to drop to an abnormally low voltage, thereby preventing damage to the power tube in the absence of the operating bias.   When a class C amplifier is amplitude-modulated, special bias considerations apply (see Sec. 5.3*a*).

Tube Types.   Triodes, tetrodes, and pentodes are employed in class C amplifiers.   Tetrodes are often employed in preference to triodes because of their low internal grid-to-plate capacitance and because of the much lower grid power excitation requirements.   In uhf applications, the foregoing considerations are of particular importance.   Very few high-power pentodes are manufactured.

Types of Coupling.   Both inductive and capacitive coupling have been shown as the means of coupling power into the grid circuit and out of the plate circuit.   Inductive coupling introduces more attenuation to the harmonics coupled to the load than does capacitive coupling.   In addition, a Faraday shield can be placed between the plate tank coil and the inductive link to minimize the capacitive coupling between the plate circuit and the load.   This shielding further minimizes the transfer of harmonics to the load.   When the plate circuit is single-ended, the inductive link should be located at the high-voltage-supply end of the plate-circuit inductance.

A frequently used plate circuit network is the π network[1] shown in Fig. 4.14*a*.   This network (1) serves as the resonant plate circuit, (2) introduces increased attenuation to harmonics as compared to a single-tuned parallel resonant circuit, and (3) provides a means of matching to a wide range of load impedances.

Neutralization.   In the circuits shown in Fig. 4.14, the current which flows through the grid-to-plate capacitance will have a component 180° out of phase with the grid excitation voltage if the plate circuit is tuned slightly higher in frequency than the grid excitation frequency.   If the grid-to-plate capacitance is sufficiently large and if the grid losses are not excessive, the plate circuit will couple an adequate amount of power into the grid circuit to cause the amplifier to oscillate.   Neutralization consists of applying feedback between the plate and grid so as to reflect a resistive load into the grid circuit, thereby canceling the negative resistance effect produced by the current through the grid-plate capacitance.   Several different methods of neutralizing class C amplifiers over a relatively broad band of frequencies are shown in Fig. 4.14.   A simple narrow band neutralization network for a single-ended amplifier which has not been shown can be realized by connecting an inductance and capacitance in series between the grid and plate.   The magnitude of the inductive reactance is made greater than the magnitude of the capacitive reactance by an amount equal to the magnitude

[1] See E. W. Pappenfus and K. L. Klippel, Network Tank Circuits, *CQ*, September, 1950.

of the grid-to-plate capacitive reactance. The result is that the added network has an inductive reactance which exactly neutralizes the effect of the grid-to-plate capacitive reactance. This particular type of neutralization is often used in single-frequency r-f amplifiers.

To make a neutralization adjustment, the filaments or heaters should be turned on, the high voltage turned off, and grid excitation applied. Utilizing a sensitive r-f detector in the plate circuit, adjust the neutralizing network to minimize the power which can be detected if the plate circuit as the plate circuit is tuned over its operating limits. In relatively high power class C amplifiers, it may not be possible to obtain an adjustment for zero r-f power in the plate circuit. Neutralization can also be achieved by adjusting the neutralizing network to minimize the dip in grid current as the plate circuit is tuned through resonance. During the adjustment, the filaments or heaters should be on, high voltage turned off, and grid excitation applied as before.

With improper neutralization in a plate-modulated class C amplifier, the modulation trapezoidal pattern cannot be driven to a triangle, i.e., the pattern which indicates 100 per cent modulation. Consequently, the final neutralization adjustment in a plate-modulated class C amplifier frequently consists of adjusting the neutralization so that the modulation pattern is triangular with sufficient modulation.

In tubes having a screen grid, the control grid-to-plate capacitance is much lower than in triodes. Consequently, tetrodes and pentodes do not ordinarily need to be neutralized if the screen grid is at zero signal potential and if the control grid-to-plate capacitance external to the tube is very small. This latter condition, however, requires very thorough shielding between the control grid circuit and the plate circuit. The screen grid can be put at zero signal potential by bypassing the screen grid to the cathode at the tube socket. If the amplifier is to be used at frequencies where the effects of the internal screen grid lead inductance become noticeable, the screen grid should be bypassed only with a small variable capacitor which can be adjusted to resonate with the screen grid lead inductance.

Parasitic Oscillations. Undesired oscillations are referred to as parasitic oscillations. Most amplifiers and oscillators are susceptible to parasitic troubles unless special precautions are taken to prevent them. Parasitic oscillations can result in (1) abnormal tube currents which may shorten the tube life, (2) the inability to achieve 100 per cent modulation, and (3) interference with other services. These oscillations ordinarily occur in the frequency band between approximately 50 and 200 Mc or in the frequency range between approximately 200 and 1,200 kc.

High-frequency parasitic oscillations in class C amplifiers are usually due to the inductances of the grid and plate circuit leads between the tube and the tank circuits resonating with the tube interelectrode capacitances and any external shunting capacitances. High-frequency parasitics can sometimes be eliminated by tuning the plate circuit of the parasitic oscillator to a frequency lower than the grid circuit so that the amplifier will no longer oscillate. One method of accomplishing such tuning is to make the grid leads considerably shorter than the plate leads so as to cause the inductance of the plate leads to be much greater than the inductance of the grid leads. The same effect can also be obtained by placing small inductances, usually only a few microhenries, in each plate lead. Another method of suppressing high-frequency parasitics is to reduce the Q of the parasitic circuit by placing small resistors, usually 50 to 100 ohms, in each grid and plate lead next to the tube. The combination of a small inductance and a shunting resistor in each plate lead is sometimes used. In this application, the inductance is used to lower the frequency of the parasitic oscillation to some less critical frequency, e.g., below the frequencies used for television so as to avoid television interference. The resistor shunting the coil should be noninductive and should shunt only a sufficient number of turns to suppress the

parasitic. The resistor should have a value in the order of 50 to 100 ohms, and a typical inductance might consist of approximately four to six turns with an inside diameter of 0.25 in. Tuned traps as shown in Fig. 4.15 are also used to eliminate parasitics. L_p should consist of one or two turns, and L_s and C_s should resonate at

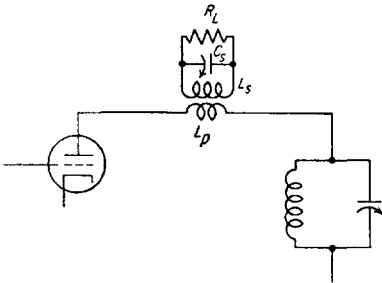the parasitic frequency. The coupling between L_p and L_s should be increased until the parasitic disappears. The value of R_L is usually 50 to 100 ohms.

Tetrodes exhibit the characteristic of being self-neutralized for a limited range of frequencies determined by the grid-to-plate capacitance of the circuit, and ordinarily high-frequency parasitic oscillations occur at frequencies above this region. Consequently, one method of suppressing a high-frequency parasitic in a tetrode is to lower the frequency of the parasitic oscillation so that it falls in the frequency range

Fig. 4.15. Tuned trap for eliminating high-frequency parasitic oscillations.

where the tube is self-neutralized. Sometimes this can be done by placing small inductances in the plate leads. The parasitic circuit may be composed of the control-grid and screen-grid circuits or the control-grid and plate circuits. If parasitics should exist in the screen circuit, they can be eliminated by inserting a small unbypassed noninductive resistor, usually between 50 and 100 ohms, in both the control-grid and screen-grid leads at the tube base. The introduction of a resistor in the screen-grid circuit reduces the effectiveness of the screen-grid shielding between the control grid and plate. Therefore, it is likely that neutralization will be required between the plate and control-grid circuits if an unbypassed screen resistor is added. If a tetrode is used at frequencies sufficiently high that the reactances of the grid and screen-grid lead inductances become appreciable, the screen grid should be bypassed to the cathode with a small variable capacitor to achieve proper neutralization. It may also be necessary to add a neutralizing circuit between the plate and control-grid circuits if the external control grid-to-plate capacitance is excessive.

Low-frequency parasitic oscillations sometimes exist in class C amplifiers whenever both the grid and plate circuits utilize r-f chokes. As an example of a typical low-frequency parasitic circuit, consider Fig. 4.14c. The grid resonant circuit would consist of the grid r-f choke in parallel with C_5, and the plate resonant circuit would consist of the plate r-f choke in parallel with C_9. The reactances of L_2 and L_3 can be considered as insignificant compared to the reactances of the r-f chokes. For low-frequency oscillations to exist, the product of the plate r-f choke inductance and the plate circuit capacitance which establishes the plate low-frequency resonant circuit must be less than the product of the grid r-f choke and the capacitance which determines the grid low-frequency resonant circuit. With the circuit of Fig. 4.14d, there should be no problem with low-frequency parasitic oscillations since the grid circuit does not contain an r-f choke. However, it is possible to have low-frequency parasitics in Fig. 4.14b if the lead inductance to the bias supply is excessive. In this case, the grid resonant circuit would consist of the bias lead inductance in parallel with C_5.

When testing a class C amplifier for parasitics, the grid excitation voltage should be removed, the bias reduced to a small value, and the plate and screen voltages lowered sufficiently so that rated dissipations are not exceeded. The plate tuning capacitor should be tuned through its limits for several grid tuning capacitor settings. Changes in either the grid or plate current as a result of tuning and the detection of r-f power on the plate leads by means of a neon lamp, absorption meter, etc., are indications of parasitics.

4.4b. Voltage and Current Relationships. **Voltage and current waveforms for a typical single-ended tetrode or pentode class C amplifier used for direct amplification, i.e., not frequency multiplication, are shown in Fig. 4.16.   To obtain high values of efficiency, the control grid bias E_{cc} is usually made between 1.5 and 3 times higher than the value required for plate current cutoff.   Biasing the tube beyond plate**

current cutoff ensures that plate current will flow less than 180° of each electrical cycle.   If the plate resonant circuit has a sufficiently high Q, the output plate voltage waveform will be very nearly sinusoidal even though plate current flows for less than 180°.   Since the resonant plate circuit is tuned to the operating frequency, the tube operates into a resistive plate load; consequently, the plate voltage is 180° out of phase with the grid excitation voltage.

Control-grid current flows for a fewer number of degrees than the plate current since grid current flows only during the limited portion of the cycle in which the control grid is driven into the positive region.

The impedance of the screen-grid circuit to the a-c component of the screen-grid current is usually very nearly equal to zero, and for most calculations it can be assumed that the signal voltage at the screen grid will be zero.

Since the control-grid current, screen-grid current, and plate current flow for less than one-half of each excitation cycle, the ratios of the peak current values to the average values may be relatively high.
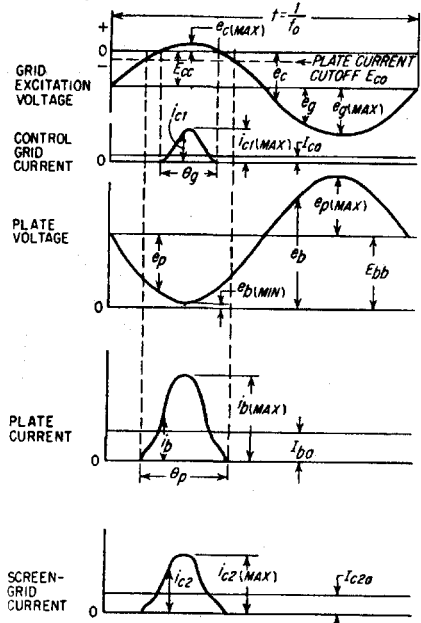
Fig. 4.16. Voltage and current waveforms for a class C amplifier.

These ratios become increasingly greater as the conduction angle is decreased.   In step 4 of the procedure for the design of a class C amplifier (see Sec. 4.4j), a simple method is given for determining the d-c, fundamental, and the 2nd and 3rd harmonics of the control grid, screen grid, and plate current waveforms.
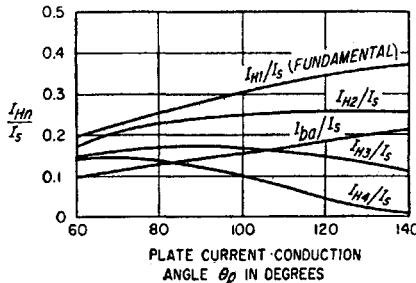
Fig. 4.17. Typical ratios of the peak values of the plate current harmonics to the peak space current I_s as a function of the plate current conduction angle.

Typical ratios of the d-c, peak fundamental, and peak harmonic components of the plate current waveform to the peak space current I_s are shown in Fig. 4.17. I_s is the sum of the instantaneous peak values of the control grid, screen grid, and plate currents.   These curves are all based on the same peak positive grid-to-cathode excitation voltage, the same values of peak grid and peak plate currents, and the same plate supply voltage.   The variables are the plate-circuit impedance, grid bias, and the amplitude of the grid excitation voltage.   It can be seen that for small plate current conduction angles both I_{ba}/I_s and I_{H1}/I_s increase almost linearly as the plate

current conduction angle increases; while at plate current conduction angles greater than approximately 120° there is a noticeable increase in the rate at which the slope of the I_{H1}/I_s curve decreases. For this reason, the plate circuit efficiency is reduced if the plate current conduction angle is made greater than approximately 120°. If the plate current waveform is analyzed, it will be found that the center portion of the pulse contributes more to the fundamental component of the plate current than to the d-c component and the sides of the pulse contribute more to the d-c component than to the fundamental component. Therefore, for high plate efficiency, a narrow plate current pulse with steep sides is desirable. Such a waveform can be obtained if the grid bias voltage E_{cc} is made two or three times greater than the plate current cutoff bias E_{co}.
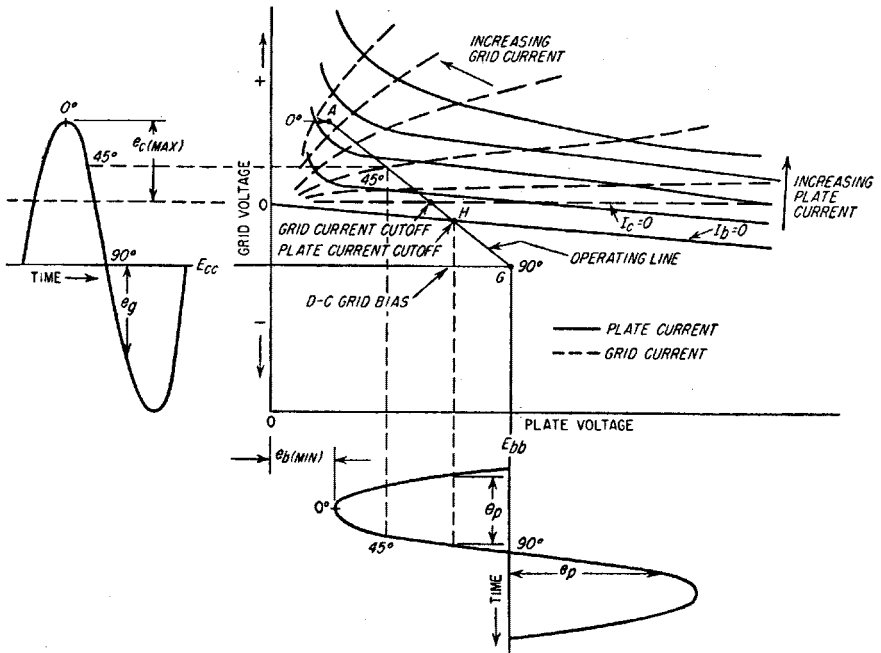
FIG. 4.18. Constant current characteristics for a triode showing an operating line and associated voltage waveforms.

4.4c. Constant Current Characteristics. The constant current characteristics are a modified plot of the familiar E_b-I_b characteristics. By changing the manner in which the tube curves are plotted, a more graphic picture of class C operation can be visualized. The control grid voltage is plotted on the ordinate and the plate voltage on the abscissa. A family of curves is plotted for the control-grid, screen-grid, and plate currents. The individual curves in each family represent a constant value of current. In Fig. 4.18 are typical constant current characteristics for a triode with typical grid and plate class C voltage waveforms included (see Sec. 4.4*j* for procedure in graphical analysis).

4.4d. Plate Circuit Power Relationships. The useful power developed in the plate circuit can be expressed in terms of the fundamental component of the plate signal voltage and the fundamental component of the plate current. If the loaded plate tank circuit Q_o is greater than approximately 12, the plate voltage can be considered to consist entirely of the fundamental. The term Q_o indicates the ratio of the energy

stored in the tank circuit to the energy dissipated in the tank circuit plus the energy coupled from the tank circuit during one r-f cycle. When Q_o is appreciably less than 12, the plate voltage waveform will be noticeably distorted due to the increased harmonic content. However, the assumption that the plate circuit waveforms are sinusoidal will usually not result in significant errors in the calculation of the plate circuit power relationships.

The *plate power output P_o* can be expressed by Eq. (4.24).

$$P_o = \frac{e_{p(\max)}I_{H1}}{2} \qquad (4.24)$$

where $e_{p(\max)}$ = peak value of r-f plate voltage, that is, $e_{p(\max)} = E_{bb} - e_{b(\min)}$
        I_{H1} = peak value of the plate current fundamental [see Eq. (4.45)]
The peak value of the fundamental current component I_{H1} must be found by analyzing the plate current waveform. A simple method is described in step 10 of the procedure for the design of a class C amplifier (see Sec. 4.4j). The value of P_o less the loss in the plate tank circuit and the coupling network is the power in watts which would be indicated on an r-f wattmeter.

The *plate power input P_{dc}* that must be supplied from the d-c power source is given by Eq. (4.25).

$$P_{dc} = E_{bb}I_{ba} \qquad (4.25)$$

where I_{ba} = d-c plate current [see Eq. (4.44)]

The *plate efficiency η_p*, that is, the ratio of the plate power output to the plate power input, expressed as a percentage, is

$$\eta_p = \frac{e_{p(\max)}I_{H1}}{2E_{bb}I_{ba}} \times 100 \qquad (4.26)$$

Neglecting the higher-order harmonics, the difference between the input power and the output power is the power lost in plate dissipation P_{bs} and can be determined from Eq. (4.27).

$$P_{bs} \simeq E_{bb}I_{ba} - \frac{e_{p(\max)}I_{H1}}{2} \qquad (4.27)$$

4.4e. Plate-load Resistance. The operating line on the constant current characteristics does not explicitly represent the plate-load resistance r_b as does a load line on the E_b-I_b characteristics. However, for any one operating line, there is only one value of load resistance which will satisfy the conditions imposed by the operating line, and its value can be calculated from Eq. (4.28).

$$r_b = \frac{e_{p(\max)}}{I_{H1}} \qquad (4.28)$$

$$\simeq \frac{(E_{bb} - E_{cc2})^2}{2P_o} \qquad \text{(tetrode or pentode)} \qquad (4.28a)$$

where E_{cc2} is the screen-grid voltage.
The process of assuming a plate-load resistance and establishing the associated operating line is a trial-and-error procedure and relatively difficult because of the graphical methods involved. As is shown in Sec. 4.4j, the recommended procedure in a class C amplifier design is to determine graphically I_{H1} and calculate the required value of r_b.

In the simplest type of plate tank circuit the external load resistance, e.g., an antenna or an additional stage, which is coupled to a single-tuned plate-resonant circuit, can be considered as a resistance R_r in series with the plate-circuit inductance.

In Fig. 4.19, R_s is the effective series resistance of the inductor itself, i.e., the d-c resistance plus the resistance due to skin and proximity effects. Ordinarily the Q of the unloaded tank circuit is sufficiently high that R_s can be neglected. The resistance r_b as measured at the terminals of the plate circuit at resonance can be calculated from Eq. (4.29) provided Q_o is greater than approximately 12. The value of r_b as determined from Eq. (4.29) should be equal to the desired value of r_b as established from Eq. (4.28).
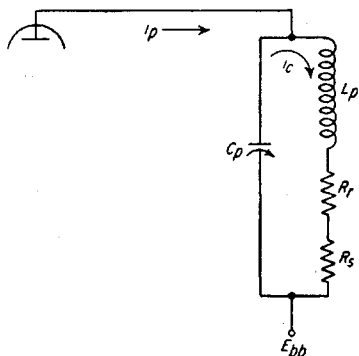
$$r_b = \frac{L_p}{(R_r + R_s)C_p} \qquad (4.29)$$

The Q_o of the loaded plate circuit is given by Eq. (4.30).

$$Q_o = \frac{2\pi f_o L_p}{R_r + R_s} \qquad (4.30)$$

where f_o is the resonant frequency

Combining Eqs. (4.29) and (4.30) results in a more usable expression for r_b which is given by Eq. (4.31).

$$r_b = \frac{Q_o}{2\pi f_o C_p} \qquad (4.31)$$

FIG. 4.19. Equivalent plate circuit of a loaded class C amplifier.

From Eq. (4.31) it can be seen that there are an infinite number of combinations of Q_o and C_p which will result in the same value of r_b. However, Q_o should not be too high, or the circulating current in the tank circuit will be excessive and cause undesired losses which may result in damage to the capacitor or inductor. The rms value of the current which circulates through the capacitor and inductor can be determined from Eq. (4.32).

$$i_c = 0.707 I_{H1} Q_o \qquad (4.32)$$

where i_c = rms value of the circulating current

I_{H1} = peak value of plate current fundamental [Eq. (4.45)]

To prevent excessive power losses it is sometimes necessary to decrease Q_o to a value as low as 3 or 4. This causes the harmonic content to increase in the plate circuit. In this type of operation the harmonic content in the output, if undesirable, can be minimized by the use of a filter in the coupling network. In push-pull stages, Q_o need only be equal to or greater than approximately 6 for adequate harmonic suppression since all even harmonics of the plate voltage waveform are canceled out.

The bandwidth Δf of the loaded plate circuit can be expressed in terms of Q_o and f_o as shown by Eq. (4.33).

$$\Delta f = \frac{f_o}{Q_o} \qquad (4.33)$$

where Δf = bandwidth between the half-power frequencies

Therefore, if Q_o is too high, the higher modulation frequencies will be suppressed in an amplitude-modulated class C amplifier.

In practice, a wide variety of plate circuits are used. For example, in Fig. 4.14 the π network[1] and single-tuned circuits with both inductive and capacitive coupling to the load are employed. Also, the plate resonant circuit may consist of double-tuned circuits (see Sec. 13) which are undercoupled, critically coupled, transitionally coupled, or overcoupled.[2]

[1] Pappenfus and Klippel, *loc. cit.*
[2] R. B. Dome, "Television Principles," chap. 4, McGraw-Hill Book Company, Inc., New York, 1951.

4.4f. Screen-grid Dissipation. The value of the screen-grid dissipation can be determined from Eq. (4.34).

$$P_{c2} = E_{cc2}I_{c2a} \qquad (4.34)$$

where E_{cc2} = screen-grid supply voltage
I_{c2a} = d-c value of screen-grid current

4.4g. Grid Excitation Requirements. In a class C amplifier, the power which must be supplied to the grid circuit can be divided into three categories: (1) grid dissipation P_{c1}, (2) dissipation in the biasing source P_1, and (3) grid tank and coupling losses.

It has been shown by Thomas[1] and verified by others that the grid dissipation P_{c1} can be calculated to a reasonable degree of accuracy by Eq. (4.35).

$$P_{c1} = e_{c(max)}I_{ca} \qquad (4.35)$$

where $e_{c(max)}$ = peak positive grid-to-cathode voltage (see Fig. 4.16)
I_{ca} = d-c grid current

The power P_1 dissipated in the bias resistor in the case of grid-leak bias or the power which tends to charge the bias battery when fixed bias is used is given by Eq. (4.36) or (4.37).

$$P_1 = I_{ca}^2 R_1 \qquad (4.36)$$
$$= E_{cc}I_{ca} \qquad (4.37)$$

where R_1 = value of grid-leak biasing resistor

For low and intermediate frequencies, the power lost in a well-designed grid-tank circuit and input coupling network is usually about 10 per cent of the grid excitation power. The total power P_t required from the preceding stage can, therefore, be expressed approximately by Eq. (4.38).

$$P_t \simeq 1.1(P_{c1} + P_1) \qquad (4.38)$$

At higher frequencies, e.g., above approximately 50 Mc, the grid and grid circuit losses tend to increase significantly with increasing frequency. The required excitation power at uhf frequencies in some cases is *ten or twenty times greater* than the value determined from Eq. (4.38).

4.4h. Efficiency of a Class C Amplifier. The maximum efficiency of a class C amplifier having an excess of grid driving power is greater than that of an amplifier in which the grid driving power is limited. Shown in Fig. 4.20 is a typical plot of grid driving power versus power output. Note that for the same power output the grid driving power must be increased more than 100 per cent to increase the plate efficiency from 75 to 85 per cent. Associated with the increase in driving power is an increase in the grid bias. In effect, the increase in bias causes a reduction of θ_p and thereby an increase in efficiency as pointed out in Sec. 4.4b.

A problem arises when the grid driving power is increased without modifying the other circuit parameters. The power output increases with grid drive up to the point where the minimum instantaneous plate voltage $e_{b(min)}$ swings below the maximum grid-to-cathode voltage at the peak of excitation. If the input is further increased, a valley appears in the plate current pulse and the grid current will very likely be excessive. In a tetrode or pentode, the plate current waveform is similarly distorted when the minimum instantaneous plate voltage $e_{b(min)}$ swings below the screen-grid voltage. In this case, excessive screen-grid current may flow. The same effect will be realized if the plate-load impedance is increased a sufficient amount to cause $e_{b(min)}$ to swing below the control-grid voltage or the screen-grid voltage. There-

[1] H. P. Thomas, "Determination of Grid Driving Power in R. F. Power Amplifiers," *Proc. IRE*, vol. 21, pp. 1134–1141, August, 1933.

fore, in class C amplifiers, the plate should be loaded at all times to prevent over-dissipation in the control grid and screen grid.

If low driving power is important, the minimum instantaneous plate potential $e_{b(\min)}$ should be at least $1.5e_{c(\max)}$, and the value of θ_p should be approximately 150°. If, however, plate efficiency is important, θ_p should be on the low side of 120° with $e_{b(\min)}$ made approximately equal to $e_{c(\max)}$.

4.4i. Harmonic Operation of a Class C Amplifier. A class C amplifier can be used as a harmonic amplifier if the plate circuit is tuned to a harmonic of the input signal
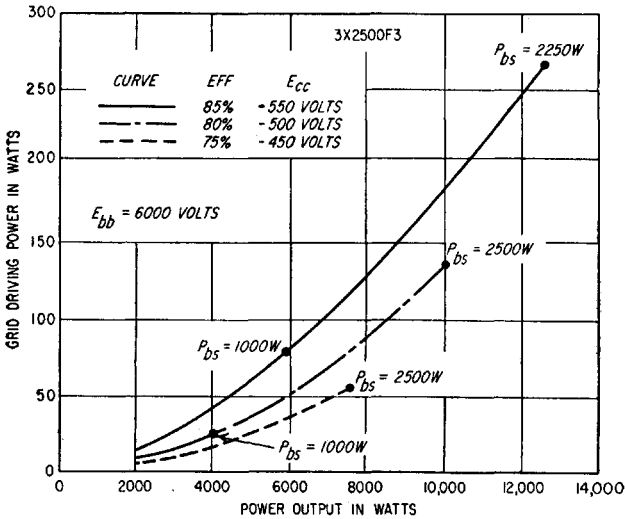
FIG. 4.20. Power relationships for a typical class C amplifier.

and if the grid circuit is tuned to the fundamental frequency. For the most efficient operation, the angle of plate current conduction should be reduced and the value of plate-load impedance increased with respect to fundamental operation. Tabulated in Table 4.2 are recommended values of θ_p and the relative plate-load impedances for the amplification of the 2nd, 3rd, and 4th harmonics. Although more grid excitation power is required for harmonic operation, the power output will be less than that achievable with fundamental operation (see Sec. 4.4).

TABLE 4.2. TABULATION OF RECOMMENDED PLATE CURRENT CONDUCTION ANGLES
AND RELATIVE PLATE-LOAD RESISTANCES FOR FUNDAMENTAL AND HARMONIC
OPERATION

	Fund.	Harmonic			
		2nd	3rd	4th	5th
θ_p (electrical degrees).........................	120–150	90–120	80–110	70–90	60–75
r_b (assuming 1.0 for fundamental operation, approx.).................................	1.0	1.5	2.5	3.3	4.0

4.4j. Class C Amplifier Design. The analysis of a class C amplifier is ordinarily carried out on a single-tube basis. If parallel or push-pull operation is to be used, the power output, grid drive, and power-supply requirements are multiplied by the

number of tubes used. A typical design procedure with associated considerations is as follows:

1. *Determine the power output P_o.* This will include the power lost in the plate tank circuit and loss in coupling network, and if operation is at frequencies above approximately 50 Mc, the dielectric loss in the insulating material of the tube envelope must be considered.

2. *Calculate the approximate plate dissipation P_{ba} per tube.*

$$P_{ba} \simeq \frac{kP_o}{n} \qquad (4.39)$$

where k = a factor expressing ratio of plate dissipation to power output (see Table 4.3)
n = number of tubes used in amplifier

TABLE 4.3. TYPICAL RATIOS OF THE PLATE DISSIPATION
TO THE POWER OUTPUT PER TUBE

k	*Harmonic*
0.33	Fund.
1.00	2nd
2.33	3rd
3.44	4th
4.35	5th

3. *Choose a tube which fulfills the performance requirements.* The plate dissipation capability of the tube must be greater than the value determined in step 2. Several other considerations in determining the proper tube are:

 a. Frequency of operation
 b. Plate, screen-grid, and filament voltages required
 c. Cooling requirements
 d. Physical characteristics
 e. Grid driving power
 f. Vibration and shock

4. *Establish the plate supply voltage E_{bb}.* In the case of a power tetrode or pentode operating in the vhf and uhf frequency bands, the combination of (1) higher screen-grid voltage, (2) reduced plate voltage, and (3) heavy loading of the plate circuit will minimize tube envelope dielectric heating due to the reduced plate voltage swing. The plate efficiency will be reduced somewhat, but the tube life will be increased.

Occasionally, it is desirable to operate a tube at voltages other than those shown on the published constant-current characteristic curves. A conversion method which is relatively simple and based on the $\frac{3}{2}$ power law can be used provided secondary emission does not affect the current values. To convert the tube characteristics to a new scale, the screen-grid, control-grid, and plate voltages must be multiplied by the desired scale factor, and the associated constant current curves must be multiplied by the same scale factor raised to the three-halves power. For example, if the tube characteristics are given for a screen-grid voltage of 300 volts and operation at 600 volts is desired, the grid and plate voltages must be multiplied by 2 and the constant current curves must be multiplied by $2^{1.5}$, or 2.83. Table 4.4 gives a number of conversion voltage and current scale factors.

5. *Calculate the approximate average plate current I'_{ba} per tube.*

$$I'_{ba} \simeq \frac{P_o/n + P_{ba}}{E_{bb}} \qquad \textbf{(4.40)}$$

This should not exceed the maximum plate current rating of the tube.

TABLE 4.4. TABULATION OF SEVERAL VOLTAGE AND CURRENT SCALE FACTORS FOR
USE IN PLOTTING CONSTANT CURRENT CHARACTERISTICS TO DIFFERENT
SCALES

Voltage Scale Factor	Current Scale Factor
0.25	0.125
0.50	0.35
0.75	0.65
1.00	1.00
1.50	1.84
2.00	2.83
2.50	3.95
3.00	5.20

6. *Determine the maximum instantaneous plate current $i_{b(\max)}$ per tube.*

$$i_{b(\max)} \simeq 4I'_{ba} \tag{4.41}$$

Equation (4.41), although not rigorous, expresses the approximate relationship between the peak plate current $i_{b(\max)}$ and the average plate current. It is necessary to make an approximation of this type since a tentative value of $i_{b(\max)}$ is necessary to perform the graphical analysis.

7. *Locate point A on the constant current plate characteristics.* Point A (see Fig. 4.21) establishes the peak positive grid-to-cathode voltage $e_{c(\max)}$, the peak instantaneous plate, screen-grid, and control-grid currents, and the minimum instantaneous plate potential $e_{b(\min)}$. Point A must lie on the constant plate current line which has a value equal to that determined in step 6. In the case of triodes having ample grid excitation power, point A should be located near the "diode line," i.e., the straight line on the tube characteristics defined by the relationship $E_c = E_b$. However, if driving power is limited, the minimum instantaneous plate potential $e_{b(\min)}$ should be somewhat higher than the peak-positive grid-to-cathode voltage $e_{c(\max)}$. The result, therefore, is the displacement of point A from the diode line. With tetrodes, point A should be located on the constant current plate line near the point where the curve turns sharply upward. If point A were to be located at a lower plate voltage, the screen and control grid currents would be excessive and the increase in power output would be relatively small.

8. *Establish the bias supply voltage E_{cc} and locate point G on the constant current characteristics.* Point G is determined by E_{bb} and E_{cc}. However, the value of the grid bias is dependent on the desired plate current conduction angle θ_p. Therefore, with reference to Table 4.2, it is necessary to select arbitrarily a value for θ_p and determine graphically the value of E_{cc} which causes the plate current conduction angle to be equal to the selected value. The use of Eq. (4.42) to determine the approximate value of E_{cc} will minimize the number of trial-and-error steps in the graphical determination of E_{cc}.

$$E_{cc} \simeq - \frac{|E_{co}| + e_{c(\max)} \cos (\theta_p/2)}{1 - \cos (\theta_p/2)} \tag{4.42}$$

where E_{co} = grid bias which will cause plate current cutoff at operating plate supply voltage

As an example, consider the tube whose characteristics are given in Fig. 4.21 and assume that (1) point A has been previously located, (2) the plate supply voltage is equal to 2,000 volts, and (3) the desired value of the plate current conduction angle θ_p has been established as 140° from Table 4.2. To apply Eq. (4.42), the values of

$e_{c(max)}$ and E_{co} must be known. Point A establishes $e_{c(max)}$ as being equal to 255 volts, and the plate current cutoff bias for a 2,000-volt plate supply voltage is -375 volts. Therefore,

$$E_{cc} \simeq -\frac{375 + 255 \cos (140/2)}{1 - \cos (140/2)}$$

$$\simeq -702 \text{ volts}$$

Point G is located at the intersection of the 2,000-volt plate voltage line and the -702-volt grid voltage line. A tentative operating line is then drawn between points
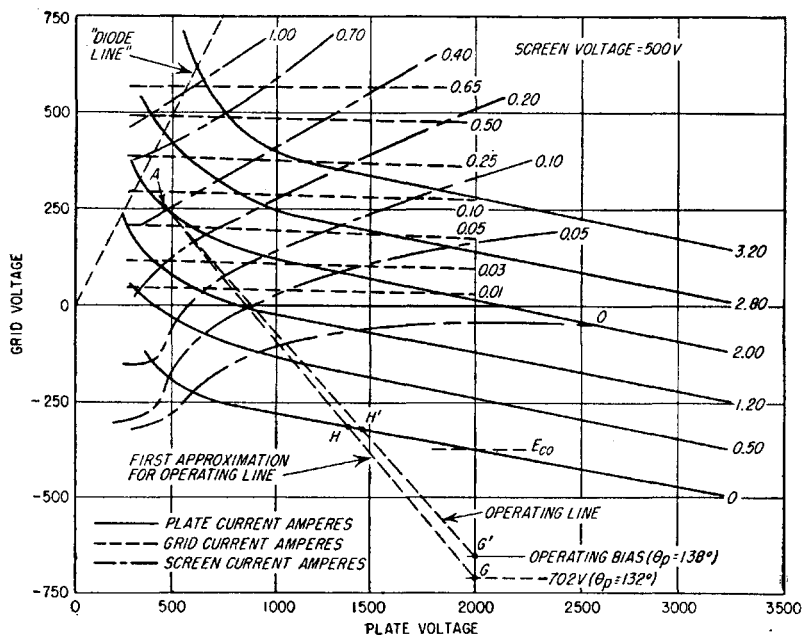
Fig. 4.21. Typical constant current characteristics for a power tetrode.

A and G. The conduction angle θ_p for the calculated value of E_{cc} can be determined by means of Eq. (4.43), and the calculated value can be compared with the value of θ_p originally selected.

$$\theta_p = 2 \cos^{-1} (HG/AG) \tag{4.43}$$

where point H = intersection of operating line and zero plate current line
The terms AG and HG are linear distances measured along the operating line and may be in any units, e.g., inches, centimeters, etc. The value of θ_p as established by the tentative operating line AG and Eq. (4.43) is equal to 132°. Since this angle is less than the assumed value of 140°, it is necessary to make the bias less negative in order to increase the angle of conduction. Another operating line is, therefore, drawn from point A to point G' which is slightly more positive than G. Based on the values of $H'G'$ and AG', another calculation of the value of θ_p must be made. If the calculated angle is significantly different than 140°, it will be necessary to make another estimate as to the proper position of point G. This process must be repeated until the calculated value of θ_p is approximately equal to the assumed value, i.e., 140°. The line AG' in Fig. 4.21 will provide a plate current conduction angle equal to 138°.

9. *Calculate the peak grid excitation voltage* $e_{g(max)}$. This is equal to the absolute value of the grid bias E_{cc} plus the peak positive grid-to-cathode voltage $e_{c(max)}$.

10. *Determine the d-c and fundamental components of plate current and the d-c components of both the screen-grid and control-grid currents.* Equations (4.44) to (4.47) give the d-c current, peak value of the fundamental current, and the approximate peak values of the 2nd and 3rd harmonic currents. The same equations can be used for the plate, screen-grid, and control-grid currents; however, the appropriate values must be substituted for A, B, C, etc.

$$I_{dc} = 0.0833(0.5A + B + C + D + E + F) \qquad (4.44)$$
$$I_{H1} = 0.0833(A + 1.93B + 1.73C + 1.41D + E + 0.52F) \qquad (4.45)$$
$$I_{H2} \simeq 0.0833(A + 1.73B + C - E - 1.73F) \qquad (4.46)$$
$$I_{H3} \simeq 0.0833(A + 1.41B - 1.41D - 2E - 1.41F) \qquad (4.47)$$

where A, B, C, etc. = values of currents read off operating line at 15° intervals of grid excitation voltage

This method is usually referred to as *Chaffee's harmonic analysis.*[1]

Points B to F can be located by calculating the values of the grid excitation voltage at 15° intervals and graphically establishing the associated values of control-grid, screen-grid, and plate currents along the operating line. Another method (Fig. 4.22) is to locate the points on the operating line so that the linear distances along the operating line are defined as follows:

$$GB = 0.966\ GA \qquad GD = 0.707\ GA$$
$$GC = 0.866\ GA \qquad GE = 0.500\ GA$$
$$GF = 0.259\ GA$$

After points B, C, D, E, and F have been located on the operating line, the associated control-grid, screen-grid, and plate current values can be established. This usually requires a certain amount of interpolation. To obtain the plate current d-c and fundamental values, the plate current values at the referenced points can be substituted into Eqs. (4.44) and (4.45). Equation (4.44) is also used to determine the average control-grid current I_{ca} and average screen-grid current I_{c2a}, except that the control-grid and screen-grid current values are taken from the proper points along the operating line for use in the equation.

After the d-c currents have been determined, they should be compared with the maximum current ratings specified by the tube manufacturer. If the calculated current values are too high, either the plate current conduction angle or the peak instantaneous current point A should be decreased.

The r-f power output can be determined from Eq. (4.24).

11. *Determine the power which must be supplied by the plate circuit power supply.*

$$P_{dc} = nE_{bb}I_{ba} \qquad (4.48)$$

where n = number of tubes

I_{ba} = average plate current per tube (Eq. 4.44)

12. *Determine the plate circuit efficiency.*

$$\eta = \frac{P_o \times 100}{P_{dc}} \qquad (4.49)$$

13. *Determine the plate dissipation per tube [Eq. (4.27)].*

14. *Determine the screen dissipation per tube [Eq. (4.34)].*

15. *Determine the total grid driving power P_t using Eq. (4.38) as modified by the number of tubes involved.*

[1] E. L. Chaffee, A Simplified Harmonic Analysis, *Rev. Sci. Instr.*, vol. 7, p. 384, October, 1936.

16. *If grid-leak bias is used, determine the value of the grid resistor.*

$$R_1 = \frac{|E_{cc}|}{nI_{ca}}$$ (4.50)

where n = number of tubes

I_{ca} = average control grid current per tube

17. *The plate and grid tank circuits should be designed in accordance with Sec. 4.4e.*

Example 4.3

Design a class C r-f power amplifier to deliver 300 watts into a 50-ohm load at 20 Mc.

1. The power output that the plate must deliver is 300 watts plus approximately 10 per cent for losses. Therefore, P_o = 330 watts.

2. The approximate required plate dissipation can be determined from Eq. (4.39).

$$P_{bs} \simeq \frac{0.33 \times 330}{1}$$

$$\simeq 109 \text{ watts}$$

3. Assume that a 4-125A tube type is to be used. Its plate dissipation rating is 125 watts.

4. The maximum ratings for the 4-125A are as follows:

D-c plate voltage	3,000 volts
D-c screen voltage..................	400 volts
D-c grid voltage (negative)..........	500 volts
D-c plate current...................	225 ma
Plate dissipation...................	125 watts
Screen dissipation..................	20 watts
Grid dissipation....................	5 watts

Assume a plate voltage of 2,500 volts and a screen-grid voltage of 350 volts.

5. Determine the approximate average plate current I_{ba}' from Eq. (4.40).

$$I_{ba}' \simeq \frac{330 + 109}{2,500}$$

$$\simeq 0.176 \text{ amp}$$

6. Determine the maximum instantaneous plate current using Eq. (4.41).

$$i_{b(max)} \simeq 4 \times 176$$
$$\simeq 704 \text{ ma}$$

7. Locate point A on the constant current characteristics shown in Fig. 4.22. Along the interpolated 704-ma constant plate current curve, it appears the plate voltage can be reduced to a value as low as 320 volts without causing an abnormal increase in the screen-grid current. Therefore, let $e_{b(min)}$ be equal to 320 volts, which establishes $e_{c(max)}$ equal to 120 volts.

8. Calculate a tentative value of E_{cc} using Eq. (4.42). With reference to Table 4.2, a satisfactory value of θ_p would be 140°. Therefore

$$E_{cc} \simeq -\frac{82 + 120 \cos (140/2)}{1 - \cos (140/2)} = -181 \text{ volts}$$

Locate point G on the tube characteristic and use Eq. (4.43) to establish whether or not the calculated value of θ_p agrees with the assumed value of 140°.

$$\theta_p = 2 \cos^{-1} (HG/AG) = 139°$$

Therefore, the location of point G can be considered to be satisfactory.

9. The peak grid voltage is equal to 181 + 120, or 301 volts.

10. Determine the current points A, B, C, etc., on Fig. 4.22 and calculate the electrode currents [Eqs. (4.44) and (4.45)]. The values as read from Fig. 4.22 are tabulated in Table 4.5.

TABLE 4.5. TABULATION OF THE CURRENT VALUES ON THE OPERATING LINE SHOWN
IN FIG. 4.22

	A	B	C	D	E	F
Plate current.....................	0.702	0.700	0.620	0.400	0.10	0
Screen current...................	0.200	0.155	0.060	0.009	0	0
Grid current.....................	0.050	0.045	0.030	0.013	0	0
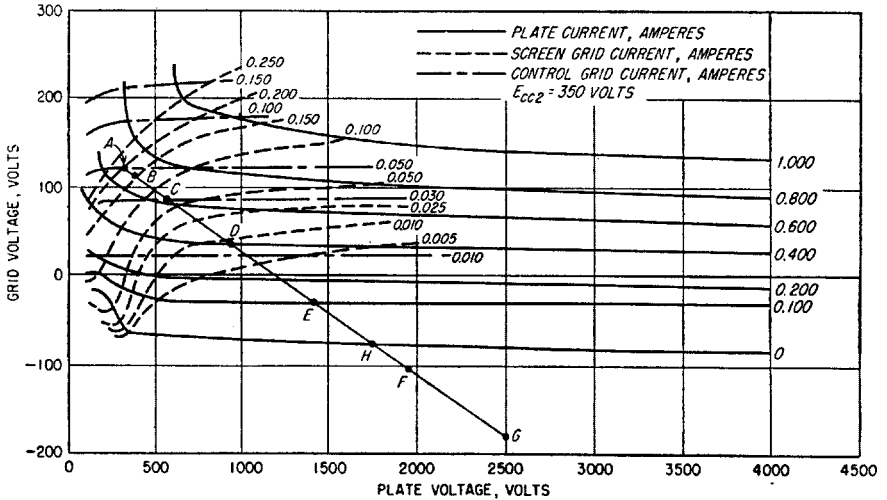
FIG. 4.22. 4-125A constant current characteristics.

Plate currents

$$I_{ba} = 0.0833[0.5(0.702) + 0.700 + 0.620 + 0.400 + 0.100]$$
$$= 181 \text{ ma}$$
$$I_{H1} = 0.0833[0.702 + 1.93(0.700) + 1.73(0.620) + 1.41(0.400) + 0.100]$$
$$= 316 \text{ ma}$$

Screen-grid current

$$I_{c2a} = 0.0833[0.5(0.200) + 0.155 + 0.060 + 0.009)]$$
$$= 27 \text{ ma}$$

Control-grid current

$$I_{ca} = 0.0833[0.5(0.050) + 0.045 + 0.030 + 0.013]$$
$$= 9.41 \text{ ma}$$

Calculate the power output [Eq. (4.24)].

$$P_o = \frac{e_{p(\max)}I_{H1}}{2} = \frac{2,180 \times 0.316}{2} = 344 \text{ watts}$$

This is slightly higher than the required power output of 330 watts; however, recalculation is not necessary since the desired power output can be obtained with a slight reduction in plate circuit loading.

11. Determine the power input to the plate circuit.

$$P_{dc} = 2,500 \times 0.181 = 453 \text{ watts}$$
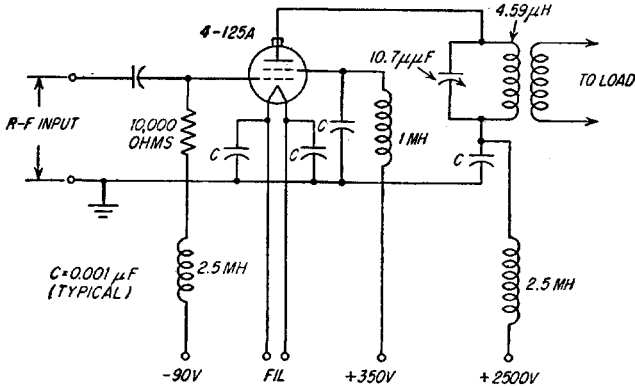
Fig. 4.23. Class C power amplifier for Example 4.3.

12. Determine the plate circuit efficiency.

$$\eta = 344/453 \times 100 = 75.9\%$$

13. Determine the plate dissipation.

$$P_o - P_{dc} = 109 \text{ watts}$$

14. Determine the screen-grid dissipation.

$$E_{cc2}I_{c2a} = 350 \times 0.027 = 9.45 \text{ watts}$$

15. Determine the grid dissipation P_{c1}, the power supplied to the bias source P_1, and the total driving power P_t.

$$\text{Grid dissipation} = e_{c(\max)}I_{ca}$$
$$= 120 \times 0.00941$$
$$= 1.13 \text{ watts}$$
$$\text{Power supplied to the bias source} = E_{cc} \times I_{ca}$$
$$= 181 \times 0.00941$$
$$= 1.70 \text{ watts}$$

From Eq. (4.38)

$$\text{Total grid driving power} = 1.1(1.13 + 1.70)$$
$$= 3.11 \text{ watts}$$

The driver stage should be capable of supplying somewhat more than the calculated power, e.g., 5 to 10 watts.

16. Assume that a fixed bias of 90 volts is used and the rest of the bias is developed with a resistor R_1.

$$R_1 = \frac{181 - 90}{0.00941} = 9,670 \text{ ohms}$$

Therefore, a 10,000-ohm resistor can be used.

17. Design the tank circuit. Assume that a loaded Q of 12 is to be used. The plate load resistance r_b is, therefore, determined from Eq. (4.28).

$$r_b = \frac{2,500 - 320}{0.316} = 6,900 \text{ ohms}$$

and from Eq. (4.31)

$$C_p = \frac{Q_o}{2\pi f_o r_b} = \frac{12}{2\pi \times 20 \times 10^6 \times 6,900}$$
$$= 13.8 \ \mu\mu f$$

The output capacitance of the tube is listed as 3.1 $\mu\mu f$; therefore the tuning capacitor would have to be equal to 10.7 $\mu\mu f$. To obtain resonance at 20 Mc, the plate inductance must be equal to 4.59 μh. Assuming link coupling in the output, the coupling should be adjusted until the calculated d-c plate current I_{ba} flows. The final circuit is as shown in Fig. 4.23.

Modulation

5.1. Introduction. Modulation may be defined as the alteration of one or more of the characteristics of a signal as a function of another signal. Commonly, the signal being modulated is a sine wave of constant amplitude and is usually referred to as the carrier. The signal which varies some parameter of the carrier is known as the modulating signal. The parameters of a sine wave which may be varied are the amplitude, the frequency, and the phase. Other types of modulation may be applied to special signals, e.g., pulse-width and pulse-position modulation of recurrent pulses.

Modulation, demodulation or detection, and heterodyne action are very closely related processes. Each process involves generating the sum and/or difference frequencies of two or more sinusoids by causing one signal to vary as a direct function (product) of the other signal or signals. The multiplication of one signal by another can only be accomplished in a nonlinear device. This is readily seen by considering any network where the output signal is some function of the input signal e_1. For example,

$$e_o = f(e_1) \tag{5.1}$$

In any perfectly linear network, this requires that

$$e_o = ke_1$$

and, assuming two different input signals

$$e_o = k(E_a \cos \omega_a t + E_b \cos \omega_b t) \tag{5.2}$$

where k = constant

In this case the output signal contains only the two input-signal frequencies. However, if the output is a nonlinear function of the input it can, in general, be represented by a series expansion of the input signal. For example, let

$$e_o = k_1 e_1 + k_2 e_1^2 + k_3 e_1^3 + \cdots + k_n e_1^n \tag{5.3}$$

When e_1 contains two frequencies, e_o will contain the input frequencies and their harmonics plus the products of these frequencies. These frequency products can be expressed as sum and difference frequencies. Thus, all modulators, detectors, and mixers are of necessity nonlinear devices. The principal distinction between these devices is the frequency differences between the input signals and the desired output signal or signals. For example, amplitude modulation in general involves the multiplication of a high-frequency carrier by lower-frequency modulation signals to produce sideband signals near the carrier frequency. In a mixer, two high-frequency signals are multiplied to produce an output signal at a frequency which is the difference between the input-signal frequencies. In a detector for amplitude modulation, the carrier is multiplied by the sideband signals to produce their different frequencies at the output.

5.2. Amplitude Modulation. In amplitude modulation the instantaneous amplitude of the carrier is varied in proportion to the modulating signal. The modulating signal may be a single frequency, or, more often, it may consist of many frequencies of various amplitudes and phases, e.g., the signals comprising speech. For a carrier modulated by a single-frequency sine wave of constant amplitude, the instantaneous signal $e(t)$ is given by

$$e(t) = E(1 + m \cos \omega_m t) \sin (\omega_c t + \phi) \tag{5.4}$$

where E = peak amplitude of unmodulated carrier

 m = modulation factor as defined below

 ω_m = frequency of modulating voltage, radians/sec

 ω_c = carrier frequency, radians/sec

 ϕ = phase angle of carrier, radians

The instantaneous carrier amplitude is plotted as a function of time in Fig. 5.1. The *modulation factor m* is defined for unsymmetrical modulation in the following manner.

$$m = \frac{E_{max} - E}{E} \qquad \text{(upward or positive modulation)} \qquad (5.5)$$

$$m = \frac{E - E_{min}}{E} \qquad \text{(downward or negative modulation)} \qquad (5.6)$$

The maximum downward modulation factor, 1.0, is reached when the modulation peak reduces the instantaneous carrier envelope to zero. The upward modulation factor is unlimited.
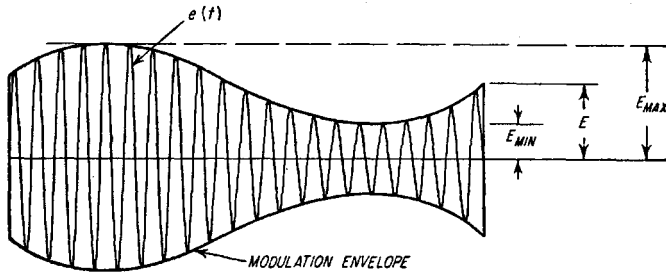
FIG. 5.1. Amplitude-modulated carrier.

5.2a. Signal Spectrum and Modulation Sidebands. The modulated carrier described by Eq. (5.4) can be rewritten as follows:

$$e(t) = E(1 + m \cos \omega_m t) \sin (\omega_c t + \phi)$$
$$= E \sin (\omega_c t + \phi) + \frac{mE}{2} \sin [(\omega_c + \omega_m)t + \phi] + \frac{mE}{2} \sin [(\omega_c - \omega_m)t + \phi] \quad (5.7)$$

Thus, the amplitude modulation of a carrier by a sine wave has the effect of adding two new sinusoidal signals displaced in frequency from the carrier by the modulating frequency. These sinusoids resulting from modulation of the carrier are known as *sidebands*. The spectrum of the modulated carrier is shown in Fig. 5.2a. For a complex modulating signal $G(t)$ the modulated carrier spectrum is given by

$$F(\omega) = \mathfrak{F}\{E[(1 + mG(t)] \sin (\omega_c t + \phi)\}$$
$$= \frac{E}{2\pi} \int_{-\infty}^{\infty} [1 + mG(t)] \sin (\omega_c t + \phi)e^{-i\omega t} dt \qquad (5.8)$$

where $F(\omega)$ = Fourier transform of the time function = $\mathfrak{F}[f(t)]$

When the modulating signal is a regularly recurrent time function and, therefore, expressible as a Fourier series, each frequency term forms a pair of sidebands displaced symmetrically from the carrier by the modulating frequency. Thus, when the modulating signal is

$$e_m(t) = A \cos \omega_a t + B \cos (\omega_b t + \theta)$$

the modulated carrier is given by

$$e(t) = E[1 + m_1 \cos \omega_a t + m_2 \cos (\omega_b t + \theta)] \sin (\omega_c t + \phi)$$
$$= E \sin (\omega_c t + \phi)$$
$$+ \frac{Em_1}{2} \sin [(\omega_c + \omega_a)t + \phi] + \frac{Em_1}{2} \sin [(\omega_c - \omega_a)t + \phi]$$
$$+ \frac{Em_2}{2} \sin [(\omega_c + \omega_b)t + \phi + \theta] + \frac{Em_2}{2} \sin [(\omega_c - \omega_b)t + \phi - \theta] \quad (5.9)$$

where m_1 = modulation factor for $\cos \omega_a t$
$\qquad m_2$ = modulation factor for $\cos \omega_b t$
This is illustrated in Fig. 5.2b.   There is no interaction between the various modulating frequencies provided that the modulation process is linear; i.e., even though the
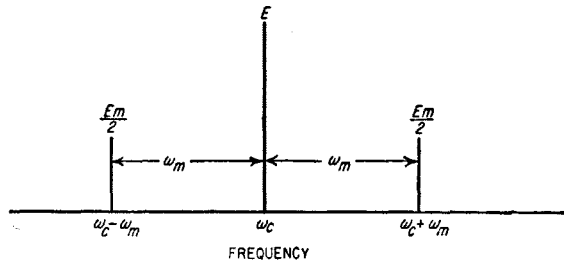
(a) CARRIER MODULATED BY A SINUSOID OF FREQUENCY ω_m
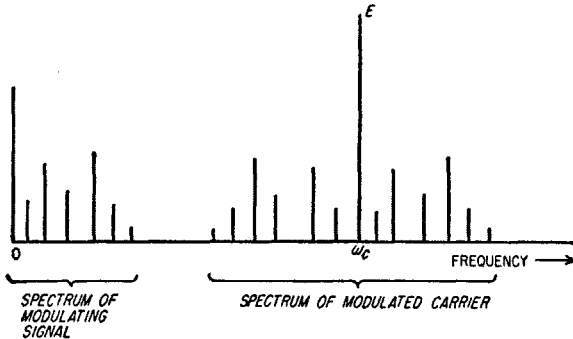
(b) CARRIER MODULATED BY A COMPLEX SIGNAL
COMPOSED OF SEVERAL SINUSOIDS

Fig. 5.2. Frequency spectrum of an amplitude-modulated carrier.

modulator is a nonlinear device, the modulated parameter of the carrier varies as a linear function of the modulating signal.

5.2b. Vector Representation of an Amplitude-modulated Carrier.   The manner in which the carrier is amplitude-modulated by the sideband frequencies is readily visualized by considering each signal frequency component $A_n \sin (\omega_n t + \phi)$ to be the projection of the rotating vector $A_n e^{j(\omega_n t + \phi)}$ on the imaginary axis (Fig. 5.3a).   The vector representation of $e(t)$ of Eq. (5.7) is shown in Fig. 5.3b.   By considering the sideband vectors as they would appear when viewed from the carrier vector, the equivalent vector diagram of Fig. 5.3c is obtained.   Here it is seen that the two sideband signals modulate the magnitude of the carrier vector at the modulating frequency.   Amplitude modulation alone exists as long as the two sideband frequency

components have equal amplitude and maintain the phase relationship with respect to the carrier indicated by Eq. (5.7) and Fig. 5.3c, i.e., the phase angle between the upper sideband and the carrier is always opposite in sign and equal in magnitude to the phase angle between the lower sideband and the carrier. If the amplitude or phase relationships are altered by any networks through which the signals pass, phase modulation of the carrier will result.

5.2c. Sideband Power. Amplitude modulation of a carrier increases the total signal (carrier plus sidebands) power P_t by the amount of power present in sideband signals. Thus,

$$P_t = KE^2 \left(1 + \sum_{n=1}^{n=n} \frac{m_n{}^2}{2}\right) \tag{5.10}$$

where m_n = modulation factor of nth pair of sidebands
E = peak amplitude of unmodulated carrier voltage
K = constant

The total signal power present in a carrier amplitude modulated by any complex waveform is given by

$$P_t = P_c \left(1 + \frac{m^2}{k^2}\right) \tag{5.11}$$

where P_c = average carrier power with modulation (in unsymmetrical modulation the average carrier power will change with modulation)

m = modulation factor at peak of modulating waveform

k = ratio of peak to rms value of modulating voltage or current

From Eqs. (5.10) and (5.11) it is seen that amplitude modulation of a carrier may increase the total signal power considerably. The power present in the sidebands is supplied by the modulator.

5.2d. Modulation Distortion. Although the modulation of a carrier can only be accomplished in a nonlinear device, the envelope of the carrier can be made identical to the modulation signals if the modulation network characteristics are such that only first-order product terms of the carrier and the modulation signals occur (extraneous signals present at other

(a) VECTOR REPRESENTATION OF A SINUSOID

(b) VECTOR REPRESENTATION OF A CARRIER AND ITS MODULATION SIDEBANDS

(c) EQUIVALENT VECTOR REPRESENTATION OF A CARRIER AND ITS MODULATION SIDEBANDS

FIG. 5.3. Vector representation of a modulated carrier.

frequencies in the output are filtered out). If the modulation envelope of a modulated carrier is not identical to the modulating waveform, the modulation process has introduced distortion. Three types of distortion may occur:

1. *Amplitude distortion (nonlinear distortion).* The introduction of frequencies into the modulation envelope not present in the modulating waveform.

2. *Frequency distortion.* Variations in the relative amplitudes of the sideband signals compared to the relative amplitudes of the frequency components of the modulating waveform.

3. *Phase distortion.* Variations in the relative phases of the various sideband signals compared to the relative phases of the frequency components of the modulating waveform.

5.3. Methods of Amplitude Modulation. There are many methods by which a carrier can be amplitude-modulated. The most frequently used modulation methods are discussed in the following subsections.

5.3a. Plate Modulation of a Class C Amplifier. The most widely used means of obtaining amplitude modulation of a carrier is by plate modulation of a class C amplifier. Plate modulation possesses the advantages of ease of adjustment and high-percentage modulation with good linearity. It possesses the disadvantage of requiring a modulator capable of delivering an average power approximately equal to one-half the unmodulated plate input power of the class C amplifier. During the positive peak of a 100 per cent modulating signal the total plate input power (carrier plus modulation) and the r-f output power are four times their values for zero modulation.

Typical circuits used for plate modulation are illustrated in Fig. 5.4. The voltage applied to the plate of the class C amplifier is made to vary in proportion to the modulating signal by adding a modulating voltage in series with the plate supply voltage. The analysis and design of class C amplifiers is described in detail in Sec. 4.4. The manner in which variation of the plate supply voltage can accomplish linear modulation of the r-f output voltage in a tuned amplifier is best understood by plotting the operating line of the tube on the constant-current characteristic curves for the tube (see Sec. 4.4*j*). This is illustrated in Fig. 5.5 for a typical transmitting triode. The operating line is shown for zero modulation and for the peak and trough of a 100 per cent modulating signal when a combination of fixed and grid-leak bias is used. The operating line AA' in Fig. 5.5 represents the locus of all of the instantaneous values of grid and plate voltages and currents during the positive half cycle of the applied r-f grid voltage with zero modulation. The remainder of the cycle is obtained by extending the line AA' an equal distance below the point A', where point A' is established by the plate supply voltage and the bias with no modulation. The specified operating line AA' can be satisfied by only one value of plate-circuit impedance at the frequency of the r-f voltage as explained in Sec. 4.4*e*. At the crest of the modulation, the new operating line BB' forms the locus of all instantaneous values of grid and plate voltages and currents. The operating line CC' forms the locus of all instantaneous values of grid and plate voltages at the trough of the modulation cycle. On this line the instantaneous plate voltage is zero and the instantaneous plate current is very small. The average grid current increases rapidly as the plate voltage is reduced to zero, and the negative grid-leak bias increases sharply. The exact location of CC' on the zero plate voltage line is difficult to establish graphically for high percentages of modulation because of the inaccuracy of the tube characteristic curves at zero plate volts.

To obtain the best modulation linearity, a portion of the bias for the grid is obtained from a fixed bias supply and part is obtained from a grid-leak resistor. The values of grid-leak bias and fixed bias which will provide the best modulation linearity over the complete modulation cycle are determined closely by finding the values of fixed bias and grid-leak bias which locate the operating line BB' at the crest of the modulation cycle so as to provide an output power which is $(1 + m)^2$ times the unmodulated carrier power. Thus, in the example of Fig. 5.5, a change in the grid-leak bias of 50 volts is required between AA' and BB' to obtain linear modulation. When the values of grid-leak bias and fixed bias necessary to provide modulation linearity between operating lines AA' and BB' are determined, the location of operating line CC' at the trough of the modulation and all other intermediate operating lines are established. In practice, some readjustment of the fixed bias and grid-leak bias may be

desirable to obtain the best modulation linearity over the entire modulation cycle. By making the time constant of the grid resistor R_c and its r-f bypass C_c small compared to the time of one cycle of the highest modulation frequency, the self-bias developed across R_c by the average grid current I_c will follow any variations in I_c caused by the modulation of the tube plate voltage. The average value of grid current decreases as the plate supply voltage is increased for a fixed r-f grid voltage
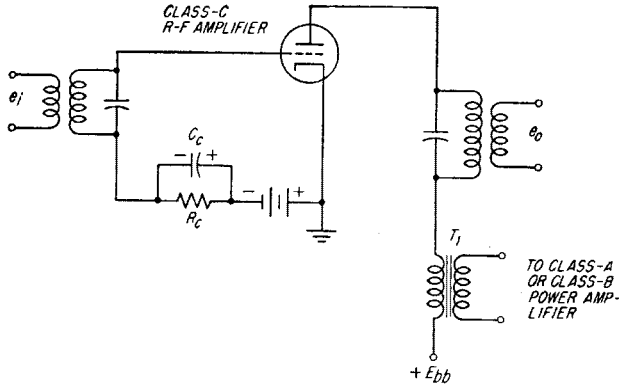
(a) TRANSFORMER COUPLING
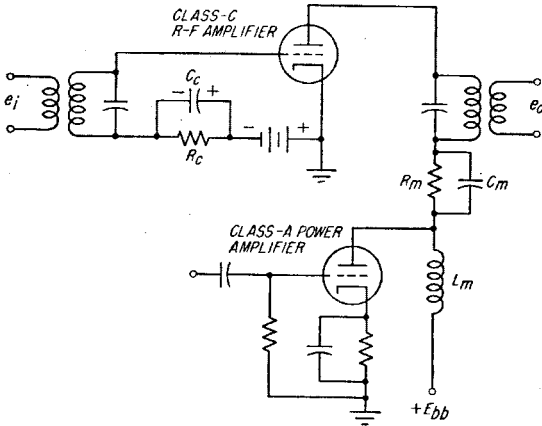
(b) IMPEDANCE COUPLING

Fig. 5.4. Methods of plate modulation of a class C r-f amplifier.

and bias. As a result of the reduction in average grid current at higher plate-supply voltages, the self-bias is reduced and the grid bias is lowered. The reduced grid bias allows the same r-f grid voltage to increase the peak plate current and, therefore, the component of plate current at the amplifier resonant frequency. The plate-current conduction angle is also increased with increased plate supply voltage, causing an additional rise in the fundamental component of plate current. By proper adjustment of the values of fixed grid bias and grid-leak bias the fundamental component of plate current can be made to increase very linearly with plate supply voltage, thus resulting in linear modulation of the r-f carrier. The effect of fixed bias and self-bias on the

linearity of modulation can be seen in Fig. 5.6a.   Here, the peak amplitude of the output r-f voltage is plotted as a function of the instantaneous plate supply voltage for the tube characteristic of Fig. 5.5 for fixed bias and for the optimum ratio of fixed and grid-leak bias.   By using only grid-leak bias, the r-f output voltage will increase more rapidly with increasing plate voltage than desired for linear modulation.   The linearity of modulation can be optimized in practice by adjusting the relative values of fixed bias and grid-leak bias while observing the modulation Lissajous figure as shown in Figs. 5.6b and 5.6c.
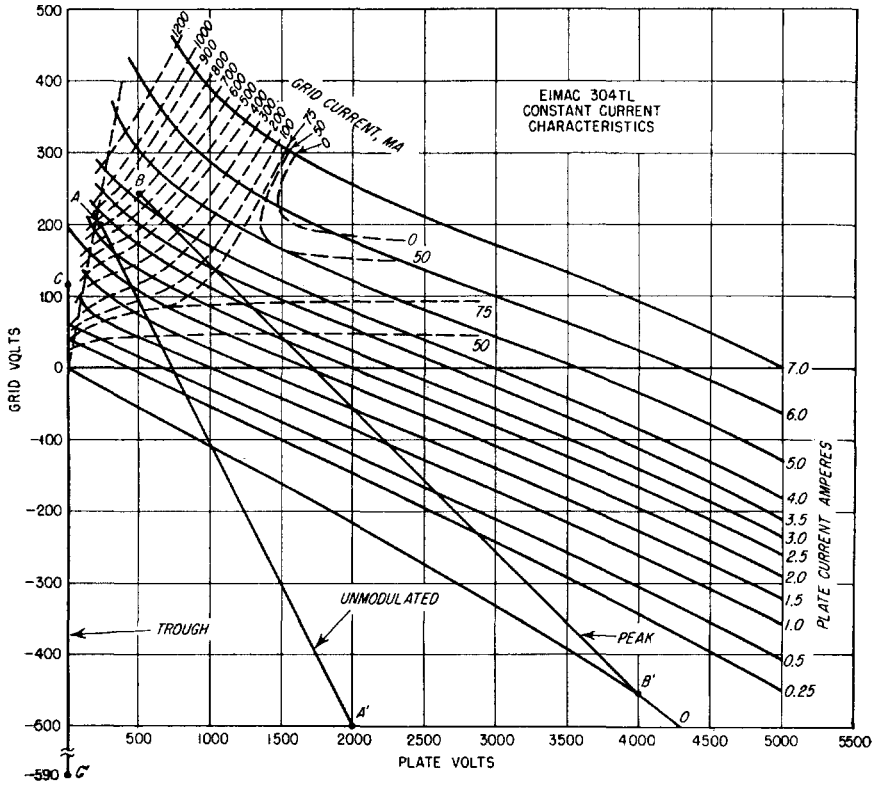
FIG. 5.5. Operating lines for a plate-modulated triode class C r-f amplifier having optimum ratio of fixed bias to grid-leak bias.

When plate-modulated by a sinusoidal signal, the plate input power, plate dissipation, and plate output power of a class C r-f amplifier are given by Eqs. (5.12), (5.13) and (5.14), respectively, assuming constant tube efficiency.

$$P_i = P_o \left(1 + \frac{m^2}{2}\right) \qquad \text{(average plate circuit input power)} \qquad (5.12)$$

$$P_d = P_o (1 - \eta) \left(1 + \frac{m^2}{2}\right) \qquad \text{(average plate dissipation)} \qquad (5.13)$$

$$P_{rf} = P_o \eta \left(1 + \frac{m^2}{2}\right) \qquad \text{(average output power)} \qquad (5.14)$$

where P_o = plate input power under conditions of zero modulation
η = tube efficiency under condition of zero modulation
m = modulation factor

In practice the plate efficiency η is not constant throughout the modulation cycle. The efficiency is a maximum near zero modulation and decreases somewhat during both the crest and trough of the modulation cycle. As a result, the average plate current, and hence the d-c plate-input power, will increase somewhat during linear plate modulation of a class C amplifier. The plate dissipation will also be somewhat higher than indicated by Eq. (5.13). The increase in plate input power during modulation is supplied by the modulator. For single-frequency sinusoidal modula-
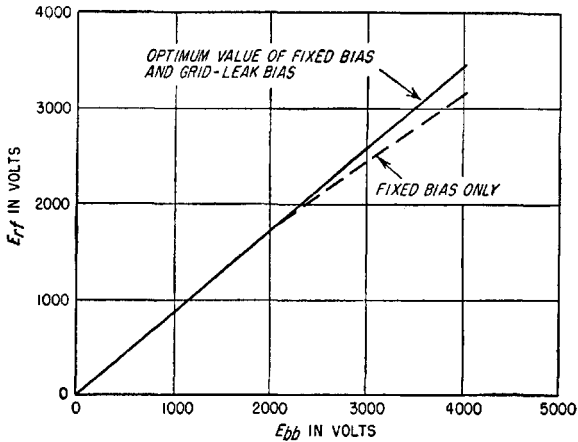
(a) MODULATION LINEARITY AS A FUNCTION OF TYPE OF BIAS EMPLOYED

$$MODULATION\ FACTOR\ M = \frac{B-A}{B+A}$$
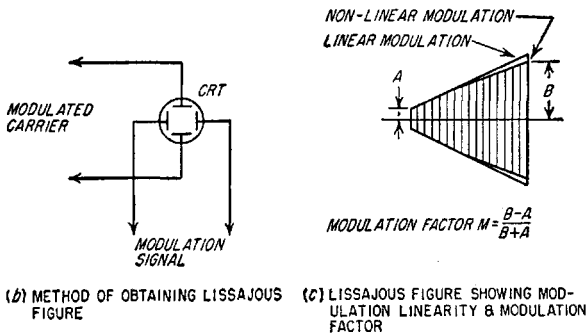
(b) METHOD OF OBTAINING LISSAJOUS FIGURE

(c) LISSAJOUS FIGURE SHOWING MODULATION LINEARITY & MODULATION FACTOR

FIG. 5.6. Modulation linearity of a typical plate-modulated triode class C r-f amplifier.

tion of 100 per cent, the modulator is required to supply somewhat more than one-half the plate-input power required by the amplifier in the absence of modulation.

Since the plate dissipation of the class C amplifier tube is increased by modulation, the carrier power level must be somewhat below the maximum value allowable without modulation. The maximum plate dissipation of a plate-modulated class C amplifier without modulation is ordinarily specified by the tube manufacturer to be two-thirds of the maximum plate dissipation for the tube. This prevents exceeding the full rated plate dissipation during 100 per cent sinusoidal modulation.

The following procedure can be used to determine graphically the operating lines for a class C triode amplifier which is to be linearly modulated:

1. Select a power output such that the plate dissipation does not exceed two-thirds

of the maximum rated value for the tube, assuming that 100 per cent modulation is desired. Select a plate supply voltage so that the maximum permissible plate voltage for the tube is not exceeded during the crest of modulation. Determine the operating line for the tube to be used for the zero modulation condition by the design procedure given in Sec. 4.4j.

The result of this step will be the determination of the plate supply voltage, peak a-c plate voltage, peak a-c plate current, average plate current, plate-circuit impedance, grid-bias voltage, average grid current, peak r-f grid voltage, grid driving power, and plate dissipation for the zero modulation condition. These are represented by an operating line such as AA' in Fig. 5.5.

2. Determine the plate voltage for the maximum modulation factor to be utilized.

$$E_{b(\text{max})} = (1 + m)E_{bo}$$

where E_{bo} = zero modulation value of E_b

Construct a vertical line along this value of E_b on the constant-current characteristic.

3. Determine the maximum value of the peak r-f plate voltage $e_{p(\text{max})}$ assuming linear modulation.

$$e_{p(\text{max})} = (1 + m)e_{po}$$

where e_{po} = zero modulation value of peak r-f plate voltage

4. Determine the minimum instantaneous value of plate voltage during the crest of the modulation.

$$e_{b(\text{min})} = E_{b(\text{max})} - e_{p(\text{max})}$$

Construct a vertical line along this value of E_b on the constant-current characteristics.

5. Determine the peak value of the fundamental component of plate current required for linear modulation at the crest of modulation.

$$I_{H1(\text{max})} = \frac{e_{p(\text{max})}}{r_b}$$

where r_b = plate-load resistance at resonant frequency of plate circuit

6. Determine a trial value for the peak instantaneous plate current.

Locate the intersection of the $e_{b(\text{min})}$ line with the constant-grid-current curve corresponding to the grid current at the peak of the grid voltage swing on the zero modulation operating line (point A in Fig. 5.5). Denote this intersection as point B.

7. Construct a trial operating line.

Subtract the value of the peak r-f grid voltage $e_{g(\text{max})}$ from the instantaneous value of grid voltage e_c at point B. Mark the intersection of this new value of grid bias and the line corresponding to $E_b = E_{b(\text{max})}$ as point B'. Construct a trial operating line between points B and B'. Calculate the peak value of the fundamental component of plate current for this operating line from Eq. (4.45).

$$I_{H1(\text{max})} = \tfrac{1}{12}(a + 1.93b + 1.73c + 1.41d + e + 0.52f) \qquad (4.45)$$

If the value of $I_{H1(\text{max})}$ found from Eq. (4.45) for this operating line is within a few per cent of the value required for linear modulation (step 5), the selected operating line is the desired one for linear modulation within the accuracy allowed by this method. If $I_{H1(\text{max})}$ is not within a few per cent of the required value, the operating line BB' should be shifted vertically up or down as required, maintaining the same slope and same limits of $e_{b(\text{min})}$ and $E_{b(\text{max})}$ until $I_{H1(\text{max})}$ calculated from the new operating line is within the proper tolerance.

8. Determination of the proper values of fixed bias and grid-leak bias.

The required shift in grid bias between the two operating lines for linear modulation has been determined by step 7. The value of grid-leak resistance which would pro-

vide this change in bias when the tube plate voltage is varied from E_{bo} to $E_{b(max)}$ can be determined by an accurate calculation of the average value of grid current for each operating line and the use of Eq. (5.15).

$$R_c = \frac{|E_{cc1}| - |E_{cc2}|}{I_{c1} - I_{c2}} \tag{5.15}$$

where E_{cc1} = bias voltage for zero modulation
E_{cc2} = bias voltage for maximum crest of modulation
I_{c1} = average grid current for zero modulation
I_{c2} = average grid current for maximum crest of modulation

Because the change in average grid current is usually about 10 per cent of I_{c1}, approximate equations for the determination of I_{c1} and I_{c2}, such as Eq. (4.44) which is
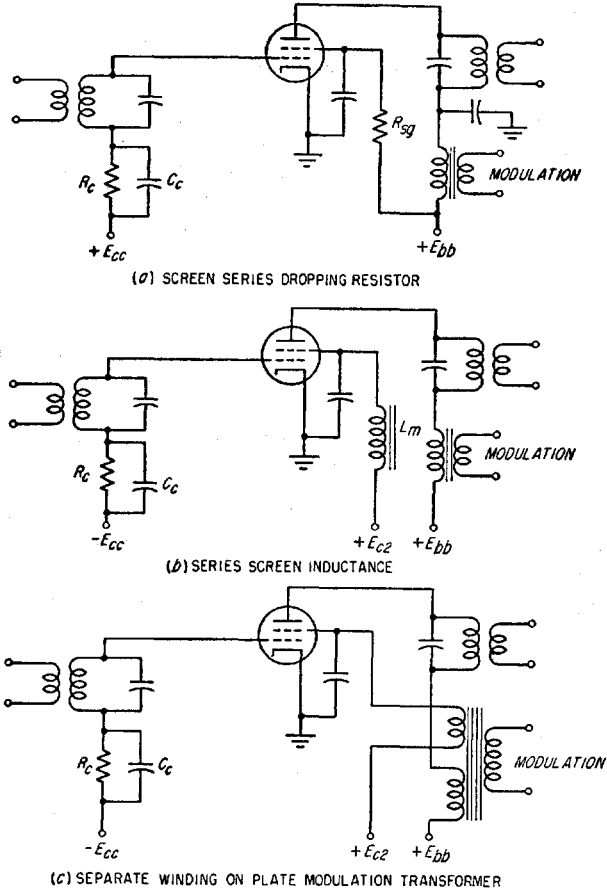
Fig. 5.7. Methods of obtaining plate and screen modulation in class C amplifier tetrodes and pentodes.

accurate to only approximately 5 per cent, cannot be used to provide values of I_{c1} and I_{c2} of sufficient accuracy for use in Eq. (5.15). A more accurate determination of average grid current is quite laborious. The proper values of fixed bias and grid-leak bias are most easily determined experimentally when the amplifier is con-

structed.   The procedure is as follows: With $R_c = 0$, set the plate voltage to E_{bo}
with a fixed bias equal to that specified at point A'.   Modulate the amplifier with a
sinusoidal voltage giving the desired modulation factor at the input to the tube.
Observe the modulation envelope Lissajous figure (see Fig. 5.6b and c).   It will
appear nonlinear as indicated in Fig. 5.6a for fixed bias.   Remove the modulation,
increase R_c, keeping the time constant R_cC_c within the proper range, and decrease
the value of fixed bias until the same value of bias is present as before.   Apply modula-
tion and check the linearity of the envelope Lissajous figure.   Vary R_c and the value
of fixed bias, always maintaining the same value of total bias without modulation,
until the optimum linearity of modulation is obtained.

The over-all efficiency of a plate-modulated class C amplifier can be increased
somewhat by applying modulation to the amplifier driving the grid of the class C
amplifier.   By varying the amplitude of the r-f driving voltage in accordance with
the plate modulation, the amount of grid-leak bias can be made to increase as plate
voltage is increased, while still maintaining linear plate modulation.   By proper
adjustment of R_c and the modulation factor of the driver carrier, the plate conduction
angle can be maintained nearly constant at an optimum value, thus keeping the plate
efficiency essentially constant throughout the modulation cycle.

In class C r-f amplifiers utilizing tetrodes and pentodes, plate modulation must be
accompanied by screen modulation if a high modulation factor is desired.   The
required amount of screen-voltage modu-
lation can usually be achieved by any of
the following methods: (1) screen voltage
obtained by a series resistor from the
unmodulated plate supply voltage; (2)
screen voltage obtained from a separate
screen supply through an inductance; (3)
screen modulated by a separate winding
on the plate modulation transformer.
These three methods are illustrated in Fig.
5.7.   In the series resistor and the series
reactor methods, screen-voltage modula-
tion is achieved by the inverse variation in
screen current with plate voltage.   This is
illustrated in Fig. 5.8 for a typical trans-
mitting tetrode.   To minimize any phase
shift between the plate and screen modu-
lation voltages in methods (1) and (2),
the screen bypass capacitor should be made
only as large as necessary to adequately
bypass the screen at the radio frequency.

Fig. 5.8. Variation in screen current with plate voltage for a typical power tetrode.

The series screen resistor method is most
frequently used because of simplicity and
because it eliminates the necessity of a separate screen supply.   The value of series
screen resistor to be used is given by

$$R_{sg} = \frac{E_{bb} - E_{c2}}{I_{c2}}$$ (5.16)

where E_{bb} = plate supply voltage
  E_{c2} = desired screen voltage without modulation
  I_{c2} = d-c screen current when screen voltage is equal to E_{c2} and in absence of
    modulation
The peak amplitude of the screen modulating voltage required for 100 per cent modula-
tion is usually about $0.75E_{c2}$.   It is normally specified by tube manufacturers for

given operating conditions. It is difficult to determine graphically the operating conditions for a plate-modulated tetrode or pentode class C amplifier because a new family of constant-current characteristics is required for each value of screen voltage.

The modulator for a class C amplifier is a power amplifier. Power amplifiers are discussed in Secs. 4.1 and 4.2. If impedance coupling is used between the modulator and the r-f amplifier as shown in Fig. 5.4b, the power amplifier must be operated class A since a single-ended amplifier is required. Under zero modulation conditions, the quiescent plate current of the power amplifier tube will be high and the combined efficiency of the r-f amplifier and the modulator rather low. For this reason, impedance coupling is seldom used except in low-power applications. When impedance coupling is used, the d-c voltage at the plate of the r-f amplifier must be reduced below E_{bb} for the peak modulating voltage from the class A amplifier to equal the r-f amplifier plate voltage for 100 per cent modulation.

When transformer coupling is used between the modulator and the r-f amplifier, the modulator is usually a push-pull power amplifier operating class B or class AB1 or AB2 (see Sec. 4.2). Under these operating conditions, the zero-signal plate current of the modulator tubes is very small and the zero modulating signal power consumed in the modulator is approximately 5 to 20 per cent of the maximum output power. The modulation transformer should ensure that the load impedance presented by the plate of the r-f amplifier is transformed to the optimum impedance for the power amplifier at the transformer primary. For modulating frequencies much lower than the carrier frequency the impedance presented by the r-f amplifier is

$$R_L = \frac{E_{bb}}{I_b} \tag{5.17}$$

where E_{bb} = plate supply voltage

I_b = d-c plate current without modulation

When the upper modulating frequencies are close enough to the carrier frequency so that the plate resonant circuit has an appreciable impedance to the modulating frequencies, Eq. (5.17) must be modified to include this impedance in series with the tube. This condition may require special consideration in the design of the modulator frequency response to compensate for this factor and special filters in the output resonant circuit to suppress the upper modulating frequencies below an acceptable value. The modulator and modulation transformer must be capable of delivering a peak power equal to approximately one-half the unmodulated plate input power to the r-f amplifier if 100 per cent sinusoidal modulation is to be employed, and the modulation transformer must be designed to handle a d-c secondary current of I_b.

5.3b. *Grid Modulation.* By varying the grid bias of a class C amplifier with a modulating voltage, it is possible to amplitude-modulate an r-f carrier. A method is illustrated in Fig. 5.9. Grid modulation is possible in a class C amplifier because the peak value of the plate current pulse, which determines the peak value of the fundamental component of plate current, varies with grid bias voltage.

FIG. 5.9. Grid modulation of a class C amplifier.

The problems associated with grid modulation of a class C amplifier are most readily visualized by observing the operating lines on the constant-current characteristics of the tube. The operating lines for zero modulation and the positive crest of 100 per cent modulation are shown in Fig. 5.10 as lines AA' and BB', respec-

tively. At the positive crest of 100 per cent modulation, the r-f plate voltage and r-f plate current component must be twice their respective values without modulation.
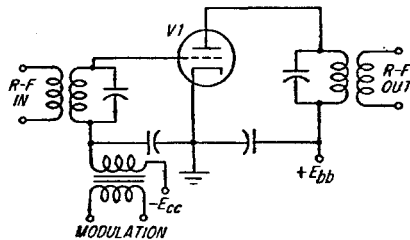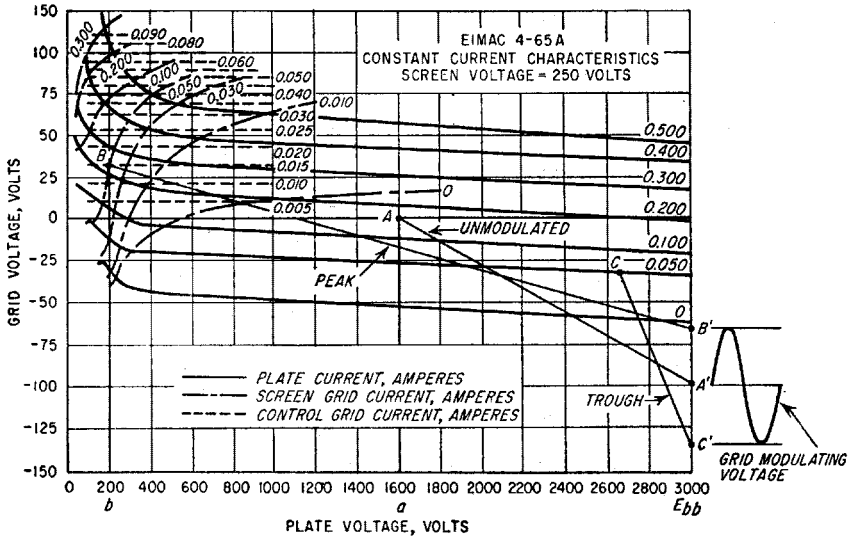
FIG. 5.10. Operating lines of a grid-modulated class C amplifier.

Since the plate supply voltage is constant, this limits the peak r-f plate voltage without modulation to less than one-half the plate supply voltage. Because of this low value of peak r-f voltage in the absence of modulation, the instantaneous value of plate voltage during the portion of the r-f cycle that plate current flows is quite high and the amplifier plate dissipation is high, resulting in low efficiency. Plate efficiencies of 35 to 45 per cent are typical when the carrier is unmodulated. At the positive peak of the modulation cycle, the peak r-f voltage is nearly equal to the plate supply voltage, and the plate efficiency is high, viz., 70 to 85 per cent.
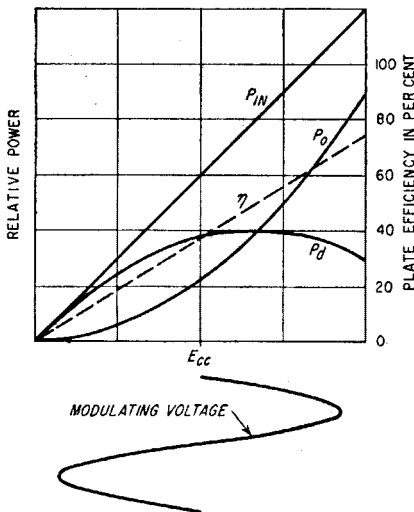
FIG. 5.11. Plate dissipation P_d of an ideal grid-modulated class C amplifier as a function of the modulation voltage.

Although the r-f output power varies as the square of the instantaneous grid modulating voltage for linear modulation, the plate input power varies approximately linearly with the instantaneous grid modulating voltage since the plate supply voltage is constant and the average plate current varies approximately linearly with grid bias. Therefore, the plate dissipation of the class C amplifier does not vary as the square of the instantaneous modulating voltage as in plate modulation. This condition is illustrated in Fig. 5.11 where the power input P_{in}, power output P_o, and plate

dissipation P_d are plotted as a function of the instantaneous grid modulation voltage for an ideal class C amplifier where the fundamental component of plate current and the average plate current vary linearly with grid bias. In practice, the linear relationship is not exact, and the instantaneous plate dissipation at the peak of the modulation cycle varies between about 0.9 and 1.25 times the plate dissipation without modulation. For the assumptions of linear variations in plate current as a function of instantaneous grid bias and 75 per cent efficiency at the positive crest of the modulation, the plate dissipation P_{do} of the amplifier without modulation is given by

$$P_{do} = P_{rfo}(\tfrac{1}{3} + \tfrac{4}{3}m) \qquad (5.18)$$

where m = modulation factor to be utilized

P_{rfo} = r-f carrier power without modulation

Equation (5.18) is useful in selecting the tube to be used when the carrier power and percentage modulation are specified.
For the same assumptions, the average plate dissipation $P_{d(avg)}$ over a complete modulation cycle is given by

$$P_{d(avg)} = P_{rfo}\left(\frac{1}{3} + \frac{4}{3}m - \frac{m^2}{2}\right) \quad (5.19)$$

A comparison of Eqs. (5.18) and (5.19) shows that the plate dissipation of a grid-modulated class C amplifier is always higher in the absence of modulation and decreases when modulation is applied. Therefore the required plate dissipation rating of the tube is specified by the zero modulation condition of operation and the modulation factor to be employed. Equations (5.18) and (5.19) are plotted in Fig. 5.12 as a function of the modulation factor m. Although the linear relationships assumed in Eqs. (5.18) and (5.19) are not exact, the equations are useful in initially determining the required plate dissipation rating of the tube to be used. The carrier power that can be obtained from a plate-modulated class C amplifier is approximately 3 times the carrier power that can be obtained from a grid-modulated class C amplifier utilizing the same tube.
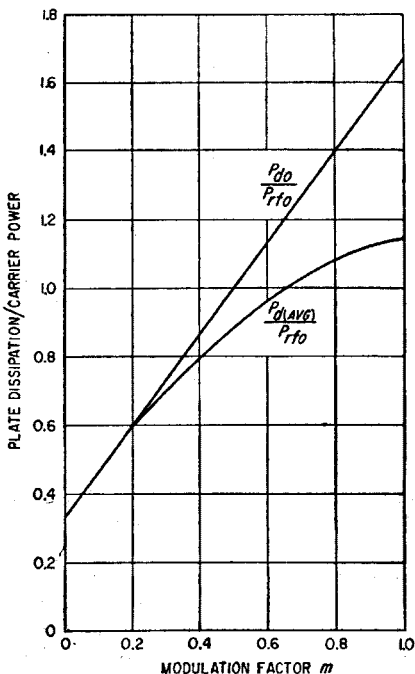
FIG. 5.12. Approximate plate dissipation of a grid-modulated class C amplifier as a function of the modulation factor.

The principal advantage of grid modulation is the requirement of low modulator voltage amplitude and power. This advantage is especially pronounced when the bandwidth of the modulation is wide enough that a plate modulation transformer cannot be used and a class-A modulator tube and impedance coupling must be utilized.

Good modulation linearity with high percentages of modulation is difficult to achieve in a grid-modulated class C amplifier. During the portion of the modulation cycle that the peak grid voltage is positive and grid current flows, the peak r-f grid voltage will decrease because of the additional loading presented to the r-f source by the grid unless the driver output is well regulated, i.e., has a low output impedance.

A reduction in r-f drive causes the r-f output from the class C amplifier to be reduced, resulting in a modulation characteristic which is nonlinear during the positive half of the modulation cycle. This is illustrated by the dotted curve OA in Fig. 5.13.

If the peak value of the r-f grid voltage is maintained constant, the class C amplifier is capable of very good linearity during the positive half of a modulation cycle for modulation factors as high as unity (OA' in Fig. 5.13). By sacrificing amplifier efficiency, the positive portion OA of the modulation characteristic can be made quite linear by operating the amplifier in the negative grid region throughout the entire cycle of the applied r-f grid voltage.

The modulation characteristic of a class C amplifier is nonlinear near the peak of the negative half of the modulation cycle because the transconductance of the tube decreases rapidly near cutoff. In this region a given change in grid bias does not
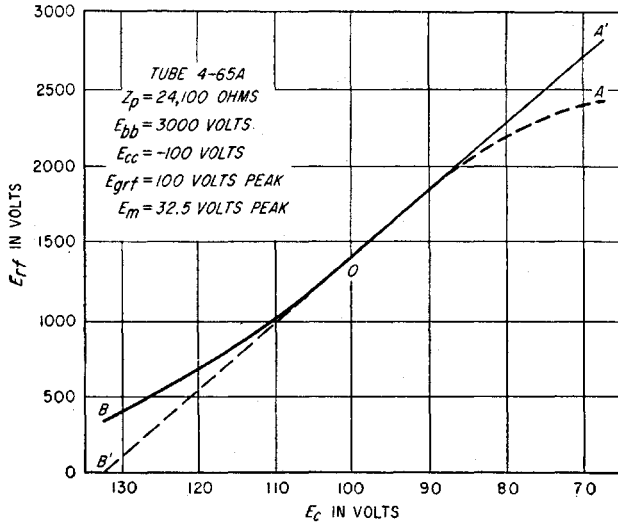
FIG. 5.13. Modulation linearity of a typical grid-modulated class C amplifier. The solid line is for the example in Fig. 5.10.

produce as large a change in the peak plate current as at less negative grid voltages, and as a result, the fundamental r-f component of plate current does not vary linearly with the grid modulation voltage (BO in Fig. 5.13). A linear modulation characteristic can be approximated by introducing suitable distortion into the modulator to cause the negative half cycle of the modulation signal applied to the grid of the class C amplifier to be amplified nonlinearly so that at the negative crest of the input modulation voltage, the modulation voltage at the grid of the class C amplifier is sufficiently negative to cut off the amplifier plate current.

The procedure to be followed in the design of a grid-modulated class C amplifier is as follows:

1. Determine the tube to be used from Eq. (5.18) for the specified power output and modulation factor.

2. Determine E_{bb}.

3. From the constant-current characteristics specify the approximate peak value of the r-f plate voltage during the positive peak of the modulation waveform (operating line to be established later). This is illustrated by the voltage difference between E_{bb} and line Bb in Fig. 5.10.

4. Determine the peak value of the r-f plate voltage without modulation

$$e_{po} = \frac{e_{p(max)}}{1 + m} \tag{5.20a}$$

where e_{po} = peak r-f plate voltage without modulation
  $e_{p(max)}$ = peak r-f plate voltage during positive peak of modulation
  m = modulation factor
  This is shown for a modulation factor of 1.0 by line Aa in Fig. 5.10.

5. With E_{bb} and e_{po} determined above, try several operating lines between these values with various values of peak r-f grid voltage and grid bias until the combination giving maximum plate efficiency with a plate dissipation equal to the maximum rating of the tube is determined. By adjusting the ratio of peak r-f grid voltage to grid bias such that the plate conduction angle is approximately 140° (for fundamental operation), the approximate operating line can be rapidly established as discussed in Sec. 4.4j. Several trials will be required to locate the optimum operating conditions.[1] The desired operating line is shown for the 4-65A tetrode having 65 watts plate dissipation by the operating line AA' in Fig. 5.10.

6. Determine the plate-load impedance for the operating line found in step 5.

$$r_b = \frac{e_{po}}{I_{H1o}}$$

where I_{H1o} = peak value of fundamental component of plate current

7. Find the operating line for the positive crest of modulation having a peak r-f plate voltage equal to $e_{p(max)}$, an r-f grid voltage as determined in step 5, and a fundamental component of plate current $I_{H1(max)}$ given by

$$I_{H1(max)} = (1 + m)I_{H1o} \tag{5.20b}$$

This operating line will be for the same load impedance as found in step 6 and will have a power output $(1 + m)^2$ times the power output without modulation. The change in grid bias required to shift to this new operating line is the peak value of the modulation voltage which must be applied to the grid to obtain the modulation factor m. This new operating line is shown for a modulation factor of 1.0 by BB' in Fig. 5.10. This is done by assuming a bias voltage, calculating $I_{H1(max)}$ from Eq. 4.45 and checking to determine whether the product $r_b I_{H1(max)}$ is equal to $e_{p(max)}$; if not, new values of bias must be assumed until $r_b I_{H1(max)}$ is equal to $e_{p(max)}$.

The minimum power output at the modulation trough is determined by finding the operating line having the impedance r_b of step 6 and the peak r-f grid voltage of step 5 at a new value of grid bias which is more negative than the bias at point A' by the difference in bias between A' and B' (line CC' in Fig. 5.10). By plotting the peak r-f plate voltage as a function of grid bias for the fixed values of r_b and r-f grid voltage, the modulation characteristic of the amplifier can be obtained. From this, the amount of distortion introduced by the grid modulation, assuming a constant driver output, can be determined and the required modulator compensation to provide linear modulation can be found.

The modulator used to amplitude-modulate the grid bias of a class C amplifier may be any one of the types of power amplifiers discussed in Secs. 4.1 and 4.2 provided that a modulation transformer having a suitable frequency response is obtainable. If the modulation frequencies are beyond the capabilities of transformers, a class A amplifier capacitively or resistively coupled to the grid circuit must be used.

[1] Tube performance can be rapidly calculated for any selected operating line by use of the tube-performance computer overlay manufactured by Eitel-McCullough, Inc., San Bruno, Calif.

The impedance Z_g presented to the modulator by the class C amplifier grid circuit is given by

$$Z_g = \frac{E_m}{I_m - I_o} \tag{5.21}$$

where E_m = peak value of modulation voltage

I_m = d-c grid current at peak of modulation voltage

I_o = d-c grid current without modulation

The currents I_m and I_o may be calculated from the operating lines for the class C amplifier for these two conditions of operation (see Sec. 4.4j). The average power P_m which the modulator must deliver to the grid circuit is

$$P_m \simeq \frac{E_m(I_m - I_o)}{2} \tag{5.22}$$

5.3c. Screen Modulation. The r-f output from a tetrode or pentode class C amplifier can be modulated by varying the screen potential. A method of accomplishing
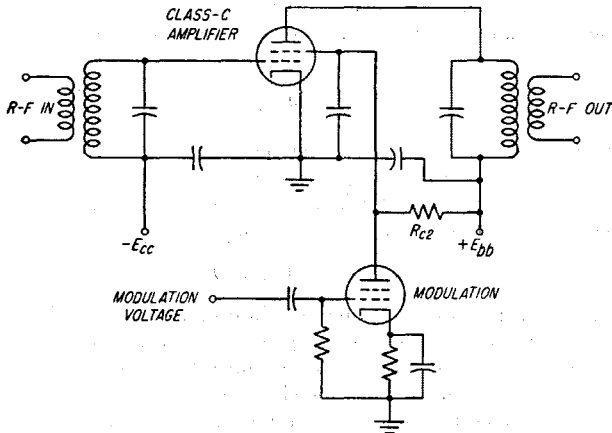
Fig. 5.14. Screen modulation of a class C r-f amplifier.

screen modulation is shown in Fig. 5.14. The plate efficiency is low when the carrier is unmodulated and rises to peak efficiency at the positive crest of modulation. The determination of the operating conditions for screen modulation is similar to that for each different value of screen potential. This can be done by assuming the currents of all tube elements to vary as the $\frac{3}{2}$ power of the tube voltages. From one constant-current characteristic for a particular value of screen current, new characteristics can be constructed by changing all electrode potentials by the same scale factor X and changing all tube currents by the factor $X^{3/2}$. For example, if characteristics are available at a screen potential of 250 volts and the characteristics at 500 volts screen potential are desired, the plate, screen, and grid voltage scales would all be multiplied by 2 and the grid, screen, and plate current values would all be multiplied by 2.82.

The modulator power requirements are higher for screen modulation than for grid modulation of the same tube, and the modulation linearity is, in general, poorer than for grid modulation. Screen modulation is, therefore, seldom used in practice except in low-power transmitters where modulation linearity is unimportant.

5.3d. Cathode Modulation. Cathode modulation of a class C amplifier is illustrated in Fig. 5.15. By placing the modulation transformer in series with the cathode the

grid-cathode and plate-cathode voltages are varied simultaneously. By adjustment of the grid tap on the modulation transformer, the relative amounts of grid and plate modulation can be varied. With the tap at ground, the grid-cathode modulation voltage equals the plate-cathode modulation voltage and nearly pure grid modulation is obtained, since the amount of modulation voltage required to produce 100 per cent grid modulation will produce only a small amount of plate modulation. As the grid tap is moved closer to the cathode, the grid-cathode modulation is reduced and the percentage of plate modulation can be increased by increasing the modulation voltage. When the tap is at the cathode, the grid is unmodulated and 100 per cent plate modulation can be obtained. Capacitor C_{c2} couples a portion of the cathode modulation voltage to the grid.

The performance obtainable is intermediate between that of pure grid modulation and pure plate modulation. The plate efficiency may vary from about 35 to 40 per cent for nearly complete grid modulation to 75 to 85 per cent for nearly complete plate modulation. In practice, a ratio of grid-to-plate modulation yielding about
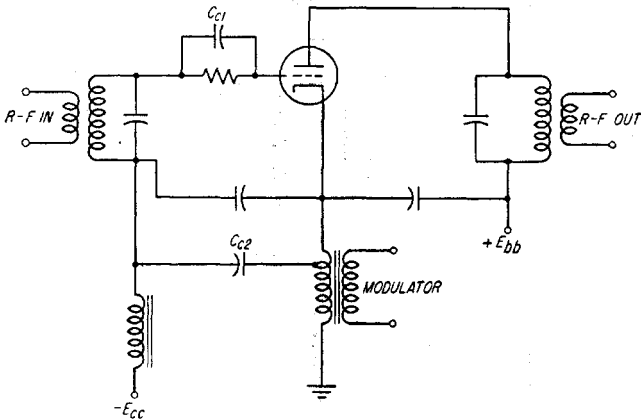
FIG. 5.15. Cathode modulation of a class C r-f amplifier.

56 per cent efficiency is normal. This will require a modulator capable of delivering a power output of approximately 20 per cent of the plate input power to the class C amplifier.

The r-f drive and grid bias requirements depend upon the ratio of grid modulation to plate modulation. Fixed bias and a low-impedance driver are desirable for best linearity when grid modulation predominates. A certain amount of grid-leak bias is desirable for best linearity when plate modulation predominates.

The graphical analysis of a cathode-modulated class C amplifier can be obtained by using the constant-current characteristics for the tube. Once the ratio of grid-cathode modulation voltage to plate-cathode modulation voltage has been established, the power output, plate efficiency, and modulation linearity can be obtained by simultaneously following the procedures for plate and grid modulation as described in Secs 5.2a and 5.2b, i.e., adjust grid voltage and plate voltage in the proper ratio to obtain operating lines at various intervals of the modulating voltage.

If the cathode modulation of the class C amplifier is linear, which is usually a good approximation, the cathode impedance of the amplifier at the modulation frequencies is given approximately by

$$Z_k \simeq K \frac{E_{bb}}{I_b}$$
(5.23)

where K = that fraction of total modulation which is developed by variation in plate-cathode voltage [i.e., for total modulation factor of m, mK is contributed by modulation of plate voltage and $(1 - K)m$ is contributed by modulation of grid voltage]

E_{bb} = plate supply voltage

I_b = average plate current

The modulation transformer should match the desired load impedance for the power amplifier to Z_k. The power P_k which the modulator must deliver to the class C amplifier is given by

$$P_k = \frac{Km^2}{2} P_i + P_m \tag{5.24}$$

where P_i = plate input power to tube

P_m = grid modulation power given by Eq. (5.22)

5.3e. Suppressor Modulation. In a pentode class C r-f amplifier, the output r-f signal can be amplitude-modulated by applying the modulating signal to the suppressor. A circuit illustrating suppressor modulation is shown in Fig. 5.16.
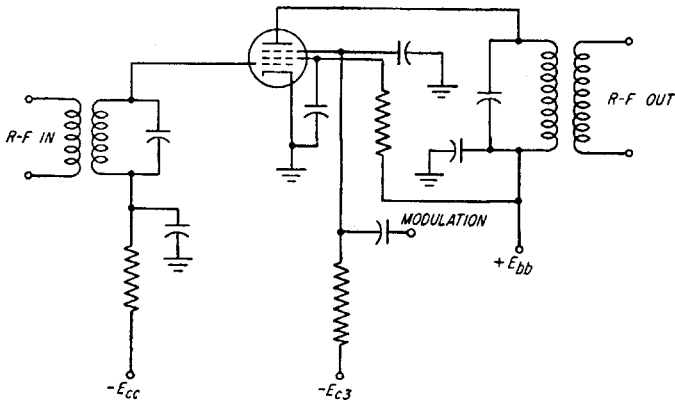
Fig. 5.16. Suppressor modulation of a pentode class C r-f amplifier.

By varying the suppressor grid voltage, the amplitude of the plate current pulses is made to follow the modulating waveform, thus producing modulation of the r-f carrier. (The amplitude of the suppressor voltage controls the ratio of plate current to screen current, and the total tube current remains essentially constant for a fixed screen potential.) Since the screen current varies inversely with plate current during suppressor modulation, screen dissipation is considerably higher with modulation and may limit the maximum allowable modulation.

The plate efficiency is essentially the same as for grid modulation, i.e., 35 to 40 per cent. The linearity of modulation is better than for a grid-modulated amplifier because the load presented to the driver by the grid circuit is independent of the modulation. This improves the rounded modulation characteristic shown by OA in Fig. 5.13, although the suppressor control characteristic introduces a certain amount of nonlinearity at both the positive and negative peaks of modulation when the percentage modulation is high. Good linearity can usually be obtained for modulation as high as 80 to 90 per cent. The use of a screen dropping resistor, with the screen bypassed for the r-f but not for the modulation frequencies, will minimize the problem of screen dissipation by reducing the screen voltage when the screen current increases.

5.3f. The Van der Bijl Modulator. The Van der Bijl modulator (Fig. 5.17) consists of a class A r-f amplifier having a small amplitude r-f carrier applied to the grid

in series with a relatively large amplitude modulating voltage. Modulation of the output r-f signal is accomplished by variation in the amplification of the input r-f signal by the changes in grid bias.
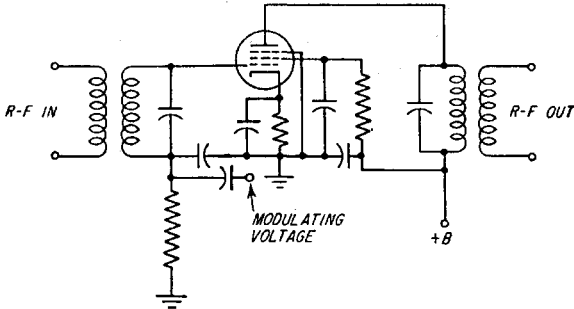
FIG. 5.17. Typical circuit of a Van der Bijl modulator.

In a pentode amplifier having a plate resistance much higher than the a-c plate-load resistance r_b, the peak r-f output voltage E_o is given by

$$E_o = g_m r_b E_i \qquad (5.25)$$

where E_i = peak r-f input voltage
g_m = grid-plate transconductance
In the region of operation where g_m varies linearly with changes in grid bias, the modulation will be linear. A linear varia-
tion in g_m with grid bias implies a variation in plate current as the square of the varia-
tion in grid bias. In most pentodes, this region of linear variation in g_m is large enough to allow approximately 50 per cent linear modulation. The modulation factor obtained when g_m is varied linearly be-
tween two limits is given by

$$m = \frac{g_{m2} - g_{m1}}{g_{m2} + g_{m1}} \qquad (5.26)$$

where g_{m2} = maximum value of trans-conductance
g_{m1} = minimum value of trans-conductance
The efficiency of a Van der Bijl modulator is quite low, but the r-f excitation and modulation power required is negligible and the modulator is easily adjusted.

5.3g. Diode Modulation. Diodes may be used as modulators in the same manner that they are employed as mixers and detectors. Consider the diode modulator of Fig. 5.18. Assume that the input sig-
nal consists of $E_c \cos \omega_c t$ and $E_m \cos \omega_m t$. If the carrier voltage E_c is much larger than the modulation voltage E_m, which is usually the case, the conductance of a perfect
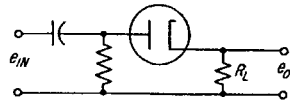
(a) MODULATOR CIRCUIT
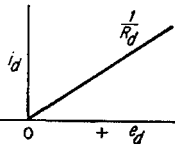
(b) DIODE CHARACTERISTIC
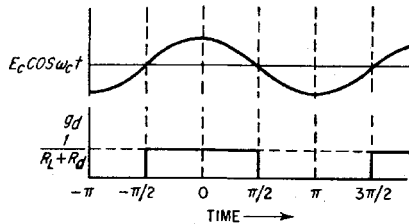
(c) DIODE AND LOAD CONDUCTANCE AS A FUNCTION OF CARRIER VOLTAGE

FIG. 5.18. Diode modulator.

diode and its load is given by

$$g_d = \frac{1}{R_d + R_L} \quad \text{for } E_c > 0$$
$$g_d = 0 \quad \text{for } E_c < 0$$

The diode conductance is shown as a function of the applied voltage in Fig. 5.18c. It may be represented by a Fourier series as

$$g_d = g_o + g_1 \cos \omega_c t + g_2 \cos 2\omega_c t + g_3 \cos 3\omega_c t + \cdots + g_n \cos n\omega_c t \quad (5.27)$$

where

$$g_n = \frac{1}{\pi} \int_0^{2\pi} g_d \cos n\omega_c t \, d\omega_c t \quad (5.28)$$

$$g_o = \frac{1}{2\pi} \int_0^{2\pi} g_d \, d\omega_c t \quad (5.29)$$

The current in the load resistance R_L is given by

$$I_o = (E_c \cos \omega_c t + E_m \cos \omega_m t) g_d \quad (5.30)$$

The modulation sideband components of load current for the carrier are given by

$$I_{om} = E_m g_1 \cos \omega_m t \cos \omega_c t$$
$$= \frac{2E_m}{\pi(R_d + R_L)} \cos \omega_m t \cos \omega_c t$$
$$= \frac{E_m}{\pi(R_d + R_L)} [\cos (\omega_c - \omega_m)t + \cos (\omega_c + \omega_m)t] \quad (5.31)$$

The carrier component of load current is given by

$$I_{oc} = E_c g_o \cos \omega_c t$$
$$= \frac{E_c}{2(R_L + R_d)} \cos \omega_c t \quad (5.32)$$

Since the modulation factor is unity when one sideband current equals one-half I_{oc}, the modulation factor is given by

$$m = \frac{4}{\pi} \frac{E_m}{E_c} \quad (5.33)$$

The modulation will be linear if the percentage modulation is low. The harmonics of the carrier and the sidebands about these harmonics which are created by the diode nonlinearity must be filtered out by the load circuit if only the carrier and its sidebands are desired. Although the efficiency is low, the diode modulator finds frequent application in low-level modulators because of its simplicity.

5.4. Suppressed Carrier Modulation and Single-sideband Generation. The information transmitted by a modulated carrier is contained wholly in the modulation sidebands. The transmission of the carrier in no way aides the transmission of the desired intelligence. Since the power contained in the carrier is twice that in the sidebands even with 100 per cent modulation, the transmission of only the sideband signals may represent a considerable saving in transmitted power if the carrier is eliminated at a low level and the sidebands only are linearly amplified to the desired output level. The information contained in the sidebands can be completely recovered at the receiver by adding a sinusoidal signal at the carrier frequency to the sideband signals before detection. When all of the sidebands of a modulated carrier are transmitted but the carrier itself is eliminated, the transmission is referred to as double-sideband suppressed carrier modulation (dsb).

Each frequency component of the modulating signal at the transmitter produces two sideband frequency components spaced equally on each side of the carrier fre-

quency by the modulating frequency (see Sec. 5.2a). These sideband pairs are redundant since they each contain the same information. If either sideband and the carrier are eliminated, the original information contained in the modulating signal can still be recovered at the receiver by the use of a local oscillator at the carrier frequency, although some amplitude and phase distortion will be present. This type of signal transmission is known as single-sideband suppressed carrier modulation (ssb).

By considering the sideband signals as vectors rotating about the carrier vector at their difference frequencies (see Sec. 5.2b), the manner in which the modulating signal is recovered from a suppressed carrier transmission is apparent from Fig. 5.19. For both single-sideband and double-sideband modulation, the output from an amplitude or envelope detector is equal to the absolute magnitude of the vector sum of the reference carrier and the sideband signals. If the reference carrier is added to the sidebands of dsb modulation in exactly the same phase as that of the original carrier, no distortion of the envelope occurs during amplitude detection provided that the amplitude of the referenced carrier is large enough so that the modulation factor does not exceed 1.0. If the phase of the reference carrier differs appreciably from that of the original carrier considerable distortion is introduced by the resultant phase

(a) SINGLE SIDEBAND MODULATION   (b) DOUBLE SIDEBAND MODULATION
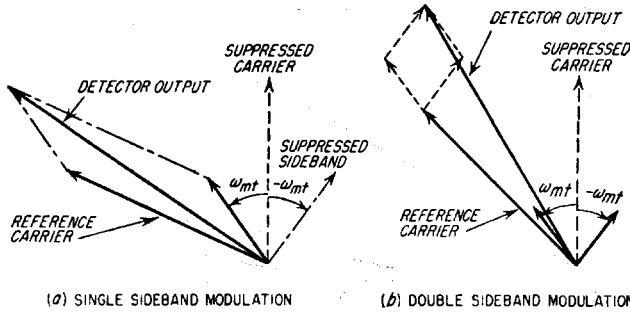
Fig. 5.19. Vector relationships in carrier-suppressed modulation sytems.

modulation of the envelope vector. If the reference carrier phase differs from that of the original carrier by 90°, the amplitude modulation of the envelope vector at the modulation frequency is completely eliminated. The result is phase modulation and 2nd harmonic amplitude modulation. For these reasons dsb is not utilized unless the carrier is transmitted at a reduced level (usually about -20 db) for synchronization of the reference carrier.

When single-sideband modulation is employed, the phase of the reference carrier is unimportant. The envelope distortion, which arises from the phase modulation of the amplitude vector, is independent of the phase of the reference carrier relative to the suppressed carrier. The distortion is minimized by making the reference carrier very large compared to the modulation sidebands. The harmonic distortion present in the envelope of single-sideband modulation is shown in Fig. 5.20 as a function of the effective modulation factor at the detector input, i.e., the ratio of the peak modulation sideband amplitude to the peak reference carrier amplitude. A difference between the frequency of the reference carrier and the frequency of the suppressed carrier will have the effect of shifting all of the modulating frequencies by a constant amount in single-sideband modulation. However, in double-sideband modulation, a frequency difference between the reference carrier and the suppressed carrier causes a variation in the amplitude and phase distortion of the envelope at twice the difference frequency. The reference carrier can be phase-locked to the transmitter carrier by transmitting a low-level carrier together with the modulation sidebands rather than suppressing the carrier entirely. Single-sideband modulation

possesses the advantages over double-sideband modulation of requiring only half the bandwidth and of having a detected output which is less affected by phase and frequency variation of the reference carrier.

A frequency error in the reference carrier of 10 to 20 cps is acceptable for speech transmission; however, an error of 50 cps seriously degrades intelligibility. The transmission of music places a tighter tolerance on the reference carrier because the error destroys the harmonic relationships of overtones. Therefore, reference oscillator stabilities of from 1 part in 10^6 to 1 part in 10^7 are required at a 30 Mc carrier frequency. For this reason, it is common commercial practice to transmit a pilot carrier at a reduced level (usually -20 db from the sideband level) which is filtered out in the receiver and used to control the frequency of the reference oscillator.

Assuming that the receiver bandwidth is reduced to one-half the desired a-m value for ssb reception, the use of ssb provides an increased signal-to-noise ratio at the receiver second detector of 4.8 db for the same total transmitted power. An equally important advantage of ssb over a-m in long-range communication is the minimization
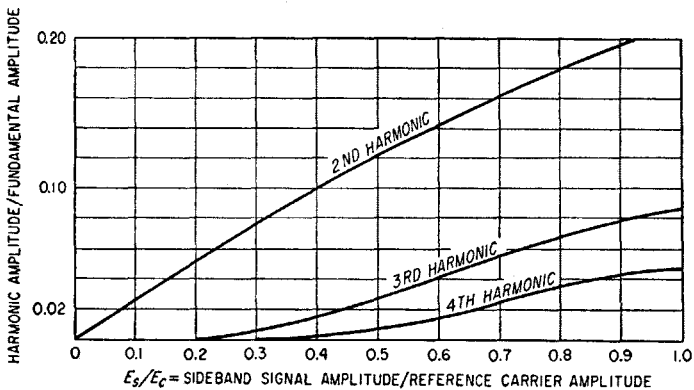
FIG. 5.20. Harmonic content of the detected envelope of a single-sideband signal and reference carrier.

of the effects of selective fading of the carrier which renders a-m unintelligible when the carrier is reduced significantly below twice the sideband amplitude by atmospheric effects.

The economic advantage of suppressed carrier modulation is a function of the ratio of the number of receivers to the number of transmitters in the system in which it is to be used. When there are only a few receivers for each transmitter, it is more economical to complicate the receivers and minimize the transmitter power requirements. However, when the number of receivers is very large compared to the number of transmitters, as is the case in broadcast stations, the economics of the situation are reversed.

5.4a. Balanced Modulators. The suppression of the carrier in either single-sideband or double-sideband modulation is accomplished by the use of a balanced modulator (Fig. 5.21). It consists of two modulators connected back-to-back in which the modulation is injected with reversed phase in one modulator compared to the other. The carrier is injected into both modulators in the same phase. The outputs from the two modulators are given by

$$e_1 = E_c \cos \omega_c t + \frac{E_m}{2} \cos (\omega_c + \omega_m)t + \frac{E_m}{2} \cos (\omega_c - \omega_m)t \tag{5.34}$$

$$e_2 = E_c \cos \omega_c t + \frac{E_m}{2} \cos [(\omega_c + \omega_m)t + \pi] + \frac{E_m}{2} \cos [(\omega_c - \omega_m)t - \pi] \tag{5.35}$$

The push-pull output from the two modulators is given by

$$e_o = e_1 - e_2 = E_m \cos (\omega_c + \omega_m)t + E_m \cos (\omega_c - \omega_m)t \qquad (5.36)$$

If the modulators are perfectly balanced, i.e., have equal carrier outputs, the carrier is completely rejected, and only the modulation sidebands remain. The nonlinear elements in Fig. 5.21 are usually diodes.

5.4b. Single-sideband Generation. There are two methods of eliminating the undesired sideband of suppressed carrier modulation to achieve single-sideband modulation. The most frequently used method is the use of selective filters to pass only the upper or lower sideband of carrier-suppressed modulation. This is illustrated in Fig. 5.22a. In many applications, however, the ratio of the lowest modulation frequency to the carrier frequency is so small that one group of sidebands cannot be separated from the other by a filter at the sideband frequencies, e.g., practical filters could not separate sidebands separated by 100 cycles at a carrier frequency of 1 Mc. When this situation prevails, the desired sideband rejection can be accomplished by first modulating a low-frequency carrier with the desired modulating signals in a balanced modulator, filtering out the unwanted sidebands where the ratio
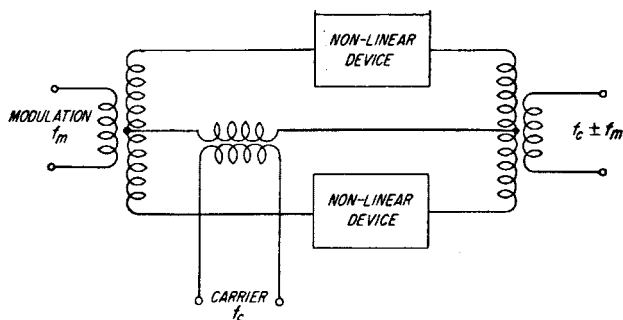
FIG. 5.21. Basic circuit of a balanced modulator.

of modulating frequency to carrier frequency is large enough to permit successful filtering and then using the desired sidebands to modulate another higher-frequency carrier. This process can be applied as many times as necessary to achieve the desired frequency band for the sidebands to be transmitted. The reference carrier at the receiver is set to the frequency which provides the same demodulated frequency components as made up the original modulation. This system of successive modulation and filtering is illustrated in Fig. 5.22b. Modulation systems of this type are used in telephone systems to provide multiple single-sideband voice channels closely spaced for transmission over a single coaxial cable.

The most satisfactory means of obtaining the desired filtering is with the Collins[1] mechanical filters. These filters provide extremely sharp selectivity, e.g., for a center frequency of 250 kc and a −3 db bandwidth of 3.2 kc, the skirt selectivity is 40 to 60 db attenuation 600 cps away from the −3 db frequency.

The second method of developing single-sideband modulation accomplishes the rejection of the unwanted sideband by phase discrimination (see Fig. 5.23). The carrier and modulation into one balanced modulator are both shifted in phase by 90° with respect to the carrier and modulation into the other balanced modulator. The outputs of the two modulators are given by Eqs. (5.37) and (5.38) when the 90° phase shift is leading.

[1] The Collins Radio Co., Cedar Rapids, Iowa.

$$e_1 = \frac{E_m}{2} \cos \left[\left(\omega_c t + \frac{\pi}{2} \right) + \left(\omega_m t + \frac{\pi}{2} \right) \right] + \frac{E_m}{2} \cos \left[\left(\omega_c t + \frac{\pi}{2} \right) - \left(\omega_m t + \frac{\pi}{2} \right) \right]$$
$$(5.37)$$

$$e_2 = \frac{E_m}{2} \cos (\omega_c + \omega_m)t + \frac{E_m}{2} \cos (\omega_c - \omega_m)t \qquad (5.38)$$

The output from the modulator is given by

$$e_o = e_1 + e_2$$
$$= E_m \cos (\omega_c - \omega_m)t \qquad (5.39)$$

Thus, the upper sidebands are eliminated by a +90° phase shift. Alternatively, the lower sidebands may be eliminated by a −90° shift. All modulating frequencies
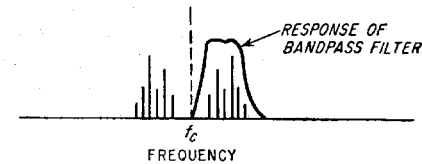
(*a*) USE OF A FREQUENCY SELECTIVE FILTER TO
OBTAIN SINGLE SIDEBAND MODULATION
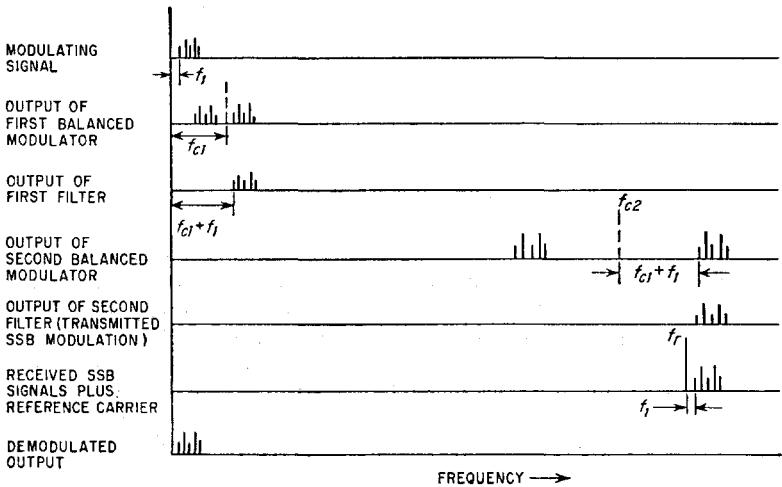
(*b*) USE OF SUCCESSIVE MODULATION AND FILTERING
TO ACHIEVE SINGLE SIDEBAND MODULATION

Fig. 5.22. Sideband filtering to achieve single-sideband modulation.

must also be shifted either +90 or −90 degrees in conjunction with the carrier phase shift. This requires the use of an all-pass filter network (see Sec. 17.3) having 90° phase shift for all frequencies of interest. If some phase distortion of the modulation can be tolerated, the modulation need not be shifted exactly 90°. Instead, the modulation can be passed through separate networks for each modulator so that the phase difference in the modulation to the two balanced modulators is 90°.

The practical limitation in this type of ssb modulator is the accuracy of phase shift

and amplitude control required to achieve the necessary level of rejection of the undesired sideband. An amplitude unbalance of 1 db or a phase error of 1 degree will reduce the rejection of the undesired signal to 40 db. To minimize adjacent channel interference, the undesired sideband signals should be reduced by at least 40 db.

The generation of single sideband modulation at low signal levels by either of the methods discussed requires the use of linear class B r-f amplifiers to obtain the desired output power. Although the spectrum of an ssb signal is identical to the modulating signal spectrum, each frequency component is translated to a much higher frequency, and, as a result, the instantaneous amplitude of the r-f envelope bears no relationship to the modulating signal amplitude as in a-m. Clipping of the modulation peaks or nonlinear amplification results in the generation of many sum and difference frequencies which can extend far outside the alloted frequency channel and cause an unintelligible splatter type of interference in adjacent channels. A 40-db ratio of desired signal to adjacent channel distortion signal is desirable. Negative r-f feedback in the high-level class B r-f amplifier stages is employed to achieve the necessary linearity. Since peak clipping of the modulating signal before modulation does not provide control of the peak r-f amplitude, a system of automatic gain reduction in
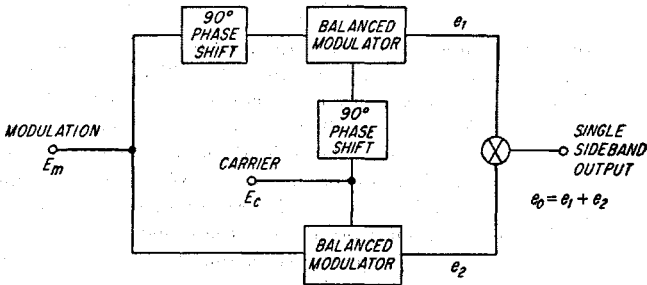
Fig. 5.23. Block diagram of a phase-discrimination single-sideband modulator.

the driver stages is sometimes employed to prevent peak clipping in the final r-f amplifier.

5.5. Angle Modulation. Information can be transmitted on a carrier by varying any of the parameters of the sinusoid in accordance with the modulating voltage. Thus, a carrier is described by

$$e(t) = E_c \sin \theta \tag{5.40}$$

where $\theta = \omega_c t + \phi$

This carrier can be made to convey information by modulating the peak amplitude E_c or by varying the instantaneous phase angle θ of the carrier. This type of modulation is known as *angle modulation*. The two types of angle modulation which have practical application are *phase modulation* and *frequency modulation*.

In phase modulation, the instantaneous phase angle θ of the carrier is varied by the amplitude of the modulating signal. As discussed in Sec. 5.6b, the principal application of phase modulation is in the utilization of modified phase modulators in systems which transmit frequency modulation. The expression for a carrier phase-modulated by a single sinusoid is given by

$$e(t) = E_c \sin (\omega_c t + \phi + \Delta\phi \cos \omega_m t) \tag{5.41}$$

where $\Delta\phi$ = peak value of phase variation introduced by modulation and is defined as phase deviation

ω_m = modulation frequency, radians/sec

In frequency modulation the instantaneous frequency of the carrier, i.e., the time derivative of the phase angle θ, is made to vary in accordance with the amplitude of the modulating signal. Thus,

$$f = \frac{1}{2\pi} \frac{d\theta}{dt} \tag{5.42}$$

When the carrier is frequency-modulated by a single sinusoid,

$$f = f_c + \Delta f \cos \omega_m t \tag{5.43}$$

where Δf = peak frequency deviation introduced by modulation
The instantaneous total phase angle θ is given by

$$\theta = 2\pi \int f \, dt + \theta_o \tag{5.44}$$

$$\theta = 2\pi f_c t + \frac{\Delta f}{f_m} \sin 2\pi f_m t + \theta_o \tag{5.45}$$

The complete expression for a carrier which is frequency-modulated by a single sinusoid is

$$e(t) = E_c \sin \left(\omega_c t + \frac{\Delta f}{f_m} \sin 2\pi f_m t + \theta_o \right) \tag{5.46}$$

The maximum frequency difference between the modulated carrier and the unmodulated carrier is known as the *frequency deviation* Δf. The ratio of the frequency deviation Δf to the modulation frequency f_m is known as the *modulation index*. It is also referred to as the *deviation ratio*. The degree of modulation in a frequency modulation system is usually defined as the ratio of the frequency deviation Δf to the maximum frequency deviation allowable or the ratio of Δf to the maximum frequency deviation of which the system is capable. Degree of modulation in a frequency modulation system is, therefore, not a property of the signal itself.

5.5a. Modulation Spectra. The frequency spectrum of a p-m or f-m carrier can be determined by an expansion of the basic equation for each waveform. Consider first an f-m carrier.

$$e(t) = E_c \sin \left(\omega_c t + \frac{\Delta f}{f_m} \sin \omega_m t \right)$$

$$= E_c \left[\sin \omega_c t \cos \left(\frac{\Delta f}{f_m} \sin \omega_m t \right) + \cos \omega_c t \sin \left(\frac{\Delta f}{f_m} \sin \omega_m t \right) \right] \tag{5.47}$$

The term $\cos \left(\frac{\Delta f}{f_m} \sin \omega_m t \right)$ is equal to the following Bessel function series:

$$\cos \left(\frac{\Delta f}{f_m} \sin \omega_m t \right) = J_o \left(\frac{\Delta f}{f_m} \right) + 2 \left[J_2 \left(\frac{\Delta f}{f_m} \right) \cos 2\omega_m t + J_4 \left(\frac{\Delta f}{f_m} \right) \cos 4\omega_m t + \cdots \right] \tag{5.48}$$

The term $\sin \left(\frac{\Delta f}{f_m} \sin \omega_m t \right)$ is equal to the following Bessel function series:

$$\sin \left(\frac{\Delta f}{f_m} \sin \omega_m t \right) = 2 \left[J_1 \left(\frac{\Delta f}{f_m} \right) \sin \omega_m t + J_3 \left(\frac{\Delta f}{f_m} \right) \sin 3\omega_m t + \cdots \right] \tag{5.49}$$

where $J_o \left(\frac{\Delta f}{f_m} \right)$, $J_1 \left(\frac{\Delta f}{f_m} \right)$, $J_2 \left(\frac{\Delta f}{f_m} \right)$, etc., are the Bessel functions of first kind and zero order, first order, second order, etc., of argument $\Delta f / f_m$.

By substituting the Bessel function series into Eq. (5.47) and making the trigonometric substitution

$$2 \sin A \cos B = \sin (A + B) + \sin (A - B)$$

Equation (5.47) becomes

$$
\begin{aligned}
e(t) = E_c \Big\{ & J_o \left(\frac{\Delta f}{f_m} \right) \sin \omega_c t \\
& + J_1 \left(\frac{\Delta f}{f_m} \right) \sin (\omega_c + \omega_m)t - J_1 \left(\frac{\Delta f}{f_m} \right) \sin (\omega_c - \omega_m)t \\
& + J_2 \left(\frac{\Delta f}{f_m} \right) \sin (\omega_c + 2\omega_m)t + J_2 \left(\frac{\Delta f}{f_m} \right) \sin (\omega_c - 2\omega_m)t \\
& + J_3 \left(\frac{\Delta f}{f_m} \right) \sin (\omega_c + 3\omega_m)t - J_3 \left(\frac{\Delta f}{f_m} \right) \sin (\omega_c - 3\omega_m)t \\
& + \cdots \Big\}
\end{aligned}
\tag{5.50}
$$

The frequency components of a frequency-modulated wave are given by Eq. (5.50). Frequency modulation of a carrier by a single sinusoid results in the creation of an
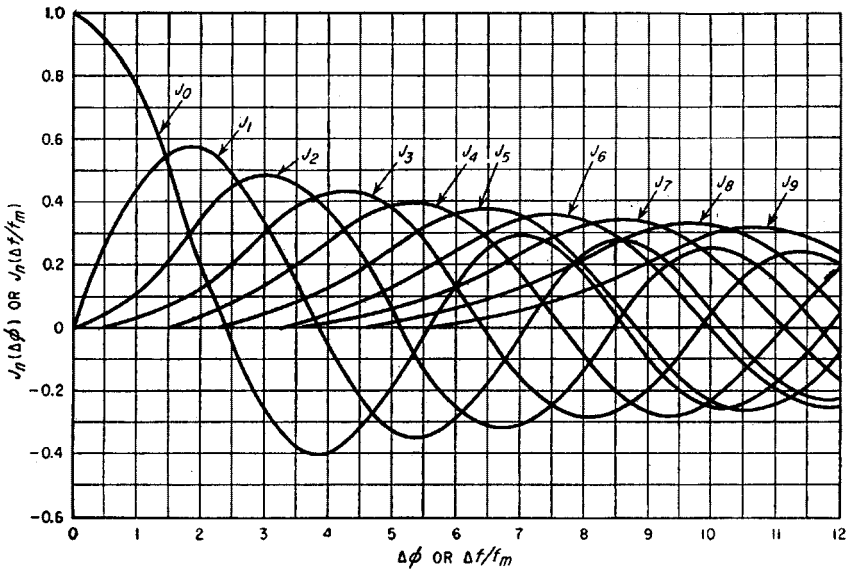
FIG. 5.24. Bessel coefficients for the carrier and first nine sidebands of a phase- or frequency-modulated wave.

infinite number of sideband signals spaced by an amount ω_m from each other and having amplitudes determined by the Bessell coefficients in Eq. (5.50). The Bessel coefficients are functions of $\Delta f/f_m$ so that as the amplitude of the modulating sinusoid varies, causing Δf to vary, the amplitudes of the various sideband signals are changed. The amplitude of the carrier is reduced by the frequency modulation in accordance with the Bessel coefficient $J_o(\Delta f/f_m)$. The Bessel coefficients for the carrier and the first nine sidebands are given in Fig. 5.24.

Frequency modulation removes power from the carrier and puts it in the sidebands. Since the amplitude of the envelope of an f-m carrier is unchanged by the modulation, the total power in the carrier and the sidebands is equal to the unmodulated carrier power. This is also shown by the Bessel function relationship of Eq. (5.51).

$$
\left[J_o \left(\frac{\Delta f}{f_m} \right) \right]^2 + 2 \left[J_1 \left(\frac{\Delta f}{f_m} \right) \right]^2 + 2 \left[J_2 \left(\frac{\Delta f}{f_m} \right) \right]^2 + 2 \left[J_3 \left(\frac{\Delta f}{f_m} \right) \right]^2 + \cdots = 1 \tag{5.51}
$$

The spectrum of an f-m carrier is shown in Fig. 5.25 for various indexes of modulation. The band of frequencies ΔF encompassing all the sideband signals of significant amplitude is given approximately by

$$\Delta F = 2(\Delta f + f_m) \tag{5.52}$$

When the modulating signal is composed of more than one sinusoid, the determination of the spectrum of the modulated carrier becomes very difficult because the modulation of the carrier by one modulating sinusoid is not independent of the modulation of the carrier by the other modulating frequencies as is the case in amplitude

(a) CONSTANT MODULATING FREQUENCY   (b) CONSTANT FREQUENCY DEVIATION
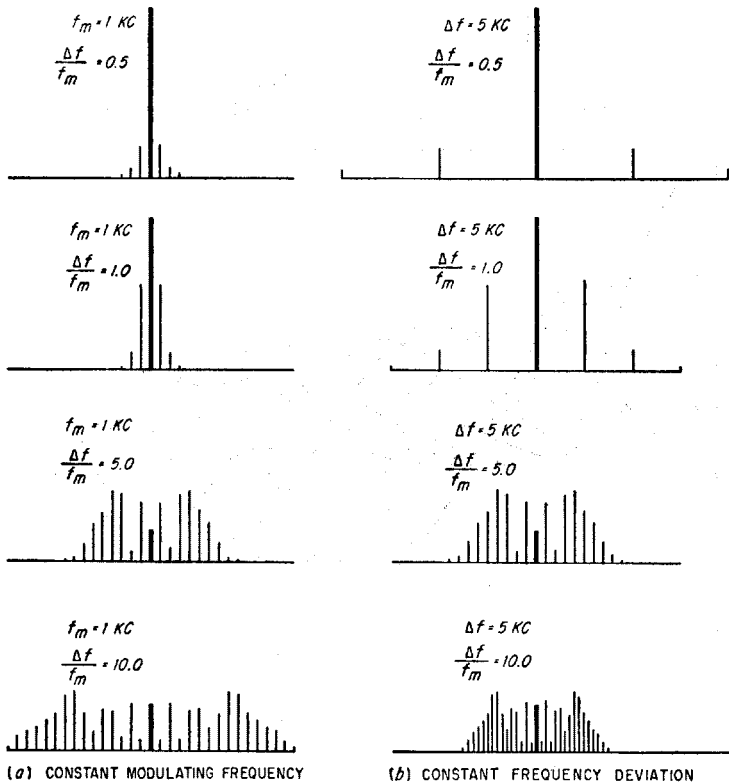
FIG. 5.25. The spectra of a frequency-modulated carrier for various combinations of frequency deviation and modulating frequency.

modulation. Sidebands are generated at frequencies corresponding to the sum and difference of each modulating frequency and its harmonics with all other modulating frequencies and their harmonics. Thus, for a modulating signal of $E_1 \cos \omega_1 t + E_2 \cos \omega_2 t + E_3 \cos \omega_3 t$, the f-m carrier will have sidebands at

$$\omega_c \pm m\omega_1 \pm n\omega_2 \pm p\omega_3$$

where m, n, and p have all possible integer values.

When a complex signal is used to frequency-modulate a carrier, the resulting sideband distribution is not necessarily symmetrical about the carrier.[1] Although fre-

[1] H. S. Black, "Modulation Theory," p. 186, D. Van Nostrand Company, Inc., New York, 1953.

quency modulation of a carrier by a complex waveform results in many additional sideband frequencies, it does not appreciably widen the frequency band occupied by the predominate sidebands. The required transmission bandwidth for an f-m system is given by Eq. (5.52) where Δf is the maximum frequency deviation of the carrier and f_m is the highest modulation frequency to be transmitted.

The spectrum of a p-m carrier can be obtained by expansion of Eq. (5.41) in a manner exactly analogous to what was done in Eqs. (5.47) and (5.50) for an f-m
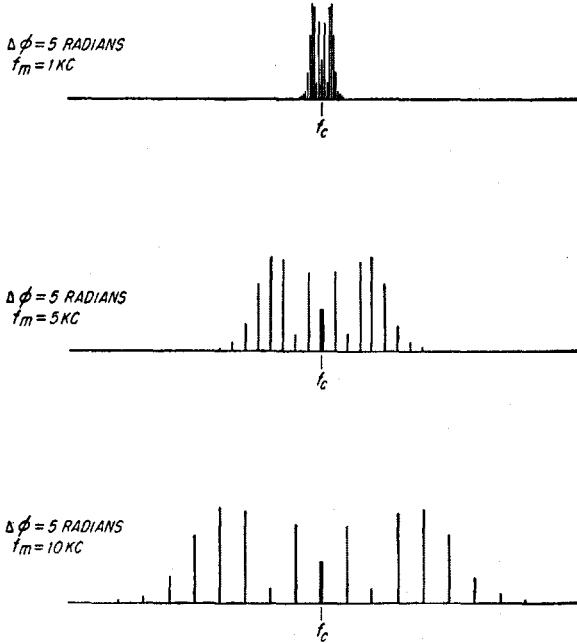
FIG. 5.26. The spectra of a phase-modulated carrier for a fixed phase deviation and variable modulating frequency.

carrier. The resulting equation expressing the instantaneous phase-modulated signal in terms of a carrier plus sidebands is given by Eq. (5.53).

$$
\begin{aligned}
e(t) \;=\; E_c\{ & J_o(\Delta\phi)\,\sin\,(\omega_c t + \phi) \\
& + J_1(\Delta\phi)\,\cos\,[(\omega_c + \omega_m)t + \phi] + J_1(\Delta\phi)\,\cos\,[(\omega_c - \omega_m)t + \phi] \\
& - J_2(\Delta\phi)\,\sin\,[(\omega_c + 2\omega_m)t + \phi] - J_2(\Delta\phi)\,\sin\,[(\omega_c - 2\omega_m)t + \phi] \\
& - J_3(\Delta\phi)\,\cos\,[(\omega_c + 3\omega_m)t + \phi] - J_3(\Delta\phi)\,\cos\,[(\omega_c - 3\omega_m)t + \phi] \\
& + J_4(\Delta\phi)\,\sin\,[(\omega_c + 4\omega_m)t + \phi] + J_4(\Delta\phi)\,\sin\,[(\omega_c - 4\omega_m)t + \phi] \\
& + \cdots \}
\end{aligned}
\tag{5.53}
$$

Equation (5.53) is identical with Eq. (5.50) except that the argument of the Bessel coefficients is now $\Delta\phi$ instead of $\Delta f/f_m$ and the sideband signals have different phase relationships. The spectrum of Eq. (5.53) is identical with that of an f-m carrier where

$$
\Delta\phi = \frac{\Delta f}{f_m}
\tag{5.54}
$$

However, in phase modulation $\Delta\phi$ is a function of the amplitude of the modulating signal only. For a constant amplitude of modulating signal the frequency deviation Δf is directly proportional to the modulating frequency f_m. This effect is shown in Fig. 5.26 where the spectra of a p-m carrier are shown for a fixed phase deviation $\Delta\phi$

and several different modulating frequencies f_m. The frequency band occupied by those sidebands containing significant energy is given approximately by

$$\Delta F = 2f_m(1 + \Delta\phi) \qquad (5.55)$$

where $\Delta\phi$ = phase deviation, radians

The required transmission band for a p-m carrier is given in Eq. (5.55) where f_m is the highest modulating frequency and $\Delta\phi$ is the peak phase deviation to be employed.

(a) NETWORK FOR THE PRODUCTION OF PHASE MODULATION IN A FREQUENCY MODULATOR

(b) NETWORK FOR THE PRODUCTION OF FREQUENCY MODULATION IN A PHASE MODULATOR
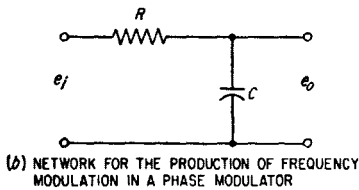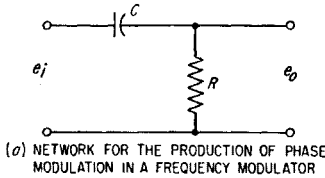
FIG. 5.27. Frequency and phase modulation conversion networks.

The essential difference between phase modulation and frequency modulation is that, assuming a constant amplitude modulating signal, the frequency deviation of a p-m carrier is proportional to the modulating frequency, while the frequency deviation of an f-m carrier is constant with a changing modulating frequency.

Phase modulation can be created in an f-m system by passing the modulating signals through a network which causes the amplitude of the modulating signals to increase linearly with frequency. This may be accomplished by using the RC network shown in Fig. 5.27a. The gain of the network is

$$\frac{e_o}{e_1} = \frac{j\omega RC}{1 + j\omega RC} \qquad (5.56)$$

At frequencies where ωRC is much smaller than 1, the network gain is a linear function of frequency.

Conversely, frequency modulation can be created in a p-m system by passing the modulating signals through a network which causes the amplitude of the modulating signals to vary inversely with frequency. This may be accomplished by using the RC network shown in Fig. 5.27b. The gain of the network is

$$\frac{e_o}{e_1} = \frac{1}{1 + j\omega RC} \qquad (5.57)$$

At frequencies where ωRC is much larger than 1, the gain of the network is an inverse function of frequency.

5.5b. Modulation Distortion. Distortion in f-m and p-m systems can be caused by numerous factors. Elimination of significant sidebands due to restricted receiver bandwidth will cause nonlinear distortion; i.e., a single modulating frequency at the transmitter will not be reproduced as a perfect sinusoid at the receiver detector, thus introducing harmonics of the modulating signal at the output.

5.5c. Frequency Multiplication of an F-M Carrier. If an f-m carrier as described by Eq. (5.46) is passed through harmonic amplifiers to multiply the carrier frequency, the resultant signal is given by

$$e(t) = E_c \sin\left(n\omega_c t + \frac{n\,\Delta f}{f_m}\sin\,\omega_m t + n\theta_o\right) \qquad (5.58)$$

where n = integer value by which carrier is multiplied

The effect of such multiplication is to increase the modulation index $\Delta f/f_m$ by the factor n.

5.6. Methods of Angle Modulation. Since either frequency modulation or phase modulation can be achieved in a given type of modulator by using the appropriate correction network as discussed in Sec. 5.5a, any modulator which inherently produces either frequency modulation or phase modulation can be used to produce both types of angle modulation.

5.6a. Frequency Modulators. In frequency modulators, the frequency of an oscillator is made to vary in accordance with the amplitude of the modulating signal. The most common method involves the use of a reactance tube.

The Reactance-tube Modulator. The reactance-tube modulator is discussed in detail in Sec. 7.6e. It consists of a pentode tube connected across the resonant circuit of an oscillator as in Fig. 5.28. By means of an *RC* phase shift network the r-f voltage between the grid and cathode of the reactance tube is caused to be very nearly 90° out of phase with the r-f voltage across the oscillator resonant circuit. The plate current of the reactance tube is in quadrature with the r-f voltage appearing between plate and cathode, and the reactance tube appears as either an inductive or a capacitive reactance (inductive in Fig. 5.28) across the oscillator resonant circuit. By varying the grid bias of the reactance tube with the modulating voltage, the

Fig. 5.28. Reactance tube connected to a Hartley oscillator.

transconductance and hence the reactance of the tube is varied, changing the frequency of the oscillator in accordance with the modulation. For the circuit of Fig. 5.28, the frequency of oscillation ω_o is given approximately by

$$\omega_o \simeq \frac{1}{\sqrt{C_o L_o \left(1 - \frac{g_m L_o}{RC}\right)}} \tag{5.59}$$

where C_o = total capacitance in parallel with inductance L_o
L_o = inductance of oscillator resonant circuit
g_m = transconductance of reactance tube
R = resistance between plate and grid of reactance tube
C = capacitance between grid and cathode of reactance tube

The ratio of the frequency deviation to the carrier frequency is usually limited to rather small values in order to maintain a linear relationship between the frequency of oscillation ω_o and the reactance tube grid bias voltage. Since the achievable frequency deviation for a given amount of distortion is directly proportional to the oscillator frequency, the frequency deviation at any desired carrier frequency cannot be increased by operating the oscillator at a subharmonic of the desired carrier frequency and using frequency multipliers. Increased frequency deviation can be achieved by heterodyning the output of an f-m oscillator operating at a high radio frequency with the output of a crystal-controlled oscillator and then multiplying the low-frequency difference by the factor necessary to obtain the desired carrier frequency as in Fig. 5.29. The disadvantage is that any instability in the center frequency of the oscillator is increased by the same factor that the frequency deviation is increased.

Since the r-f grid voltage of the reactance tube is not exactly in quadrature with the r-f plate voltage, there is some resistive loading placed upon the oscillator resonant

circuit. This loading varies with changes in tube transconductance, causing a small amount of amplitude modulation of the oscillator. This amplitude modulation can be eliminated by operating any class C amplifier or frequency multiplier following the oscillator in a "saturated" condition, i.e., with sufficient drive that the r-f output amplitude is not affected by small changes in r-f drive.

The principal difficulty encountered in reactance-tube modulated oscillators is the instability of the modulated oscillator frequency. The desired frequency stability is usually 0.01 to 0.001 per cent. To achieve the best possible frequency stability when large frequency deviations are not required, the oscillator is operated at low-power levels and at frequencies of about 100 kc to several megacycles. The output is then multiplied in frequency and the power output increased by class C harmonic amplifiers. If the required frequency deviation can be achieved at a frequency somewhat lower than the desired carrier frequency, the frequency stability of the unmodulated carrier can be improved by the system of Fig. 5.30. For the frequencies indicated in Fig. 5.30, the stability of the unmodulated carrier at 50 Mc would be fifty times greater than that obtained by straight multiplication of the oscillator output, assuming variations in the crystal oscillator frequency to be negligible.

Fig. 5.29. Method of increasing frequency deviation of a reactance tube modulated oscillator.
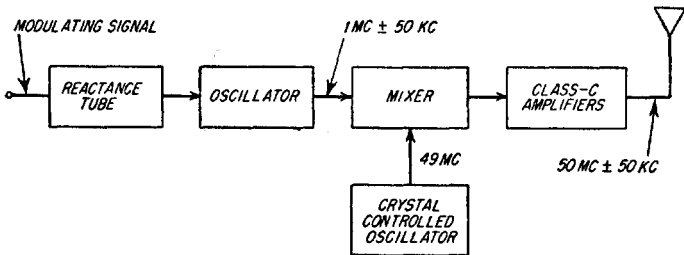
Fig. 5.30. Method of improving frequency stability of f-m system over that obtainable by multiplying oscillator output directly.

The best frequency stability is achieved by a closed-loop frequency-monitoring system in which the mean carrier frequency is compared against some absolute reference. One method of accomplishing this is shown in Fig. 5.31. The crystal oscillator reference is used instead of a discriminator at 1.0 Mc because the frequency stability of the oscillator is superior to that of the discriminator. The d-c motor controls the tuning of a variable capacitor in the oscillator resonant circuit. The frequency response of the frequency control loop should be substantially less than the lowest modulating frequency in order that the control loop not degenerate the modulation, but it must be high enough to correct for any rapid drift variations. The same stability criteria apply to the frequency control loop as in other closed-loop systems (see Secs. 18 and 19).

The Astable Multivibrator. An f-m carrier can be generated by the use of an astable multivibrator and filter as shown in Fig. 5.32a. The design of astable multivibrators is discussed in Sec. 8.5. By inserting the modulating voltage in series with the grid return of the multivibrator tubes, the gate length and, therefore, the fundamental
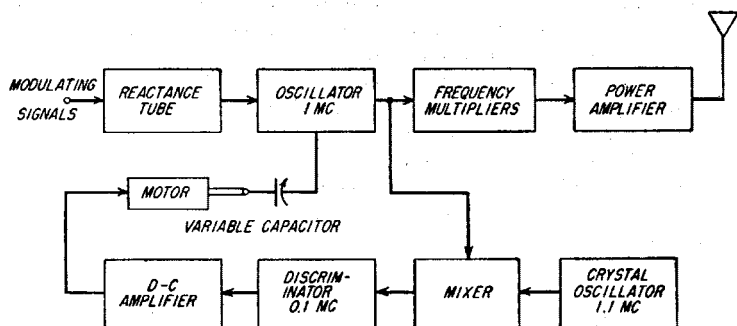
Fig. 5.31. Closed-loop control of the mean frequency of a frequency-modulated oscillator.
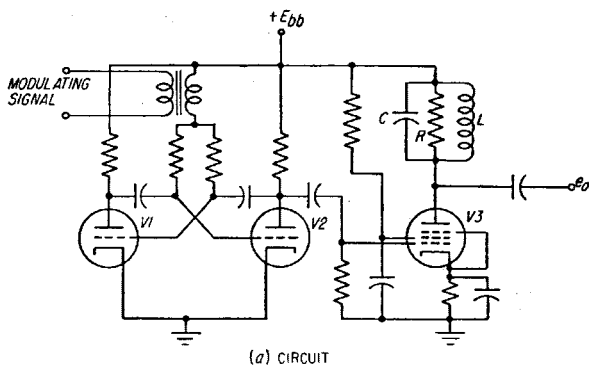
(a) CIRCUIT
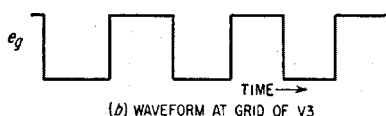
(b) WAVEFORM AT GRID OF V3
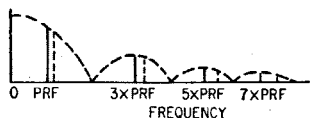
(c) SPECTRUM OF WAVEFORM AT GRID OF V3

Fig. 5.32. Generation of a frequency-modulated carrier by an astable multivibrator and filter.

frequency of the multivibrator can be varied in accordance with the modulating voltage. The fundamental frequency of the multivibrator should be high compared to the modulating frequencies to be employed. A fundamental frequency of 100 kc to 1 Mc is typical when audio frequency modulation is used.

The multivibrator output waveform is shown in Fig. 5.32b, and the frequency

spectrum of this waveform is shown in Fig. 5.32c.   In general, the frequency spectrum of the multivibrator output consists of signals at the fundamental frequency and all of its harmonics (see Sec. 22.4).   For the symmetrical multivibrator (Fig. 5.32), having square wave output, all even harmonics are eliminated (Fig. 5.32c).   As the gate length of the multivibrator is varied by the modulation voltage, the spectrum of the output square wave shifts in frequency as indicated by the dotted lines in Fig. 5.32c.   By passing the multivibrator output waveform through a bandpass filter, as shown by the *LCR* parallel resonant circuit in Fig. 5.32a, any one of the spectrum signals can be amplified and the remainder eliminated, thus providing a single carrier which is frequency-modulated.

By proper design of the multivibrator, the frequency deviation of the carrier can be made a linear function of the modulating voltage up to frequency deviations which are a relatively large fraction of the fundamental frequency of the multivibrator. The principal design consideration is that the *RC* coupling from the grid of each
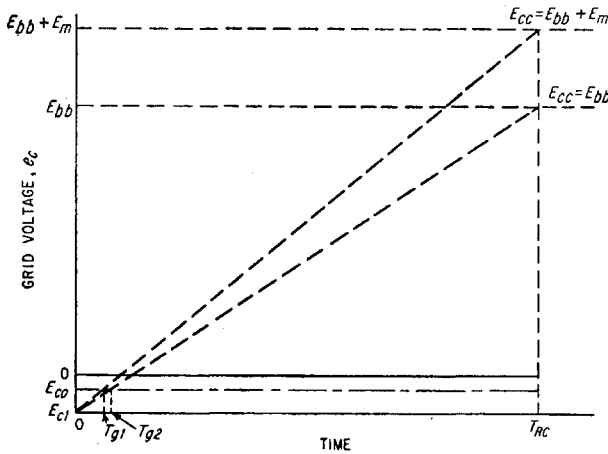
FIG. 5.33. Graphical relationship between multivibrator gate length T_g and grid supply voltage when the gate length is much shorter than the coupling time constant T_{RC}.

multivibrator tube to the plate of the other tube should have a time constant which is at least ten times the actual gate length.   When this condition exists, the grid voltage of each tube when cut off rises essentially linearly to the cutoff bias, at which time the tube is switched on.   If this rise in grid voltage is a linear function of time, the change in gate length will be an inverse function of the modulation voltage and, hence, the frequency deviation will be essentially a linear function of the modulating voltage.   This condition is illustrated in Fig. 5.33 where the gate duration T_g is shown for two values of grid supply voltage E_{cc}.   The initial slope of the grid bias waveform is

$$\frac{de_c}{dt} = \frac{E_{cc} - E_{c1}}{T_{RC}} \tag{5.60}$$

where E_{c1} = grid voltage at beginning of cutoff period for either tube
    T_{RC} = time constant of *RC* coupling circuits between tubes
For the circuit of Fig. 5.32a, where the grid is returned to E_{bb}, the duration of the cutoff period T_{g2} when the modulation voltage is zero is given approximately by

$$T_{g2} \simeq \frac{E_{co} - E_{c1}}{E_{bb} - E_{c1}} T_{RC} \tag{5.61}$$

When the modulation voltage has a value E_m, the duration of the cutoff period T_{g1} is given approximately by

$$T_{g1} \simeq \frac{E_{co} - E_{c1}}{E_{bb} + E_m - E_{c1}} T_{RC} \qquad (5.62)$$

The frequency deviation Δf is given by

$$\Delta f = f_1 - f_2 = \frac{1}{T_{g1}} - \frac{1}{T_{g2}} \qquad (5.63)$$

Thus

$$\Delta f \simeq \frac{E_m}{T_{RC}(E_{co} - E_{c1})} \qquad (5.64)$$

The multivibrator system possesses the inherent advantage over the reactance tube modulator system of much greater linear frequency deviation for a given carrier frequency. However, the fundamental frequency of the multivibrator is limited to frequencies below about 1 Mc. The same considerations of drift of the mean frequency of the carrier exist in the multivibrator system as in the reactance tube modulated oscillator system, and the same methods of stabilization are applicable.
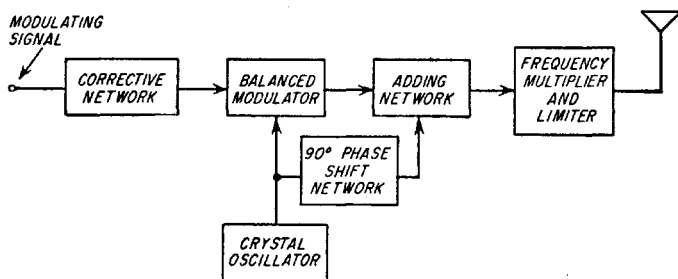
FIG. 5.34. Armstrong type of phase modulator.

5.6b. Phase Modulators. In phase modulators, the phase of an r-f carrier is made to vary in accordance with the amplitude of the modulating signal. A phase modulator can provide a frequency-modulated carrier by the use of an integral network as discussed in Sec. 5.5a. Although numerous methods of phase modulation have been devised, a simple and widely used system is the one used by Armstrong[1] in his early work in frequency modulation. This method as applied to the generation of an f-m carrier is illustrated in Fig. 5.34. If the modulating signal e_m is given by

$$e_m = E_m \cos \omega_m t \qquad (5.65)$$

the output e_s of the corrective network (Fig. 5.27b) is

$$e_s = \frac{E_m}{\omega_m RC} \sin \omega_m t \qquad (5.66)$$

The output of the balanced modulator is given by Eq. (5.68) when the carrier is given by Eq. (5.67):

$$e_c = E_c \cos \omega_c t \qquad (5.67)$$

$$e_o = \frac{E_m}{\omega_m RC} \sin (\omega_c + \omega_m)t + \frac{E_m}{\omega_m RC} \sin (\omega_c - \omega_m)t \qquad (5.68)$$

[1] E. H. Armstrong, A Method of Reducing Disturbances in Radio Signaling by a System of Frequency Modulation, *Proc. IRE*, vol. 24, pp. 689–740, May, 1936.

The addition of a carrier shifted 90° from that of Eq. (5.67) gives an output signal from the adder equal to

$$e_t = E_c \sin \omega_c t + \frac{E_m}{\omega_m RC} \sin (\omega_c + \omega_m)t + \frac{E_m}{\omega_m RC} \sin (\omega_c - \omega_m)t \qquad (5.69)$$

The vector representation of e_t is shown in Fig. 5.35a. The vector representation of e_t when the reference is $\sin \omega_c t$ is shown in Fig. 5.35b. The phase angle ϕ of the vector sum of the carrier and sidebands is seen to be

$$\phi(t) = \theta(t) + \Delta\theta(t) = \omega_c t + \tan^{-1}\left(\frac{2E_m}{E_c \omega_m RC} \cos \omega_m t\right) \qquad (5.70)$$

If the ratio $2E_m/E_c$ is much less than unity, Eq. (5.70) can be rewritten as

$$\phi(t) \simeq \omega_c t + \frac{2E_m}{E_c \omega_m RC} \cos \omega_m t \qquad (5.71)$$

The instantaneous value of the vector sum of the carrier plus the sidebands is, therefore,

$$e_t \simeq \sqrt{E_c^2 + \left(\frac{2E_m}{\omega_m RC} \cos \omega_m t\right)^2} \cos \left(\omega_c t + \frac{2E_m}{E_c \omega_m RC} \cos \omega_m t\right) \qquad (5.72)$$

Comparison of Eq. (5.72) with (5.46) shows that e_t is an f-m signal which also contains a small amount of amplitude modulation. The peak frequency deviation is

$$\Delta f = \frac{E_m}{\pi E_c RC} \qquad (5.73)$$

By passing e_t through a limiter to remove the amplitude modulation and by maintaining Δf quite small, the resulting signal is an f-m carrier whose frequency deviation Δf is a linear function of the amplitude of the modulating signal only. The maximum phase deviation achievable with good linearity is approximately 0.5 radian. Therefore the carrier frequency must be multiplied many times if a high modulation index is to be obtained. For example, assume that a peak deviation of

(a) AMPLITUDE MODULATION SIDEBANDS FROM BALANCED MODULATOR TO WHICH 90° PHASE SHIFTED CARRIER HAS BEEN ADDED   (b) SIDEBAND RELATIONSHIPS REFERRED TO PHASE SHIFTED CARRIER

FIG. 5.35. Vector representations of Armstrong phase modulator output.

± 75 kc is required at the radio frequency and that a peak phase deviation of 0.5 radian is obtainable at the modulator output. The modulation index at the modulator output is

$$\Delta\phi = \frac{\Delta f}{f_m} = 0.5$$

where f_m is, in this case, the lowest modulation frequency to be transmitted.

Since the frequency deviation Δf is proportional to the modulating signal amplitude but independent of the modulating frequency, the maximum modulation index of 0.5 can be achieved only at the lowest modulation frequency to be transmitted. If f_m is 100 cycles, the factor M by which the carrier must be multiplied to achieve the 75-kc deviation at the radio frequency is

$$M = \frac{75,000}{100 \times 0.5} = 1,500$$

The outstanding advantage of the phase-type modulation methods is that the carrier frequency can be crystal-controlled to achieve the required frequency stability, thus eliminating the necessity of the closed-loop type of automatic frequency control system usually required in the direct frequency modulation systems. The principal
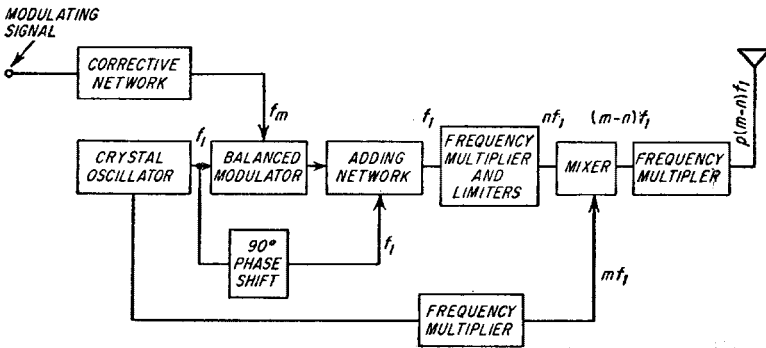
FIG. 5.36. Method for obtaining a large increase in modulation index of a frequency-modulated carrier without increasing the carrier frequency by the same factor.

disadvantage of the phase modulation systems is the small modulation index achievable with reasonable linearity and the consequent necessity of the large amount of frequency multiplication to reach the desired modulation index.

When the required frequency multiplication factor is so large that the original carrier frequency cannot be made low enough to obtain the multiplication directly, the method indicated in Fig. 5.36 can be used. After multiplying the f-m crystal-controlled carrier and the modulation index by the factor n, the carrier is heterodyned with the mth harmonic of the original carrier frequency. The low frequency difference, which retains the increased modulation index, can then be multiplied again by the required factor p to achieve the desired modulation index at the radio frequency.

5.7. Pulse Modulation. In pulse-modulated systems, one or more parameters of the pulse are varied in accordance with a modulating signal to transmit the desired information. The modulated pulse train may in turn be used to modulate a carrier in either angle or amplitude. Pulse modulation provides a method of time duplexing, since the entire modulation information of a single channel can be contained in a single pulse train having a low-duty cycle, i.e., ratio of the pulse width to the inter-

(a) UNQUANTIZED MODULATION SYSTEM

(b) QUANTIZED MODULATION SYSTEM

FIG. 5.37. Input versus output relationships of quantized and unquantized pulse-modulation systems.

pulse period, and therefore the time interval between successive pulses of a particular channel can be used to transmit pulse information from other channels.

Pulse-modulation systems can be divided into two basic types: unquantized pulse

modulation, where the pulse parameter which is varied in accordance with the modulating signal is a continuous function of the modulating signal, and quantized pulse modulation,[1] where the continuous information to be transmitted is approximated by a finite number of discrete values, one of which is transmitted by each single pulse

(a) MODULATING SIGNAL

(b) QUANTIZED PULSE CODE GROUP
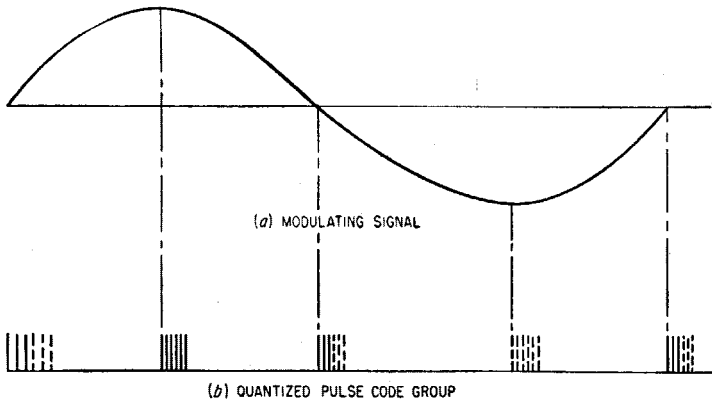
FIG. 5.38. Example of a quantized pulse-modulation system.

or group of pulses. The two methods are illustrated in Fig. 5.37. In quantized PM systems, the input function can be approximated with arbitrary accuracy by adjustment of the number of discrete values available to describe the input function. An example of a quantized PM system is shown in Fig. 5.38. The information is transmitted in pulse code groups, the sequence of pulses sent each period indicating a discrete value of the modulating signal at that instant. Typically, the pulse group might employ the binary number code, the presence or absence of each pulse in the group indicating a 1 or 0 in the binary representation of the modulating signal.

The principal methods for transmitting information by means of unquantized pulse modulation are pulse-amplitude modulation (PAM), pulse-width modulation (PWM), and pulse-position modulation (PPM).

5.7a. Pulse-amplitude Modulation. In pulse-amplitude modulation a pulse train is amplitude-modulated by a modulating signal (see Fig. 5.39a). The amplitude-modulated pulses can then be used to either amplitude- or angle-modulate an r-f carrier. The spectrum of the amplitude-modulated pulse train is given by

MODULATING
SIGNAL

PULSE TRAIN

(a) AMPLITUDE MODULATED PULSE TRAIN

$0 \ f_m \ \frac{1}{T} \ \frac{2}{T} \ \frac{3}{T} \ \frac{4}{T} \ \frac{5}{T} \ \frac{6}{T} \ \frac{7}{T} \ \frac{8}{T} \ \frac{9}{T} \ \frac{10}{T} \ \frac{11}{T}$
FREQUENCY

(b) FREQUENCY SPECTRUM OF THE MODULATED PULSE TRAIN

FIG. 5.39. Pulse-amplitude modulation.

$$e(t) = \left(1 + \frac{E_m}{E_p} \cos \omega_m t\right)\left[\frac{E_p \tau}{T} + \sum_{n=1}^{n=\infty} \frac{2E_p}{\pi n} \sin (\pi n \tau / T) \cos (2\pi n t / T)\right] \quad (5.74)$$

[1] For more details see H. S. Black, "Modulation Theory," p. 299, D. Van Nostrand Company, Inc., New York, 1953.

where E_m = peak modulating voltage

E_p = peak pulse voltage without modulation

τ = pulse width

T = interpulse period

The frequency components of the amplitude-modulated pulse train are illustrated in Fig. 5.39b. Each of the harmonic-frequency signals of which the unmodulated pulse train is composed are amplitude-modulated by the modulating signal, resulting in sideband signals displaced above and below each harmonic by the modulating frequency. The d-c component of the unmodulated pulse train is also modulated by the modulating signal, resulting in a signal at the modulating frequency as shown in Fig. 5.39b.
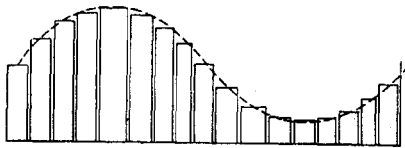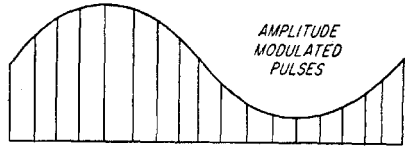
After modulating an r-f carrier for transmission of the above modulation and demodulation of the modulated carrier at the receiver by a suitable amplitude or angle detector, the modulating signal can be extracted from the modulated pulses by peak detecting in a boxcar stretcher and filtering as shown in Fig. 5.40a, or by passing the pulses through a low-pass filter to extract only the modulation component (f_m in Fig. 5.39b).

The circuit of the boxcar pulse stretcher is shown in Fig. 5.40b. The input pulses charge C through diode $V1$ to a voltage e_o equal to the peak value of the input pulse.

(a) PEAK DETECTION
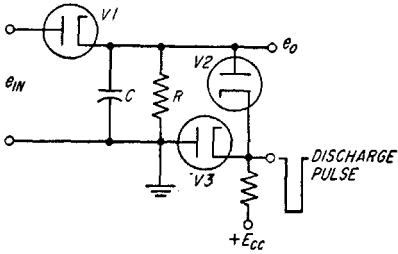
(b) BASIC CIRCUIT OF BOXCAR DETECTOR

FIG. 5.40. Demodulation of amplitude-modulated pulses.

The discharge time constant RC is made at least ten times the interpulse period T, so that the change in e_o between pulses is quite small. Just prior to the arrival of each input pulse, a negative pulse is applied to the cathodes of $V2$ and $V3$, discharging C through $V2$ to 0 volts. The diode $V3$ serves to clamp e_o at near 0 volts during the discharge period.

5.7b. Pulse-width Modulation (PWM). Pulse-width modulation is also frequently referred to as pulse-duration modulation (PDM) and pulse-length modulation (PLM). In PWM the width of each pulse in a pulse train is made proportional to the instantaneous value of the modulating signal at the instant of the pulse (see Fig. 5.41). Either the leading edge or the trailing edge or both edges of the pulses may be modulated to produce the variation in pulse width.

PWM can be obtained by a number of methods, one of which is illustrated in Fig. 5.42. By adding the modulating signal to a repetitive sawtooth, the waveform of Fig. 5.42b is obtained. This waveform is then applied to any suitable circuit which changes state when the input signal exceeds a specified threshold level to produce pulses whose width is determined by the length of time that the input waveform exceeds the threshold level. The resulting pulse output is shown in Fig. 5.42c.

Demodulation of width-modulated pulses can be accomplished in several ways. Since the average value of the pulse train varies in accordance with the modulation, as was the case in PAM, the modulation can be extracted by passing the width-

modulated pulses through a low-pass filter which passes only the desired modulation frequencies. Width-modulated pulses can also be demodulated by passing the pulses through a pulse stretcher of the type described in Fig. 5.40, modified so that the time constant for charging C through $V1$ is at least ten times the maximum pulse width. The amplitude of the voltage to which C charges will then be directly proportional to the pulse width.

MODULATING SIGNAL

MODULATED PULSE TRAIN
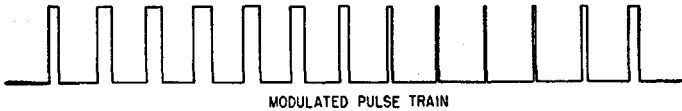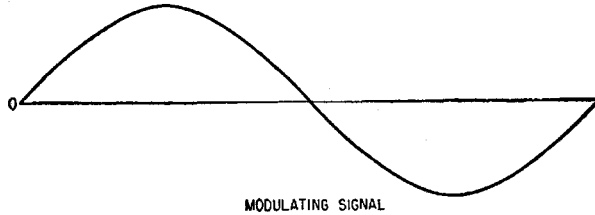
FIG. 5.41. Pulse-width modulation.

(*a*) REPETITIVE SAWTOOTH

(*b*) MODULATING SIGNAL ADDED TO SAWTOOTH

(*c*) WIDTH MODULATED PULSES OBTAINED FROM REGENERATIVE SWITCH
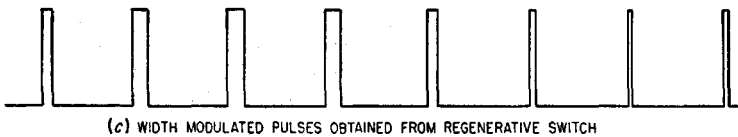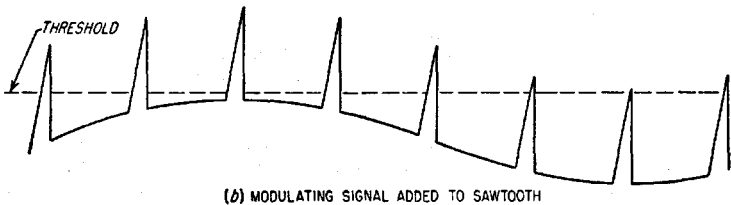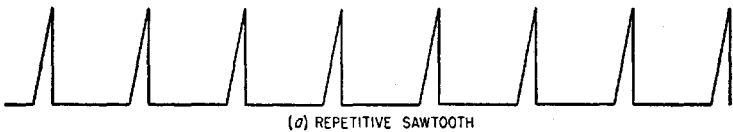
FIG. 5.42. Method of generating width-modulated pulses.

5.7c. Pulse-position Modulation (PPM). In PPM a modulating signal causes the position in time of a pulse to vary relative to its unmodulated time of occurrence (see Fig. 5.43). A reference pulse is often transmitted preceding each time-modulated pulse to synchronize the timing at the transmitting and receiving equipments.

Position-modulated pulses can be generated by a number of methods, one of which is illustrated in Fig. 5.44. The modulating signal is added to a repetitive sawtooth, and a pulse of fixed duration is generated each time the combined signal exceeds a fixed threshold level.
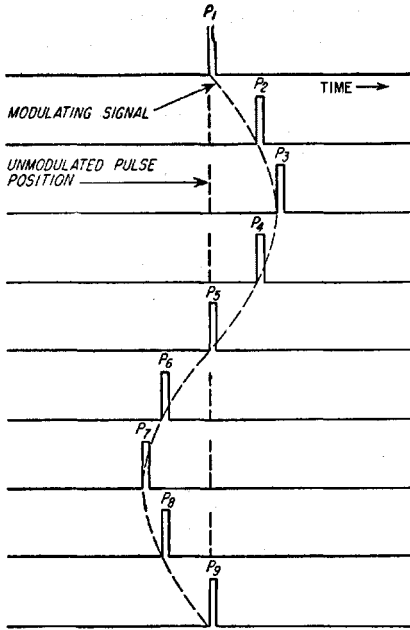
FIG. 5.43. Pulse-position modulation.

REPETITIVE SAWTOOTH

POSITION OF UNMODULATED PULSE

POSITION MODULATED PULSES
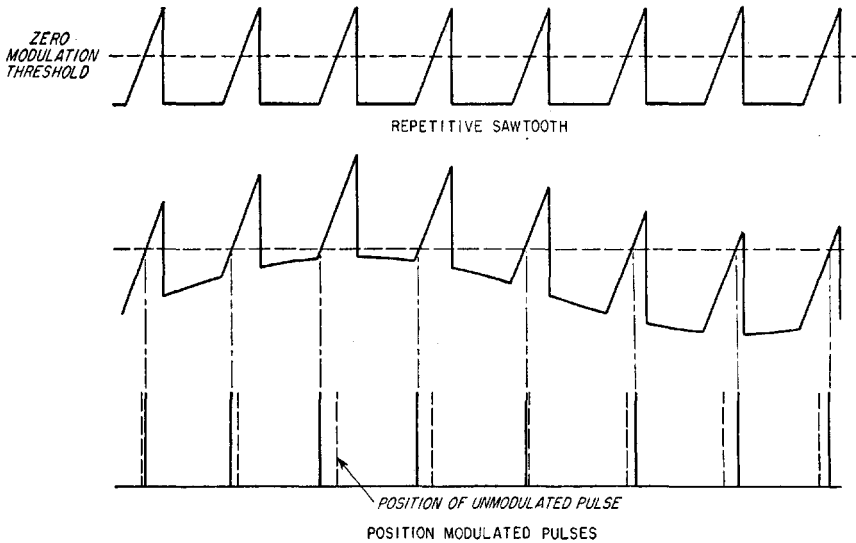
FIG 5.44. Method of producing position-modulated pulses.

The effect of pulse-position modulation upon the pulse-frequency spectrum is to frequency-modulate each of the harmonic components of the pulse spectrum and the d-c term (see Sec. 22.6).

When the peak variation in pulse timing is small compared to the interpulse period, the position-modulated pulses can be demodulated with small distortion by passing

the pulse train through a network having a frequency response with a slope of −6 db per octave throughout the range of modulating frequencies.[1]   An alternative method is to convert the position modulation to amplitude modulation by charging a peak detector to a voltage which is proportional to the time interval between a reference pulse and the position-modulated pulse (see Fig. 5.45).   A sawtooth is initiated by each reference pulse, and the peak detector is connected to the sawtooth for the time interval between the reference pulse and the signal pulse.   At the instant the signal pulse occurs, the switch between the sawtooth and the detector is opened and the
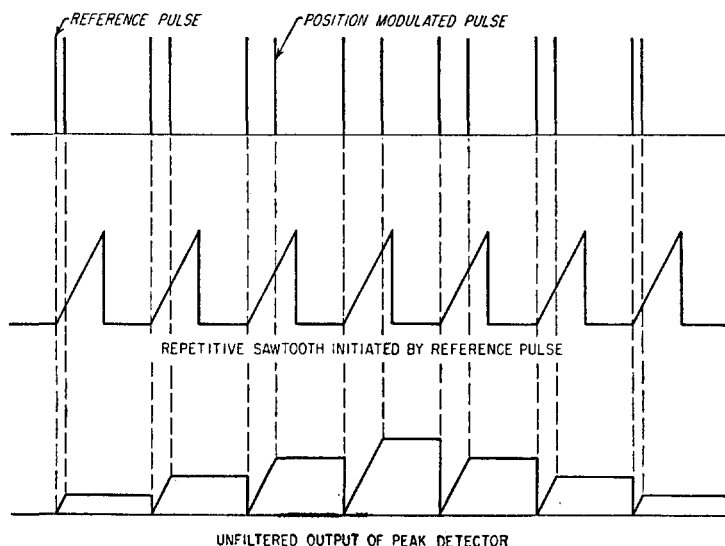
UNFILTERED OUTPUT OF PEAK DETECTOR
FIG. 5.45. Method of demodulating position-modulated pulses.

detector remains charged at that voltage until the next reference pulse again closes the switch.

5.7d. Pulse-code Modulation.   PCM is a system where each sampled value of the modulating signal is approximated by a discrete value which is represented by a coded series of pulses (see Fig. 5.38).   The modulating signal can be approximated as closely as desired by employing a pulse code containing a sufficient number of pulses. If the binary code is employed, the pulse-code modulation can transmit 2^n discrete values approximating the modulating signal to one part in $2^n - 1$, where n is the number of pulses in the code.

The outstanding advantages of PCM are twofold.   First, there is almost complete elimination of noise interference when the pulse signals are above noise by a specific threshold value since only the presence or absence of each pulse need be determined to find the exact value of the transmitted signal.   Second, the signal can be received and retransmitted as many times as desired without introducing progressive distortion of the signal.[2]

5.7e. Information Rates in Pulse-modulation Systems.   In any type of pulse-modulation system, the rate of information transmittal is limited by the rate at which the modulating signal is sampled.   For the frequency of a modulating signal to be reproduced, it cannot exceed one-half the sampling rate.   The effect of the sampling frequency f_s on the modulating-signal frequencies is illustrated in Fig. 5.46.   This

[1] *Ibid.*, p. 287.
[2] B. M. Oliver, Efficient Coding, *Bell System Tech. J.*, pp. 724–750, July, 1952.

effect is a result of the fact that the modulating signal produces sideband signals about the pulse-frequency spectrum signals as shown in Fig. 5.39 which overlap when the modulating signal frequency exceeds $f_s/2$. Thus, the low-frequency modulation
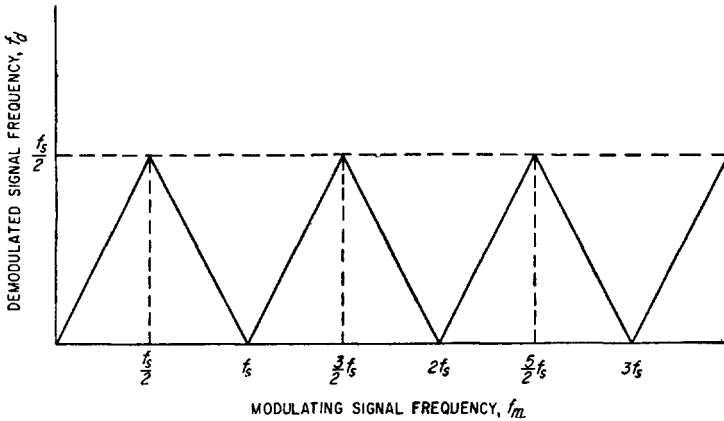
FIG. 5.46. Effect of sampling on signal frequency.

component, for example, f_m in Fig. 5.39, effectively varies between zero and $f_s/2$ as limits.

The sampling of the modulating signal also introduces a phase lag θ between the demodulated signal and the modulating signal which is a function of the ratio f_m/f_s.
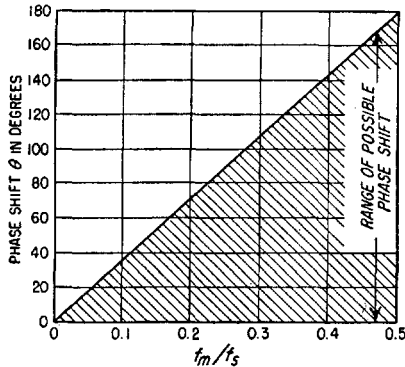
FIG. 5.47. Phase shift of the demodulated signal relative to the modulating signal as a function of the ratio of the modulating signal to the sampling frequency.

Depending upon the points in the modulating frequency cycle at which the signal is sampled, the phase shift θ lies between the limits shown in Fig. 5.47.

These effects of sampling must be considered in the application of any type of a pulse-modulation system to a particular situation (see Sec. 22.7).

Oscillators

Oscillators which produce an output voltage and current which vary substantially sinusoidally as a function of time are discussed. The six types treated are: four-terminal LC oscillators, two-terminal LC oscillators, crystal oscillators, RC oscillators, uhf oscillators, and microwave oscillators. The microwave-oscillator classification includes all types of oscillators which utilize the finite transit time of electrons between the tube electrodes for the transfer of energy from the electron current in the tube to the resonant circuits.

6.1. General Considerations in Oscillator Design. Several performance characteristics must be considered in selecting the best oscillator circuit for a particular application. These characteristics are (1) frequency, (2) frequency stability, (3) amplitude stability, and (4) power output.

6.1a. Frequency of Operation. At frequencies below 100 kc, RC oscillators have the advantages of good frequency stability, a wide tuning range since frequency varies inversely with capacitance instead of inversely as the square root as in LC oscillators, constant power output over wide tuning ranges, and no bulky inductors.

Oscillators of the LC type are used extensively at frequencies from 100 kc to 500 Mc. Low-power LC oscillators can be constructed which will operate satisfactorily, with special tubes, at frequencies as high as 4,000 Mc. However, the effects of electron transit time, tube capacitances, and lead inductances severely limit the power output at frequencies above about 1,000 Mc. At frequencies above 100 to 200 Mc, the use of mutually coupled coils becomes very difficult, and the Colpitts and tuned-plate tuned-grid circuits are used almost exclusively.

At frequencies above 1,000 Mc, magnetron and klystron oscillators are used extensively. The present upper frequency limit of magnetron and klystron operation is approximately 40,000 Mc.

6.1b. Frequency Stability. The frequency of oscillation is dependent upon the parameters of the tube (for example, μ, r_p, and interelectrode capacitance), the Q of the resonant circuit, and the load impedance in addition to the values of L and C in the resonant circuit. Certain oscillator circuits are less subject to variations in frequency due to changes in load impedance and tube parameters than others, and because of this feature they are to be preferred in special applications.

The sensitivity of the frequency of the oscillator to variations in tube characteristics is dependent upon the loaded Q of the oscillator resonant circuit. Circuit resistance due to coil loss, grid input loading, or reflected load resistance causes the frequency of oscillation to be dependent upon tube plate resistance [for example see Eq. (6.6)]. The rate of change of frequency with plate resistance is dependent upon the magnitude of the circuit resistance and certain circuit reactances and must be evaluated for each oscillator circuit individually. Since plate resistance varies with tube plate voltage and plate current, the frequency of an oscillator is dependent upon supply voltages.

At frequencies above approximately 50 Mc the grid-cathode capacitance c_{gk} is also sensitive to plate current changes because the transit-time input capacitance is a function of the tube transconductance (see Sec. 7.4h). As a result, the frequency of an oscillator will vary with changes in plate current. The frequency change caused by this effect will depend upon the per cent of the total grid-cathode capacitance that the transit-time capacitance represents.

6-2

Frequency stability is also a function of the stability of the components used in the oscillator circuit. Temperature-compensated components should be used whenever variations in such components with changes in temperature will result in undesirable frequency variations.

6.1c. Amplitude Stability. It is usually desirable to maintain the output power from an oscillator nearly constant as a function of time and frequency if the oscillator is to be tuned over a band of frequencies. The type of bias used will greatly affect the amplitude stability. Grid-leak bias is to be preferred in most cases because the bias voltage thus developed is proportional to the signal amplitude, thereby providing a form of automatic amplitude control. In cases where extreme amplitude stability is required, automatic amplitude control circuits employing negative feedback are often used. Variations in output with oscillator tuning are minimized in circuits where the frequency-determining circuit is isolated from the load, e.g., the electron-coupled oscillator, and in circuits where the feedback voltage amplitude is independent of the frequency to which the oscillator is tuned.

Fig. 6.1. Block diagram representation of an oscillator.

6.1d. Power Output. The required power output will determine in large part the type of tube to be used, and, in certain cases, it may restrict the oscillator circuits which can be used; e.g., the crystal-controlled oscillator and the reflex-klystron oscillator are capable of only relatively small power output.

The proper operating conditions for maximum efficiency and power output for an oscillator are determined in exactly the same manner as for the same tube operated as an amplifier with the single additional factor that the power output as an oscillator is less than the power output as an amplifier by the amount that must be fed back to the grid circuit to sustain oscillation. Power oscillators are normally operated class C, so that the design information in Sec. 4.4 can be applied directly provided that the required grid driving power is subtracted from the output power calculated for the tube as an amplifier.

6.1e. Criteria for Oscillation. With the exception of two-terminal LC oscillators (Sec. 6.3), any vacuum-tube oscillator can be considered as an amplifier plus a frequency selective feedback network. This is illustrated in Fig. 6.1. The required condition for sustained oscillation is

$$A\beta = 1 \qquad (6.1)$$

where A = gain of amplifier
β = transfer characteristic of frequency selective network

This means that the product of A and β must have a magnitude of 1 and a phase angle of 0° at the desired frequency if oscillation is to be sustained.

The conditions necessary for oscillation in any circuit can be established by determining A and β for the circuit and setting their product equal to 1 at a phase angle of 0°.

An equivalent method of analysis which is sometimes easier to apply is to write the mesh equations of the circuit in matrix form, expressing all voltages in terms of corresponding loop currents (see Sec. 23). The impedance determinate is then set equal to zero, and the resulting simultaneous equations solved for the conditions necessary for oscillation. For example, if the oscillator circuit is expressed as a two-mesh circuit, the circuit equations can be written in matrix form as shown in Eq. (6.2).
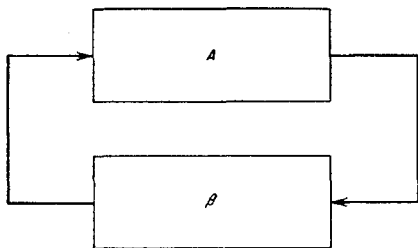
$$\begin{vmatrix} e_1 \\ e_2 \end{vmatrix} = \begin{vmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{vmatrix} \times \begin{vmatrix} i_1 \\ i_2 \end{vmatrix} \qquad (6.2)$$

where e_1 is ordinarily taken as $-\mu e_g$ and e_2 is zero. The condition for oscillation is

$$\begin{vmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{vmatrix} = 0 \qquad (6.3)$$

This is a result of the fact that the currents i_1 and i_2 must exist if the circuit is to oscillate, and the driving voltages e_1 and e_2 must be zero if no external voltages are applied. Therefore, for the network equations to be correct, the impedance determinant must also be zero.

The same criterion can be used in the nodal method of analysis, in which case the admittance determinant is set equal to zero to determine the conditions for oscillation. These methods can be used to determine the conditions of oscillation for all oscillator circuits, including two-terminal LC oscillators.
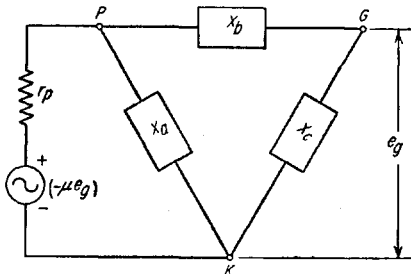
6.2. Four-terminal LC Oscillators. All LC oscillators can be divided into two general classifications: four-terminal oscillators and two-terminal, or negative-resistance, oscillators. In all four-terminal LC oscillators the resonant circuit is a four-terminal network connected between the plate and the grid of the tube. The transfer function, that is, e_{gk}/e_{pk}, of the network provides the necessary 180° phase shift at the frequency of oscillation which together with the 180° phase shift through the tube satisfies the requirements of Eq. (6.1). Although it is termed a four-terminal network, two terminals are usually common and the resonant circuit normally has only three terminals.

FIG. 6.2. Basic circuit for all four-terminal LC oscillators.

In two-terminal LC oscillators, the necessary feedback is obtained external to the tuned circuit, and the external circuit presents a negative resistance across the resonant circuit terminals. Under equilibrium conditions, the magnitude of the negative resistance of the external circuit is just equal to the positive shunt resistance of the resonant circuit.

Four-terminal LC oscillators can be classified according to the manner in which the 180° phase shift of the transfer constant for the resonant circuit is obtained. The basic circuit for all four-terminal LC oscillators is shown in Fig. 6.2. Neglecting circuit losses, the only manner in which the grid-cathode voltage can be 180° out of phase with the effective generator voltage $-\mu e_{gk}$ is for the reactance X_b to have a sign opposite to that of X_a and X_c and for the magnitude of X_b to equal the sum of the magnitudes of X_a and X_c. For this condition, there is a 180° phase shift between e_{gk} and e_{pk}, and the impedance seen looking into the four-terminal network from terminals PK is a pure resistance of infinite magnitude. The grid-cathode voltage e_{gk} is given by

$$e_{gk} = -\frac{X_c}{X_a} e_{pk} \qquad (6.4)$$

The two classes of four-terminal LC oscillators are shown in Fig. 6.3. In the class I oscillator, the reactance between plate and grid is positive, and the reactances between plate and cathode and between grid and cathode are negative. In the class II oscillator, the reactance between plate and grid is negative, and the reactances

between plate and cathode and between grid and cathode are positive.   When trans-
former coupling is used, as in the circuit shown in Fig. 6.4a, the negative mutual
coupling between primary and secondary provides the required 180° phase shift.
The sum of the reactances of the primary inductance and the mutual inductance
is normally quite small compared to the plate resistance r_p of the tube, and therefore
the terminal P of the four-terminal network can be placed as shown in Fig. 6.4b with
negligible error.   This effectively makes X_a or X_c a negative reactance, depending
upon whether the untuned winding is between plate and cathode or between grid
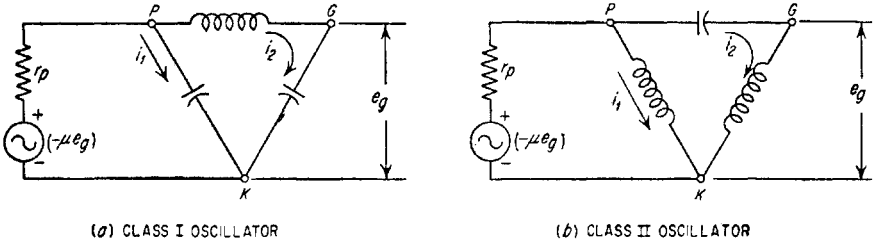
(a) CLASS I OSCILLATOR   (b) CLASS II OSCILLATOR
Fig. 6.3. Classifications of four-terminal LC oscillators.
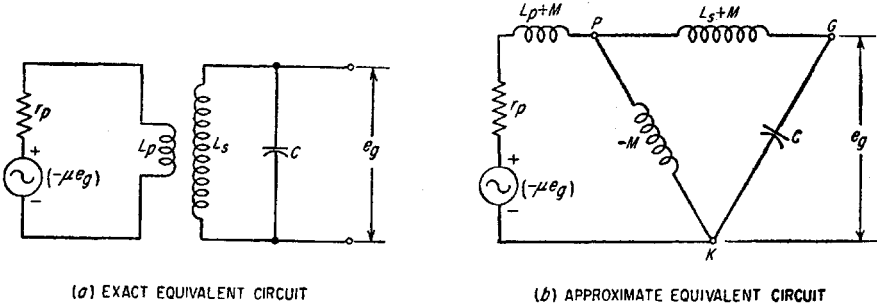
(a) EXACT EQUIVALENT CIRCUIT   (b) APPROXIMATE EQUIVALENT CIRCUIT
Fig. 6.4. Transformer-coupled four-terminal LC oscillators.

and cathode, as shown in Fig. 6.4b.   Any one of the three tube electrodes can be
at a-c ground in either the class I or class II oscillators.
   Because of grid loading, series resistance of the inductances, and loading due to
power being coupled from the oscillator, the impedances between the plate, grid,
and cathode terminals of the tube will not be pure reactances.   As a result, the cur-
rents i_1 and i_2 in Fig. 6.3 will not be of exactly equal magnitude and opposite phase,
and the resonant circuit will not appear as a pure resistance across its terminals PK.
The current and voltage relationships, taking into consideration loss in the elements
through which current i_2 flows, are shown for class I and class II oscillators in Fig. 6.5.
In the class I oscillator, the presence of resistive loading between plate and grid and/or
between grid and cathode causes the current i_2 to lag the voltage e_{pk} by less than 90°
and therefore causes e_{gk} to lag e_{pk} by less than 180°.   As a result, the effective generator
voltage $-\mu e_g$ leads the plate-cathode voltage e_{pk} as shown in Fig. 6.5a.   In the
class II oscillator, the presence of resistive loading between plate and grid and/or
between grid and cathode causes the current i_2 to lead the voltage e_{gk} by less than
90° and therefore causes e_{gk} to lead e_{pk} by less than 180°.   As a result, the effective
generator voltage $-\mu e_{gk}$ lags the plate-cathode voltage e_{pk} as shown in Fig. 6.5b.
   The finite transit time of the electrons in traveling from cathode to grid and plate
causes (1) a resistive loading to appear between grid and cathode which varies inversely
as the square of the frequency of operation and (2) the effective generator voltage
to lag the grid-cathode voltage e_{gk} by more than 180° by an amount which is directly

proportional to the frequency of operation. Thus, the phase angle of the effective
generator voltage in the plate circuit is not exactly 180° out of phase with the grid
voltage and cannot be represented exactly by $-\mu e_{gk}$. At high frequencies these effects

(a) CLASS I OSCILLATOR VECTOR DIAGRAM   (b) CLASS II OSCILLATOR VECTOR DIAGRAM
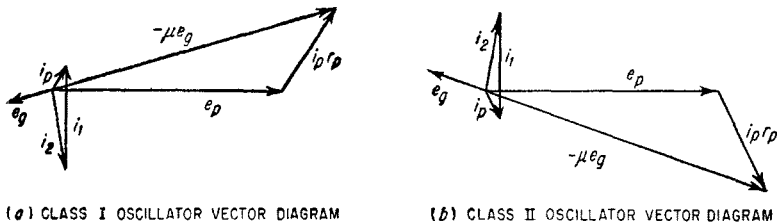
FIG. 6.5. Vector diagrams for four-terminal LC oscillators.

seriously limit the efficiency of oscillator circuits (see Sec. 6.5). In the class II oscilla-
tor the two effects are accumulative, since both cause $-\mu e_{gk}$ to lag e_{pk} more than at
low frequencies. However, in the class I oscillator, the two effects tend to offset
each other. The resistive grid loading causes i_2 and e_{gk} to have less than their desired
phase shifts with respect to e_{pk}, while the transit-time phase shift between e_{gk} and
$-\mu e_{gk}$ causes $-\mu e_{gk}$ to have more than the low-frequency 180° phase shift with respect
to e_{gk}. The partial compensation of these two effects in class I oscillators makes
them superior to class II oscillators at
high frequencies.
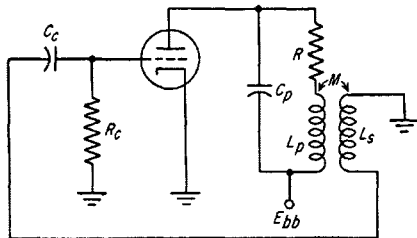
FIG. 6.6. Circuit of tuned-plate oscillator.

A number of four-terminal LC oscilla-
tors are described in detail in the follow-
ing material. At low frequencies, the
power output obtainable from a particu-
lar tube is approximately the same in any
of the various circuits. The choice of
circuit is based principally upon the re-
quirement for frequency stability and the
relative size and complexity of the circuit
components. In any of the oscillator
circuits, the effect of loading of any type upon the frequency of oscillation and the
required amplifier gain can be determined by using the proper value of effective resist-
ance in series with the circuit inductance.

6.2a. Tuned-plate Oscillator. The circuit of the tuned-plate oscillator is shown in
Fig. 6.6. An untuned inductance in the grid circuit is inductively coupled to the
resonant plate circuit. Since the tube imparts a 180° phase shift to signals at its
grid, the mutual coupling between L_p and L_s must be negative to provide an additional
180° phase shift between the voltage across L_p and the voltage across L_s at the fre-
quency of oscillation.

The design equations for the tuned-plate oscillator are given by Eqs. (6.5) and (6.6)
assuming $R_c \gg \omega_{osc}(L_s + M)$.

$$\mu \geq \frac{L_p + R r_p C_p}{M}$$

$$\geq \frac{L_p}{M}\left(1 + \frac{\omega_{osc} r_p C_p}{Q_L}\right) \tag{6.5}$$

$$\omega_{osc} = \sqrt{\frac{1}{L_p C_p}\left(1 + \frac{R}{r_p}\right)}$$

$$= \omega_o \sqrt{1 + \frac{R}{r_p}} \tag{6.6}$$

where μ = tube amplification factor
L_p = inductance in plate circuit, henrys
R = resistance in ohms in series with L_p
r_p = tube plate resistance, ohms
C_p = plate circuit capacitance, farads
M = mutual inductance in henrys between L_p and L_s

Q_L = Q of plate circuit inductance L_p = $\dfrac{\omega_{osc}L_p}{R}$

ω_{osc} = frequency of oscillation, radians/sec
ω_o = $1/\sqrt{L_pC_p}$

6.2b. Tuned-grid Oscillator. The circuit of
the tuned-grid oscillator is shown in Fig. 6.7.
An untuned inductance in the plate circuit is
inductively coupled to the resonant grid cir-
cuit. Since the tube imparts a 180° phase
shift to signals at its grid, the mutual coupling
between L_p and L_s must be negative to pro-
vide an additional 180° phase shift between
the voltages across L_p and L_s at the frequency
of oscillation.
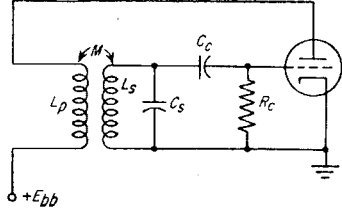
FIG. 6.7. Circuit of tuned-grid
oscillator.

The design equations for the tuned-grid oscillator are given by Eqs. (6.7) and (6.8),
assuming $R_c \gg 1/\omega_{osc}C_s$.

$$\mu \geq \frac{L_p}{M}\left(\frac{1}{1+\dfrac{Q_s}{Q_p}}\right) + \frac{M}{L_s}\left(\frac{1}{1+\dfrac{Q_p}{Q_s}}\right) + \frac{CRr_p}{M} \qquad (6.7)$$

$$\omega_{osc} = \frac{1}{\sqrt{L_sC_s\left(1+\dfrac{RL_p}{r_pL_s}\right)}}$$

or $$\omega_{osc} = \frac{\omega_o}{\sqrt{1+\dfrac{RL_p}{r_pL_s}}} \qquad (6.8)$$

where μ = tube amplification factor
L_p = inductance in henrys in plate circuit
C_s = grid circuit capacitance, farads
R = resistance in ohms in series with L_s
r_p = tube plate resistance, ohms
M = mutual inductance in henrys between L_p and L_s
L_s = inductance in henrys in grid circuit
ω_{osc} = frequency of oscillation, radians/sec

ω_o = $\dfrac{1}{\sqrt{L_sC_s}}$

Q_p = $\dfrac{\omega_{osc}L_p}{r_p}$

Q_s = $\dfrac{\omega_{osc}L_s}{R}$

6.2c. Tuned-plate Tuned-grid Oscillator. The circuit of the tuned-plate tuned-grid
oscillator is shown in Fig. 6.8a. It may be considered a tuned grounded-cathode r-f
amplifier having sufficient capacitive coupling between grid and plate to cause
oscillation.

The equations for the tuned-grid tuned-plate oscillator are very difficult to solve
explicitly for the frequency of oscillation and the minimum value of tube amplification

factor. However, the conditions necessary for oscillation are contained implicitly in Eqs. (6.9) and (6.10), and a simultaneous graphical solution of the two equations can be obtained for specified values of a and W as shown in Fig. 6.8b.

$$\frac{\omega_g}{\omega_p} = \frac{\omega_{osc}}{\omega_p} \sqrt{\frac{W/Q^2 + a + 1 - \left(\frac{\omega_{osc}}{\omega_p}\right)^2}{1 - \left(\frac{\omega_{osc}}{\omega_p}\right)^2 (1 + a)}} \tag{6.9}$$

$$\frac{\omega_g}{\omega_p} = \frac{\omega_{osc}}{\omega_p} \sqrt{1 + a(1 + |A| + W(a - 1) - \frac{W}{(\omega_{osc}/\omega_p)^2}} \tag{6.10}$$

where $a = C_{gp}/C$

$\quad C = C_g = C_p$

$\quad W = R_c/R_L$

$\quad \omega_p$ = resonant frequency of plate circuit = $1/\sqrt{L_p C_p}$

$\quad \omega_g$ = resonant frequency of grid circuit = $1/\sqrt{L_g C_g}$

$\quad \omega_{osc}$ = frequency of oscillation

$\quad Q = Q$ of plate circuit at plate resonant frequency ω_p

$\quad A$ = voltage gain of stage = $-\mu R_L/(r_p + R_L)$

The solid lines in Fig. 6.8b are plots of Eq. (6.9) for various values of plate circuit Q. The frequency of oscillation is determined from these curves for a specified ratio of ω_g/ω_p and a given value of Q. The dotted lines in Fig. 6.8b indicate the minimum stage gain required for the circuit to oscillate for any specified values of ω_g/ω_p and Q. Equations (6.9) and (6.10) are plotted in Fig. 6.8b for the condition that $R_L = R_c$ and that $C_{gp} = 0.1C$. Similar plots can be constructed from Eqs. (6.9) and (6.10) for other values of a and W. From Fig. 6.8b it is seen that the frequency is always lower than the resonant frequency of the plate circuit. It is also seen from Fig. 6.8b that lower circuit Q's require higher stage gain to sustain oscillation and that the grid circuit may be tuned considerably higher in frequency than the plate circuit for low values of Q in the plate and grid circuits.

If the Q's of the grid and plate circuits are high and the grid-plate capacitance is much smaller than the grid circuit capacitance, the frequency of oscillation is given closely by

$$\omega_{osc} \simeq \frac{1}{\sqrt{L_g C_g}} \tag{6.11}$$

For these conditions, the amplification factor required for oscillation is given by

$$\mu \geq \frac{L_g}{L_p} \frac{(1 - \omega_{osc}^2/\omega_p^2)}{(1 - \omega_{osc}^2/\omega_g^2)} \tag{6.12}$$

6.2d. Hartley Oscillator. The circuit of the Hartley oscillator is shown in Fig. 6.9. Since the grid and plate are connected at opposite ends of the tank circuit, the voltage at the grid is 180° out of phase with the voltage at the plate. The point where the cathode is returned to the tank coil determines the amount of voltage fed back to the grid. The plate, cathode, or grid may be operated at ground potential if desired.

The design equations for the Hartley oscillator, assuming infinite coil Q, are given by Eqs. (6.13) and (6.14).

$$\omega_{osc} = \omega_o = \frac{1}{\sqrt{LC}} \tag{6.13}$$

where L = total inductance between grid and plate = $L_1 + L_2 + 2M$

$$\mu \geq \frac{L + L_1 - L_2}{L - L_1 + L_2} = \frac{L_1 + M}{L_2 + M} = \frac{N_1^2}{N_2^2} \tag{6.14}$$

where L_1 and L_2 are as shown in Fig. 6.9 and N_1 and N_2 are the number of turns in L_1 and L_2 respectively.

6.2e. Colpitts Oscillator.   The circuit of the Colpitts oscillator is shown in Fig. 6.10. The circuit is similar to the Hartley in that the grid and plate are at opposite ends of the tank circuit.   However, the cathode is returned to the junction between C_1 and C_2, and the amount of feedback is determined by the relative magnitudes of C_1 and C_2.

(a)

FIG. 6.9. Circuit of the Hartley oscillator.

The design equations for the Colpitts oscillator are given by Eqs. (6.15) and (6.16).

$$\omega_{osc} = \frac{1}{\sqrt{L\dfrac{C_1 C_2}{C_2 + C_1\left(1 + \dfrac{R}{r_p}\right)}}} \tag{6.15}$$

where R = series resistance of inductance L

$$\mu \geq \frac{C_1}{C_2}\left\{ 1 + \frac{R}{r_p} \right.$$
$$\left. + \left[R + r_p\left(1 + \frac{C_2}{C_1}\right)\right]\frac{RC_2}{L} \right\} \tag{6.16}$$

When $r_p \gg R$, which is usually the case, Eq. (6.15) reduces to

$$\omega_{osc} = \omega_o = \frac{1}{\sqrt{LC}} \tag{6.17}$$

where C = total capacitance across L
$= C_1 C_2/(C_1 + C_2)$
When the coil Q is infinite, the required

(b)
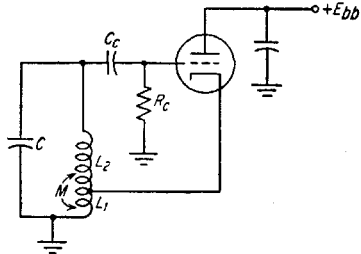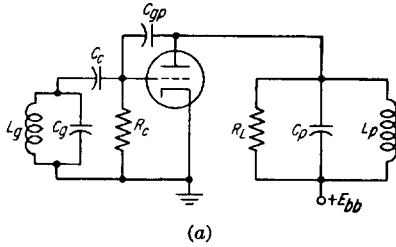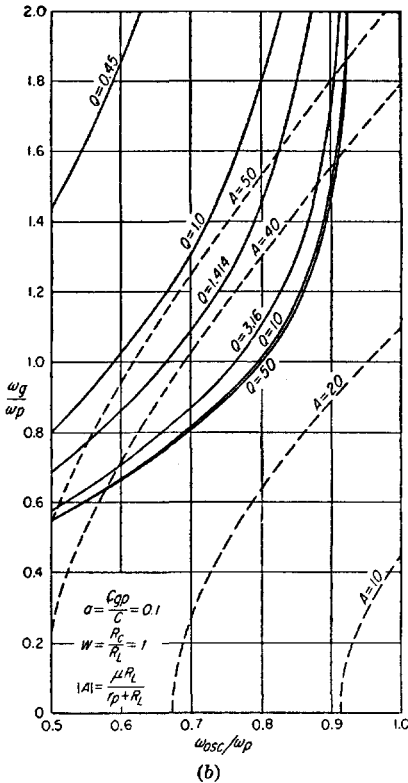
FIG. 6.8. (a) Tuned-plate tuned-grid oscillator; (b) conditions for oscillation in a TPTG amplifier or oscillator having equal grid and plate loads.

value of μ is given by

$$\mu \geq \frac{C_1}{C_2} \tag{6.18}$$

6.2f. Electron-coupled Oscillator.   The design equations for the oscillator types already discussed indicate that the frequency of oscillation is affected by the mag-

nitude of the coil series resistance. Any resistive loading placed across an oscillator tank circuit by the coupling of power from the oscillator constitutes an effective increase in coil series resistance R. Any reactance coupled in by an external circuit will also vary the oscillator frequency. Therefore, any coupling to an oscillator should present no reactance and a small amount ef resistance (low power coupling) if the frequency of the oscillator is to remain nearly constant with variations in the load.

The electron-coupled oscillator combines an oscillator and a power amplifier in one tube envelope, allowing power to be coupled from the power amplifier tank circuit without affecting the frequency stability of the oscillator. The circuit of a typical electron-coupled oscillator is shown in Fig. 6.11. The tube may be either a tetrode or pentode. If a pentode is used, the suppressor should be connected to ground. The cathode, control grid, and screen grid form a conventional oscillator circuit with the screen at r-f ground potential and acting as the plate of a triode oscillator.
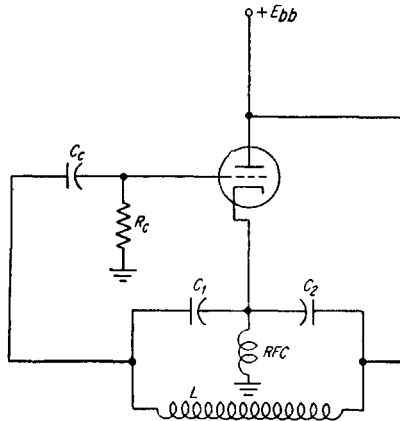
FIG. 6.10. Circuit of the Colpitts oscillator.
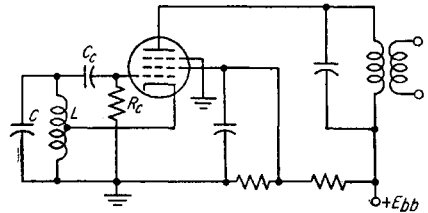
FIG. 6.11. Typical electron-coupled oscillator circuit.

Although a Hartley oscillator circuit is shown in Fig. 6.11, any of the other oscillator circuits can be used.

The screen intercepts a sufficient number of electrons from the cathode to sustain oscillation, and the remaining electrons pass on to the plate. The pulses of plate current produce a sinusoidal voltage at the oscillator frequency across the plate resonant circuit.

The design equations for the oscillators previously discussed may be applied to the electron-coupled oscillator provided that the control-grid to screen-grid μ is used instead of the control-grid to plate μ and provided that the screen-grid resistance r_{sg} is substituted for the plate resistance r_p. The control grid, cathode, or screen grid may be operated at r-f ground in any of these circuits. However, if an oscillator circuit is used where the screen grid is not operated at r-f ground potential, the screen-grid to plate capacitance will require neutralization to prevent interaction between the plate circuit and the oscillator circuit.

If the ratio of the screen to plate voltage is maintained at the proper value, the frequency of an electron-coupled oscillator can be made nearly independent of power-supply voltage changes.[1] The desired ratio of screen-to-plate voltage can be obtained by taking the screen voltage from a tapped voltage divider between the plate supply voltage and ground and adjusting the position of the tap until the frequency variation with supply change is a minimum.

6.2g. Clapp Oscillator. The frequency of an oscillator is affected by variations in the plate and filament supply voltages since changes in these voltages produce

[1] F. E. Terman, Radio Engineers' Handbook, p. 486, McGraw-Hill Book Company, Inc., New York, 1943.

changes in the effective tube interelectrode capacitances and plate resistance. Such frequency changes are minimized by paralleling the tube capacitances with external capacitors which are as large as practical. This can be done in the Colpitts oscillator (see Fig. 6.10) by making C_1 and C_2 as large as possible. The limiting values of C_1 and C_2 are determined by the μ of the oscillator tube. The μ necessary to sustain oscillation is given by

$$\mu \geq \frac{C_1}{C_2} + \frac{Rr_p}{L}(C_1 + C_2) \qquad \text{provided } r_p \gg R \qquad (6.19)$$

Typical values for the minimum reactance of C_1 and C_2 at the frequency of oscillation are 50 to 100 ohms. A much greater improvement in the frequency stability of a Colpitts oscillator can be achieved by the insertion of the capacitor C_3 in Fig. 6.12. This circuit is known as the Clapp oscillator. The frequency of oscillation is given by

$$\omega_{osc} = \sqrt{\frac{1}{L}\left(\frac{1}{C_1} + \frac{1 + R/r_p}{C_2} + \frac{1}{C_3}\right)} \qquad (6.20)$$

or

$$\omega_{osc} = \frac{1}{\sqrt{LC_3 \left[\frac{C_3}{C_1} + \frac{C_3}{C_2}\left(1 + \frac{R}{r_p}\right) + 1\right]}} \qquad (6.21)$$

Fig. 6.12. Clapp oscillator circuit.

The effective capacitance across L is C_1, C_2' and C_3 in series. Capacitor C_3 can be made small compared to C_1 and C_2, and the resonant frequency will be determined primarily by L and C_3. Therefore, variations in C_1 and C_2 and changes in r_p due to supply-voltage variations will have a very small effect upon the resonant frequency. If R is much smaller than r_p, the rate of change of oscillator resonant frequency for both the Colpitts and the Clapp oscillators due to variations in C_1, C_2, or r_p is given by

$$\frac{\Delta f}{\Delta C_1} = -\frac{fC_o}{2C_1^2} \qquad \text{cycles}/\mu\mu\text{f} \qquad (6.22)$$

$$\frac{\Delta f}{\Delta C_2} = -\frac{fC_o}{2C_2^2} \qquad \text{cycles}/\mu\mu\text{f} \qquad (6.23)$$

and

$$\frac{\Delta f}{\Delta r_p} = -\frac{fRC_o}{2C_2 r_p^2} \qquad \text{cycles/ohm} \qquad (6.24)$$

where f = frequency of oscillation, cps
  C_o = total capacitance in resonant circuit assuming $R \ll r_p$

$$= \frac{1}{\dfrac{1}{C_1} + \dfrac{1}{C_2} + \dfrac{1}{C_3}} \qquad \text{for Clapp oscillator}$$

$$= \frac{1}{\dfrac{1}{C_1} + \dfrac{1}{C_2}} \qquad \text{for Colpitts oscillator}$$

In general, the ratios C_o/C_2, C_o/C_2^2, and C_o/C_1^2 can be made many times smaller in the Clapp oscillator than in the Colpitts oscillator. The stability of the Clapp

oscillator can be made as good as several parts per million for a ± 15 per cent variation in supply voltages.[1]

The value of μ which will sustain oscillation in the Clapp oscillator is given by

$$\mu \geq \frac{C_1}{C_2}\left(1 + \frac{R}{r_p}\right) + \frac{C_1 C_2 R r_p}{L}\left(\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3}\right) + \frac{C_1}{L}R^2 \qquad (6.25)$$

6.2h. Push-pull Oscillator. The circuit of a push-pull oscillator is shown in Fig. 6.13. The grid of each tube is capacitively coupled to the end of the tank circuit opposite to the end to which the plate of that tube is connected. The requirement that the plate and grid voltages of each tube have a 180° phase difference is satisfied by connecting the grid and the plate of each tube to opposite ends of the resonant circuit.

The design equations for the push-pull oscillator are given by Eqs. (6.26) and (6.27) if the Q of the inductance is assumed infinite.

$$\omega_{osc} = \frac{1}{\sqrt{LC_o}} \qquad (6.26)$$

$$\mu \geq 1 + \frac{r_p}{R_g} \qquad (6.27)$$

where R_g = grid-to-cathode resistance
   = grid bias resistor R_c in parallel with tube input resistance (see Sec. 7.4h)
L = total inductance of resonant circuit
C_o = total capacitance in parallel with L
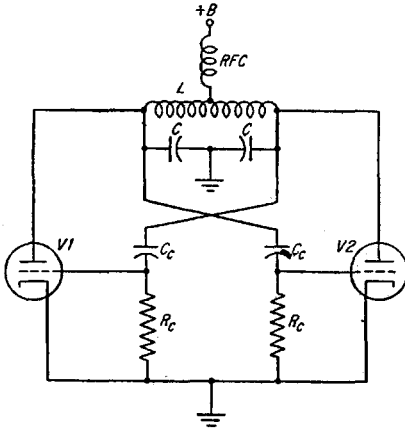
FIG. 6.13. Push-pull oscillator.

Coupling capacitors C_c should have negligible reactance at ω_{osc}. The grid-cathode capacitances of the tubes appear in series across the resonant circuit as do the plate-cathode capacitances of both tubes. The grid-plate capacitance of each tube appears directly across the resonant circuit. The total capacitance C_o shunting the inductance L is given by

$$C_o = \frac{C}{2} + \frac{c_{gk}}{2} + \frac{c_{pk}}{2} + 2c_{gp} \qquad (6.28)$$

At high frequencies parallel resonant lines may be substituted for the resonant circuit of Fig. 6.13.

6.3. Two-terminal LC Oscillators. Oscillation can be produced in a parallel resonant circuit by connecting across the circuit a negative resistance R_n whose magnitude is less than L/CR_s (see Fig. 6.14). This type of oscillator is known as a two-terminal oscillator since only two connections are made to the resonant circuit. All other oscillator types require a four-terminal network to provide the feedback necessary for oscillation. Oscillations will build up in amplitude across the circuit containing the negative resistance until nonlinearities in the device producing
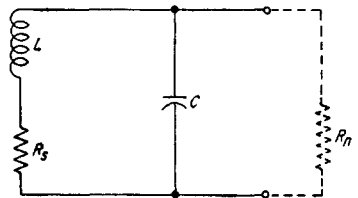
FIG. 6.14. Basic form of a two-terminal negative-resistance oscillator.

---

[1] J. K. Clapp, An Inductance-capacity Oscillator of Universal Frequency Stability, *Proc. IRE*, March, 1948.

the negative resistance cause the negative resistance to increase until it equals the positive resistance L/CR_s across the resonant circuit. At this amplitude the circuit losses are just equaled by the energy from the negative resistance source, and the amplitude of oscillation will be constant.

The frequency of oscillation of a two-terminal oscillator is given approximately by

$$\omega_{osc} \simeq \frac{1}{\sqrt{LC}}$$

where L = total inductance of equivalent resonant circuit
C = total capacity in parallel with L
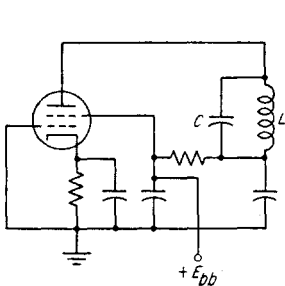
(a) DYNATRON OSCILLATOR
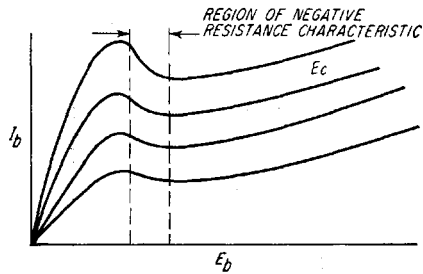
(b) REGION OF TUBE OPERATION IN A DYNATRON OSCILLATOR

FIG. 6.15. The dynatron oscillator.
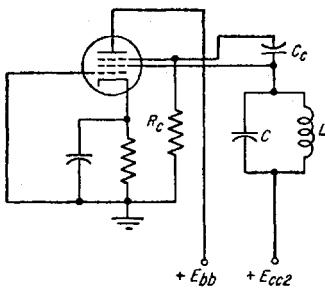
(a) TYPICAL TRANSITRON OSCILLATOR
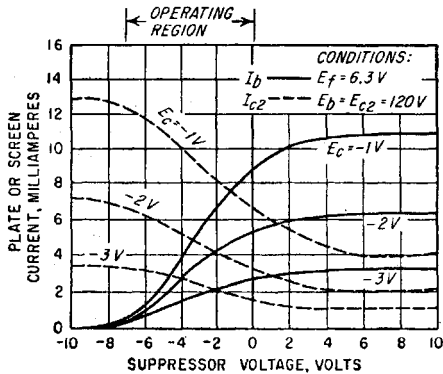
(b) REGION OF TUBE OPERATION IN A TRANSITRON OSCILLATOR

FIG. 6.16. The transitron oscillator.

6.3a. Dynatron Oscillator. The dynatron oscillator is a negative resistance oscillator which utilizes the negative dynamic plate resistance region of a tetrode. The circuit is illustrated in Fig. 6.15a, and the tube operating region is illustrated in Fig. 6.15b. The negative r_p is produced by secondary emission from the plate when the plate voltage is less than the screen voltage (see Sec. 2.5a).

6.3b. Transitron Oscillator. The transitron oscillator is another type of negative-resistance oscillator in which the negative resistance is produced between the screen and the cathode of a pentode. The circuit of a transitron oscillator is shown in Fig.

6.16a. The tube operating region is indicated in Fig. 6.16b. By coupling the suppressor to the screen through a capacitor having a small reactance at the resonant frequency of the screen circuit, the suppressor voltage is made to follow the a-c variations in screen voltage. When the suppressor voltage is increased from its quiescent negative value, the plate current will increase, and when the suppressor is made more negative, the plate current will decrease. By selection of the proper operating potentials for the tube, increases in the screen voltage cause the screen current to decrease, producing a negative dynamic screen resistance r_{sg}. This will occur when the increasing suppressor voltage causes more electrons to flow to the plate than the increasing screen voltage has attracted from the cathode, resulting in a net decrease in screen current. Also, as the screen and suppressor voltages simultaneously decrease, the negative-going suppressor voltage decreases the plate current more than the decreasing screen voltage reduces the cathode current, resulting in a net increase in screen current. The negative dynamic screen-grid resistance r_{sg} produces a negative resistance across the screen resonant circuit causing oscillation. The suppressor conduction develops grid-leak bias across R_c which stabilizes the amplitude of oscillation by automatically adjusting the d-c suppressor bias as the amplitude of oscillation varies.
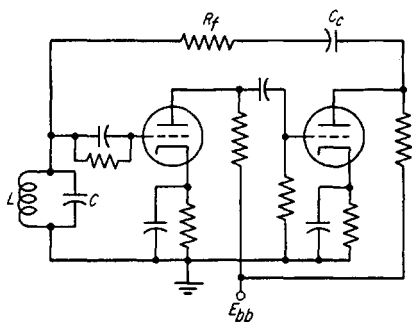
FIG. 6.17. Feedback oscillator.

By placing another parallel-resonant circuit in the plate circuit of the transitron oscillator, an electron-coupled transitron oscillator is formed, and by coupling the output signal from the plate circuit, good isolation between the load and the screen resonant circuit can be obtained. When the electron-coupled circuit is used, neutralization of the suppressor-plate capacitance is necessary to obtain complete isolation of the screen circuit from the load.

6.3c. *Feedback Oscillator.* A third type of negative resistance oscillator is the feedback oscillator. This type of oscillator consists of a two-stage RC amplifier with both input and output connected across a parallel resonant circuit as shown in Fig. 6.17. The positive feedback thus obtained effectively produces a negative resistance across the resonant circuit as long as the open loop gain exceeds unity at the frequency at which the total phase shift is 360°. As the amplitude of the oscillation increases, the nonlinearities of the amplifier reduce the effective loop gain and stabilize the amplitude of the sinusoid. The waveshape and frequency stability of such an oscillator will be improved if some form of automatic gain control is used to limit the amplitude of oscillation to a small value.

6.4. Crystal Oscillators. The frequency stability of an oscillator is dependent upon the Q or sharpness of resonance of its tuned circuit and the temperature stability of the tuned-circuit components. Changes in phase shift in the oscillator feedback path due to variations in tube plate resistance and input capacitance with supply voltage changes cause the frequency of oscillation to shift until the reactance of the resonant circuit provides exactly 360° phase shift through the amplifier and the feedback loop. The higher the resonant circuit Q the greater the change in circuit phase shift per unit of frequency change and, consequently, the smaller the frequency shift for a given initial disturbance.

The Q of conventional lumped element circuits is limited by the achievable Q in the inductances. Values of Q from approximately 100 to 300 are the maximum that can be obtained in practice. In addition, the variations in inductance and capacitance

with temperature will cause appreciable variations in frequency unless the components are carefully selected to provide necessary compensation.

Because of their high Q and high temperature stability, certain electromechanical resonators have received widespread use as resonant circuits in oscillators where frequency stability is important. Foremost among the electromechanical resonators is the quartz crystal.

Certain materials, including crystalline quartz, exhibit the property of exchanging energy between electrical and mechanical states; i.e., a mechanical force applied in the proper direction upon the material will cause an electrical charge to appear on the surfaces of the material, and conversely, an applied electrical potential will cause a mechanical displacement of the material. This is known as the *piezoelectric effect*. An alternating voltage of the correct frequency applied across the proper sides of a quartz crystal will cause it to vibrate mechanically. The mechanical vibration exhibits a resonance at a frequency which is determined by the dimensions of the quartz crystal. At this resonant frequency the exchange of energy between the electrical and mechanical states is extremely efficient and very little
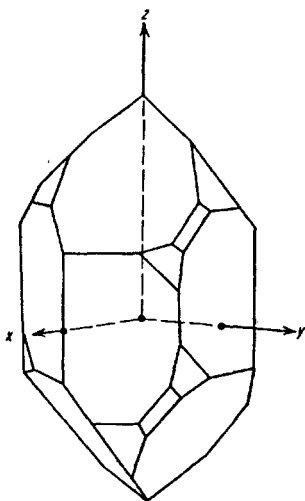
Fig. 6.18. Natural form of crystalline quartz, right-hand form.
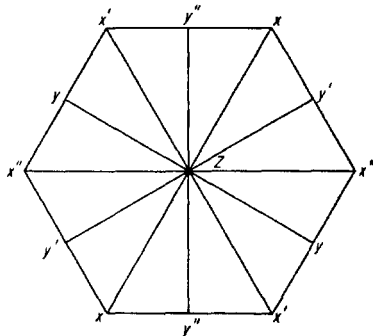
Fig. 6.19. The x, y, and z axes of a natural quartz crystal.

energy is dissipated in the crystal. The frequency of this resonance is very sharply defined, and quartz crystals have equivalent Q's of 10,000 to 100,000. Also, the physical and electrical properties of quartz have small temperature coefficients of variation, thus making quartz crystals ideal frequency-stabilizing elements for oscillator circuits.

6.4a. Crystalline Quartz. Crystalline quartz in its natural form appears as shown in Fig. 6.18. This is the right-hand form. It also is found in the left-hand form, which is the mirror image of the crystal shown in Fig. 6.18. The crystal has a hexagonal shape about the z axis, thus forming three x and y axes about the z axis as shown in Fig. 6.19.

A crystal unit suitable for use in an oscillator consists of a crystalline quartz plate equipped with suitable electrodes and mounted in a holder in a manner that will allow it to vibrate freely in the desired mode of vibration. In preparing a crystal, the mother quartz is first sawed into wafers which are oriented with respect to the x, y, and z axes of the mother quartz so as to give the desired electrical properties (see Sec. 6.4c). The plane of the crystal blank is normal to the designated axis. The wafers are diced into blanks of the appropriate size, and reduced in thickness by stages of lapping until almost the proper dimension for the desired frequency is obtained. The blanks are then carefully cleaned and brought to the final dimension by etching

for pressure-mounted crystals. For metal-plated, wire-mounted crystals, the etching process is stopped when the proper preplating thickness is reached. After etching, the blanks are thoroughly cleaned and sometimes subjected to heat cycling. The etching and heating processes minimize frequency variations with age. The blanks to be metal-plated are then base-plated and mounted in clean, moisture-free hermetically sealed holders. After mounting, the plates are adjusted to the precise final frequency by additional plating.

6.4b. Quartz-crystal Characteristics. The equivalent electrical circuit for a quartz crystal is shown in Fig. 6.20. The series L, R, and C represent the motional impedance of the quartz crystal. C_o is the parallel combination of C_e formed by the electrodes with the quartz as the dielectric and the shunt capacitance C_h of the crystal holder. The manner in which the reactance and resistance of a quartz crystal vary with frequency is illustrated in Fig. 6.21. A quartz crystal contains numerous resonances due to the various modes of mechanical motion which may be set up by an electrical driving force. However, the unwanted responses normally are sufficiently far removed from the fundamental response that they can be neglected, and the equivalent circuit of Fig. 6.20 holds.

The crystal exhibits two resonant frequencies: one where the reactance of the series arm is zero, and the other where the reactance of the series arm is equal in magnitude and of opposite sign to the reactance of C_o. The first resonant frequency is known as the series-resonant frequency f_r, where

$$f_r = \frac{1}{2\pi \sqrt{L_1 C_1}} \qquad (6.29)$$

The second resonant frequency is known as the parallel, or antiresonant, frequency f_a and is given by

$$f_a = \frac{1}{2\pi \sqrt{L_1 \dfrac{C_1 C_o}{C_1 + C_o}}} \qquad (6.30)$$

Fig. 6.20. Equivalent circuit of a quartz crystal.

Either the series-resonance or the parallel-resonance characteristic of the quartz crystal can be used to determine the oscillator frequency (see Sec. 6.4d). In the parallel-resonance case, the crystal is used as an inductance and the region of operation is restricted to the range of frequencies between the series and parallel-resonant frequencies of the crystal. The impedance seen by the crystal must be a capacitive reactance and a negative resistance for oscillation to take place. In the series-resonance case, the crystal is used as a series element in the oscillator feedback loop, and the frequency of oscillation is very close to f_r. The crystal appears as a pure resistance of low value at the frequency of oscillation provided that the crystal load impedance is a pure resistance at that frequency. Oscillation will not occur at or near f_a because the crystal resistance is very high near this frequency and, although the phase shift of the feedback loop is correct, the attenuation will be too great to allow oscillation.

The Q of a quartz crystal is defined as the Q of the series arm.

$$Q = \frac{\omega_r L_1}{R_1} = \frac{1}{\omega_r C_1 R_1} \qquad (6.31)$$

The impedance Z of the crystal as a function of frequency is given by

$$Z = R_s + jX_s$$

$$Z = \frac{\left(\frac{\omega_a}{\omega}\right) Q \left(A - \frac{1}{A}\right) - j\left\{1 + Q^2\left[1 - \left(\frac{\omega_a}{\omega}\right)^2\right]\left[A^2\left(\frac{\omega}{\omega_a}\right)^2 - 1\right]\right\}}{\omega C_o\left[1 + Q^2A^2\left(\frac{\omega_a}{\omega} - \frac{\omega}{\omega_a}\right)^2\right]} \quad (6.32)$$

where $A = \omega_a/\omega_r$
  R_s = equivalent crystal series resistance
  X_s = reactance of crystal
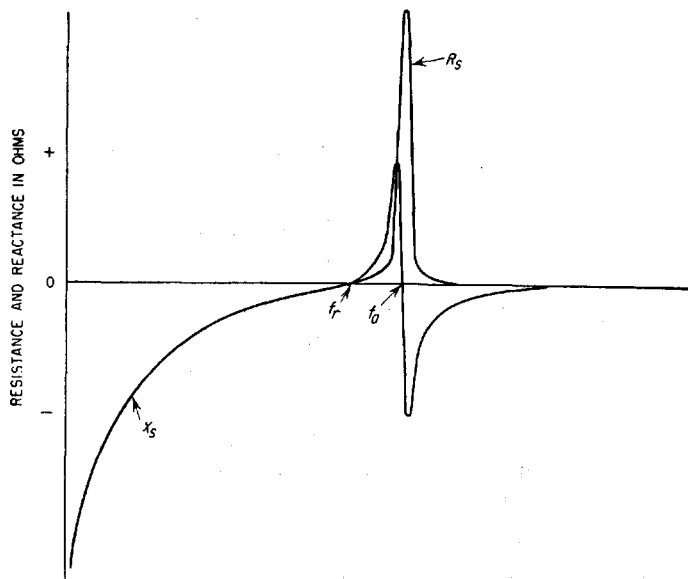
FIG. 6.21. Reactive and resistive components of crystal impedance as a function of frequency.

If the resistance R_1 is assumed to be zero, Eq. (6.32) reduces to

$$Z = \frac{-j\left(\frac{\omega_r^2}{\omega^2} - 1\right)}{\omega C_o\left(\frac{\omega_a^2}{\omega^2} - 1\right)} \quad (6.33)$$

The degree to which the mechanical resonant circuit consisting of L_1, R_1, and C_1 in series is coupled to the electrical circuit shunting C_o is known as the electromechanical coupling r_o of the crystal unit. It is defined as

$$r_o = \frac{C_o}{C_1} \quad (6.34)$$

The minimum electromechanical coupling r_e exists when the shunt capacitance C_o is a minimum. The smallest possible value of C_o is C_e, the shunt capacity of the electrodes and quartz dielectric only, excluding all holder and external circuit shunt

capacitance. For quartz crystals with plated electrodes C_e is given approximately by[1]

$$C_e = 0.402 \frac{A}{t} \quad \mu\mu\text{f} \tag{6.35}$$

where A = effective electrode area, cm^2

t = thickness of quartz plate, cm

The electromechanical coupling is increased by additional circuit shunt capacity. Varying the electromechanical coupling r_o by changing C_o has two effects upon crystal oscillator performance. First, it has a slight effect upon oscillator tuning as can be seen from Eq. (6.36).

$$f_a = f_r \sqrt{1 + \frac{\kappa-1}{r_o}} \tag{6.36}$$
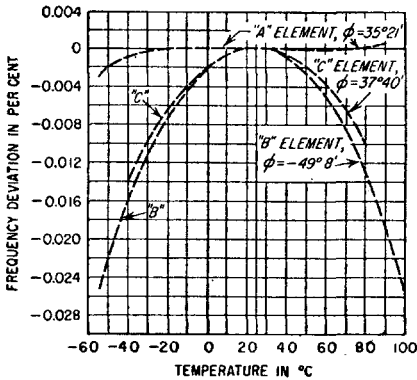
FIG. 6.22. Frequency versus temperature characteristic for several typical quartz crystal cuts.

The antiresonant frequency f_a will have a maximum value when $r_o = r_e$ and will decrease approaching f_r as a limit as r_o is increased. The change in the antiresonant frequency f_a with changes in C_o can be used to vary slightly the frequency of oscillation of a parallel-resonant-type crystal oscillator. A minimum value of r_e for any type of crystal is about 100, providing a maximum separation between f_a and f_r of approximately 0.5 per cent of f_r. Increasing C_o can lower the frequency of oscillation only a fraction of this separation before circuit losses due to the reduction in R_s stop oscillation.

The second effect of increasing r_o is to decrease the effective shunt resistance across the crystal electrodes at the parallel-resonant frequency f_a. When $f = f_a$ the resistive component of the crystal impedance is given by

$$R_s = \frac{L_1}{C_1 R_1 r_o^2 \left(1 + \frac{1}{r_o}\right)} \simeq \frac{L_1}{C_1 R_1 r_o^2} \tag{6.37}$$

Decreasing R_s increases the current through the crystal, decreasing the frequency stability of the crystal.

Typical values of Q for quartz crystals ranges from 10,000 to 100,000 depending upon the type of crystal cut. In special cases it may be considerably greater than 100,000 (see Sec. 6.4c).

Variations in the temperature of quartz crystals cause slight changes in crystal dimensions, elasticity, and density, which result in small variations in the crystal resonant frequency. The temperature stability of quartz crystals is measured in cycles per megacycle per degree centigrade variation in temperature, and the stability varies with the cut of the crystal. The frequency may either increase or decrease with an increase in temperature for a particular crystal cut, depending upon the actual temperature. This is illustrated in Fig. 6.22 where the temperature coefficient of frequency is given for several crystal cuts as a function of the ambient temperature.[1]

[1] From information received from R. A. Sykes, Bell Telephone Laboratory, Whippany, N. J.

Crystals may be placed in special ovens where the temperature is carefully con-
trolled to provide extreme frequency stability or to afford an accurate means of varying
slightly the crystal frequency.

Quartz crystals are subject to spurious responses because of the excitation of
unwanted modes of crystal motion. The flexure, extensional, and shear motions
are all excited by elastic coupling within the crystal, and electrical excitation of the
crystal will result in a certain degree of unwanted mode motion. The suppression of
unwanted mode motion is accomplished by adjusting the length and width dimen-
sions of a crystal for a given thickness dimension so as to minimize the coupling to
the unwanted modes. The critical dimensions vary as a function of the desired mode

TABLE 6.1. DESIGNATION OF CRYSTAL CUTS

Common designation	Mode of mechanical motion (see Fig. 6.23)	Standard designation*
AT—cut............	Thickness shear	A element
BT—cut............	Thickness shear	B element
CT—cut............	Face shear	C element
DT—cut............	Face shear	D element
+5°X—cut..........	Extensional	E element
−18°X—cut.........	Extensional	F element
GT—cut............	Extensional	G element
+5°X—cut..........	l-w flexure	H element
Duplex.............	l-t flexure	J element
MT—cut............	Extensional	M element
NT—cut............	l-w flexure	N element

* Standard designation is the type number placed on crystal vibrator elements used by the armed
services as established by the Armed Services Electro Standards Agency.

of motion and the type of crystal cut. Much of this information is determined
empirically.

In quartz crystal types A and B (see Sec. 6.4c) it is common practice to make the
crystal surface have a slight convex contour for fundamental frequency operation
between 0.5 and 20 Mc to aid in the suppression of unwanted responses. The convex
curvature is approximately inversely proportional to the frequency.

Because of the extreme thinness (approximately 0.006 in. at 15 Mc) of quartz
crystals at high frequencies and their consequent fragility, the operation of quartz-
crystal oscillators at frequencies above 20 Mc is usually achieved by the use of an odd
mechanical overtone of the crystal fundamental resonant frequency. Partially
plated type A crystals are used, and operation on the 3rd, 5th, and 7th overtones can
be obtained at frequencies as high as 100 Mc. The crystal is placed in the oscillator
circuit exactly as it would be for fundamental operation at the desired frequency.
The variations in the crystal parameters as a function of the order of the overtone
used are shown in Table 6.3.

The low-frequency limit for using quartz crystals as stabilizing elements in oscilla-
tors is established by the availability of large-size quartz blanks suitable for the
purpose. Standard commercial crystal units are obtainable for frequencies as low
as 15 kc.

The frequency stability of quartz-crystal oscillators is dependent upon the amplitude
of mechanical vibration of the quartz. The higher the amplitude of vibration, the
poorer the frequency stability. The amplitude of oscillation is a function of the
alternating current through the crystal, which in turn is dependent upon the capaci-

tance ratio r_o and the voltage amplitude across the crystal. Frequency deviations of 0.02 per cent or more may occur because of excessive crystal current, and after operation at high current levels, the crystal may not return to its original frequency for a period of several hours or even days after the drive has been reduced.

TABLE 6.2. FREQUENCY OF SERIES RESONANCE AND FREQUENCY-TEMPERATURE STABILITY OF VARIOUS CRYSTAL CUTS

Standard designation	Frequency of series resonance f, Mc; t = crystal thickness, mm	Temperature stability, ppm/°C (typical values)
A Element......	$f \simeq 1.66\ (1/t)$	$\simeq -0.05$ for 35°21′ cut
		$+0.80$ for 35°13′ cut
B Element......	$f \simeq 2.56\ (1/t)$	$\simeq \pm 4.4$
C Element......	$f \simeq 3.07\ (1/l)\ (l = w)$	$\simeq \pm 3.5$
D Element......	$f \simeq 2.07\ (1/l)\ (l = w)$	$\simeq \pm 2.0$
E Element......	$f \simeq 2.70\ (1/l)$ for $w/l = 0.5$	$\simeq -7.0$
F Element.......	$f \simeq 2.54\ (1/l)$ for $0.3 < w/l < 0.7$	
G Element......	$f \simeq 3.37\ (1/l)$ for $w/l = 0.86$	
H Element......	$f \simeq 4.8\ (w/l \times 1/l)$ for $w/l = 0.19$	
J Element.......	$f \simeq 5.4\ (t/l \times 1/l)$ for $t/l = 0.04$	$\simeq -3.5$
N Element......	$f \simeq 5.3\ (w/l \times 1/l)$ for $w/l = 0.19$	$\simeq \pm 2.5$

NOTE: The frequency-temperature characteristic of each crystal type will approach zero at a particular temperature, usually designed to be +25°C, for a particular orientation of crystal cut. The values given are typical values away from the temperature of minimum variation. Deviations above the critical temperature are negative and below are positive.

6.4c. Common Quartz-crystal Cuts and Their Characteristics. Tables 6.1, 6.2, and 6.3 describe the various types of quartz crystals in common use and tabulate their useful characteristics.[1] The various modes of mechanical motion which can be excited in quartz crystals are illustrated in Fig. 6.23.

There are a number of different methods in common use for mounting quartz crystals. The type of mounting to be used for a particular crystal is, in general, specified by the mode of mechanical motion for which the crystal is designed. Several of the common mounting methods are illustrated in Fig. 6.24. The air-gap mounting shown in Fig. 6.24a is used in initial testing of crystals during the grinding process and for type A and B elements in the range of 20 to 150 kc. The pressure air-gap mounting shown in Fig. 6.24b is used in place of the simple air-gap mounting

(a) THICKNESS SHEAR   (b) FACE SHEAR

(c) EXTENSIONAL   (d) l-w FLEXURE

FIG. 6.23. Modes of mechanical motion in quartz crystals.

where the ratio of crystal length to thickness exceeds 20:1, permitting the corners to be clamped without interfering with crystal motion. The pressure mounting of Fig. 6.24c and the soldered-lead-type mounting of Fig. 6.24d are used to support crystals operating in the flexure, extensional, and face-shear modes. In both types of mounting the electrodes are plated directly on the crystal faces. The solder lead mounting method provides a simple and convenient method for tuning the mechanical resonance of the support leads by means of small solder weights placed at appropriate distances

[1] From information received from R. A. Sykes, Bell Telephone Laboratory, Whippany, N.J.

from the solder cones connecting the leads to the crystal electrodes. The mounting method shown in Fig. 6.24e is used for high-frequency thickness shear-mode crystals.

6.4d. Circuits for Oscillators Using Quartz Crystals. The oscillator circuits in which quartz crystals may be used can be classified into three types: those utilizing parallel-resonance crystal operation, those utilizing series-resonance crystal operation, and those using bridge circuits.

TABLE 6.3. EQUIVALENT-CIRCUIT PARAMETERS FOR SEVERAL CRYSTAL TYPES

Standard designation	R_1, ohms	L_1, henrys	C_1, μμf	C_e, μμf	Q_{min}
A element / Partially plated: 1–20 Mc fundamental freq., 20–100 Mc overtone	$\dfrac{16 \times 10^6 N^3}{QAf_r^2}$	$\dfrac{2.62N^3}{Af_r^3}$	$\dfrac{C_e}{250N^2}$	$\dfrac{2.42Af_r}{N}$	
	where A = electrode area, cm² N = order of mechanical overtone f_r = series resonant freq., Mc				25,000 for $N = 1$ 75,000 for $N > 1$
C element (Fully plated 300–1,000 kc)	$\dfrac{1.465 \times 10^6 tf_r}{Q}$	$233T$	$\dfrac{C_e}{350}$	$\dfrac{38 \times 10^3}{tf_r^2}$	
	where t = thickness, cm f_r = series resonant freq., kc				
D element (200–500 kc)	$\dfrac{3.7 \times 10^6 f_r t}{Q}$	$588t$	$\dfrac{C_e}{400}$	$\dfrac{17.2 \times 10^3}{tf_r^2}$	10,000
	where f_r = series resonant freq., kc t = thickness, cm				

Parallel-resonant Oscillator Circuits. The quartz crystal can be used in any of the previously described oscillator circuits where no tap is required on the inductance of the frequency-determining resonant circuit. The crystal replaces the actual inductance of the resonant circuit, and the frequency of oscillation will be that frequency at which the inductive crystal reactance equals the capacitive reactance presented across the crystal terminals by the external circuit. If the crystal is assumed to have infinite Q, that is, $R_1 = 0$, and there is no external circuit loading, the actual frequency of oscillation f_o can be related to the series-resonant frequency f_r of the crystal in the following manner:

$$f_{osc} = f_r \sqrt{1 + \frac{C_1}{C_o + C_t}} \qquad (6.38)$$

where C_1 = capacitance of equivalent crystal series resonant arm
C_o = shunt capacitance of crystal and holder
C_t = capacitance of external circuit across which crystal is connected
Although the frequency variations which can be achieved by varying C_t are quite

small, such variations may amount to several parts per million and, therefore, cannot be neglected when utmost frequency accuracy is desired. For this reason, the capacitance of the oscillator circuit in which the crystal is to be used must be specified at the time of manufacture of the crystal, or the circuit capacitance must be variable and the precise frequency of oscillation adjusted by comparison with a frequency standard. Values of 20 and 32 $\mu\mu f$ have been adopted as standard values of oscillator circuit capacitance by many manufacturers.
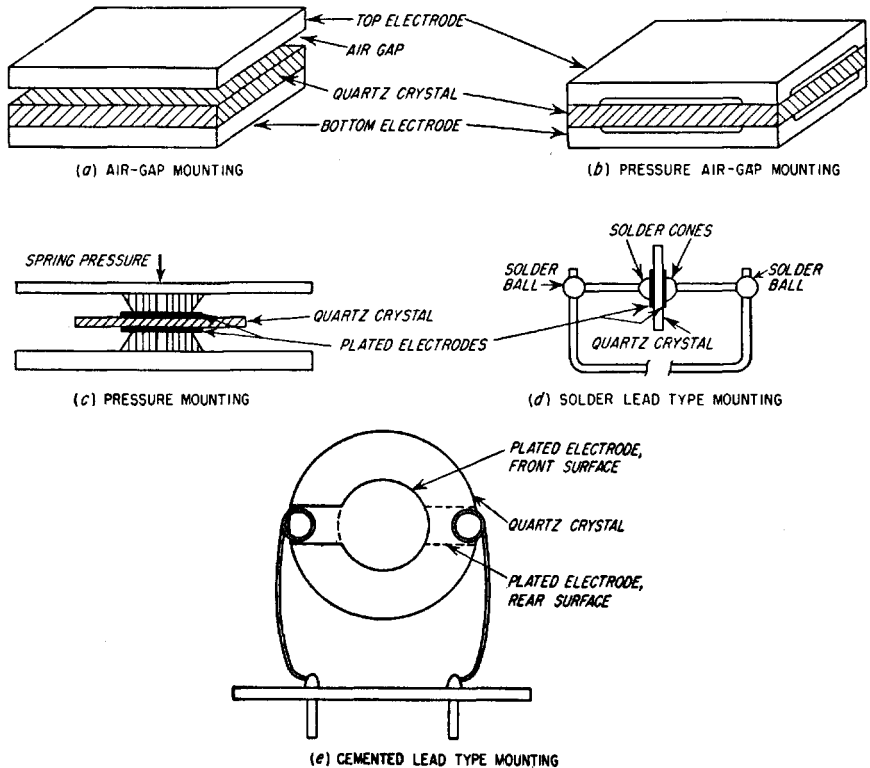
FIG. 6.24. Methods of mounting quartz crystals.

Several commonly used circuits of parallel-resonant crystal oscillators are shown in Fig. 6.25. The Pierce oscillator circuit of Fig. 6.25a is a Colpitts (see Sec. 6.2e) oscillator in which the inductance between grid and plate has been replaced by the crystal itself. If the circuit is redrawn with the crystal replaced by its equivalent circuit, the result is identical to the Clapp oscillator (see Sec. 6.2g). The Q of a quartz crystal is higher by a factor of 100 or more than the Q attainable in a physical inductance used in the Clapp or Colpitts circuits. In the modified Pierce circuit of Fig. 6.25b the plate has been placed at r-f ground instead of the cathode as in Fig. 6.25a. The crystal may replace the grid inductance of the tuned-plate tuned-grid (see Sec. 6.2c) oscillator as shown in Fig. 6.25c.

The circuits of Fig. 6.25a, b, and c may all be used in electron-coupled oscillator circuits using tetrode or pentode tubes in which the screen replaces the plate connections. The plate circuit of the electron-coupled oscillator contains a parallel resonant circuit tuned to the fundamental or to a harmonic of the crystal frequency. The "tri-tet" oscillator circuit of Fig. 6.25d is an electron-coupled modified tuned-plate

tuned-grid oscillator with the screen operating at r-f ground potential. This elimi-
nates the requirement of screen-plate neutralization when tetrode tubes are used
(see Sec. 6.2*f*).

Series-resonant Oscillator Circuits. In series-resonant crystal oscillator circuits
the crystal is connected as a series element in the feedback circuit. If the impedances
into which the crystal operates are pure resistances, oscillation will occur at the fre-
quency of zero crystal reactance and minimum crystal resistance (see Fig. 6.21) since
the loop gain will be a maximum at this frequency. This will be very close to the
series-resonant frequency f_r of the crystal. The frequency of oscillation will be
exactly f_r only if the phase shift in the remainder of the circuit is 360°.

(*a*) "PIERCE" OSCILLATOR   (*b*) MODIFIED "PIERCE" OSCILLATOR

(*c*) TUNED-PLATE TUNED-GRID OSCILLATOR   (*d*) "TRI-TET" OSCILLATOR
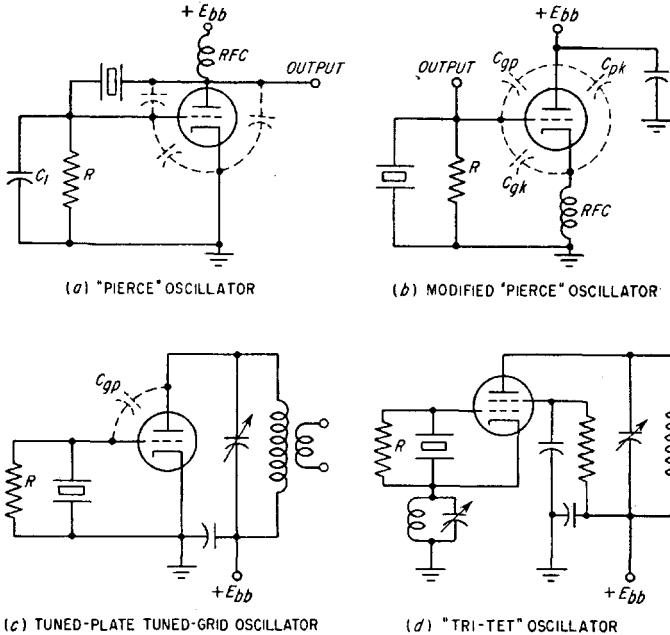
Fig. 6.25. Parallel-resonant crystal oscillator circuits.

Most series-resonant crystal circuits require more components than parallel-
resonant crystal circuits. They also require tuning of another resonant circuit and
usually do not permit grounding of either crystal electrode. Two series-resonant
crystal oscillator circuits are illustrated in Fig. 6.26. In both circuits the crystal
serves as the frequency selective feedback element between two tubes.

In the circuit shown in Fig. 6.26*b* the crystal can be operated on an odd mechani-
cal overtone and a harmonic of the overtone can be produced in the plate circuit of
the harmonic amplifier. Overtone operation is obtained by placing in parallel with the
crystal an inductance which is parallel resonant with C_o of the crystal at a frequency
slightly lower than the desired overtone frequency.

The terminating impedances for the crystal should be as low as the circuit gain
requirements will permit to achieve maximum frequency stability. Stray capaci-
tances across the crystal terminating resistances shift the frequency of oscillation
away from the series-resonant frequency f_r of the crystal and reduce the loop gain
and should, therefore, be kept to the minimum value possible.

Crystal Bridge Circuits. The greatest frequency stability of crystal oscillators is
achieved in bridge circuits where the crystal forms one arm of either a resistance

or a capacitance bridge. One circuit of each type is illustrated in Fig. 6.27. The Meacham bridge circuit of Fig. 6.27a is extremely stable and is commonly used in frequency standards below 1 Mc. Resistance bridge circuits are limited to relatively low frequencies because of the effects of stray capacities. The resistance bridge offers improved frequency stability because the amplitude of oscillation can be automatically stabilized at a low level by the use of a nonlinear power-sensitive resistive element in one arm of the bridge.

The capacitance bridge is useful at frequencies as high as 200 Mc. In capacitance bridge circuits the shunt capacitance C_o of the crystal is balanced out by the other bridge capacitances, and the bridge is unbalanced to provide a feedback signal of proper phase at a frequency very close to the series-resonant frequency f_r of the crystal.

(a) CRYSTAL OSCILLATOR WITH CRYSTAL AS A SERIES FEEDBACK ELEMENT

(b) HARMONIC OSCILLATOR USING CRYSTAL AS SERIES FEEDBACK ELEMENT
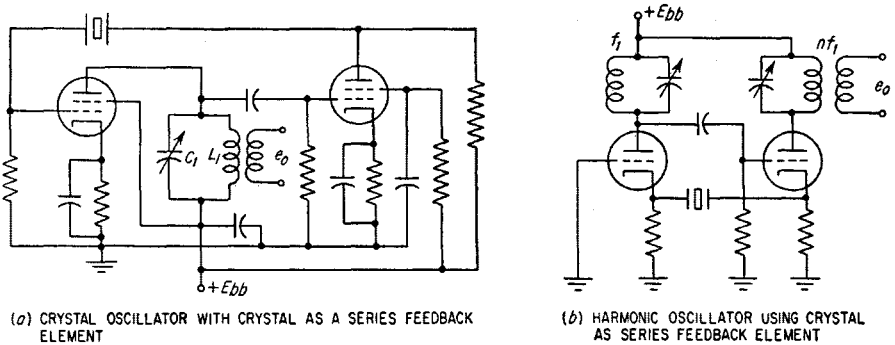
FIG. 6.26. Series-resonant crystal oscillator circuits.

6.4e. VHF Considerations. Because of the extreme thinness of crystals at high frequencies, they are usually operated on overtones at frequencies above 20 Mc.

In parallel-resonant circuits the crystal must present an inductive reactance at the frequency of oscillation, which requires a very high Q crystal for overtone operation. For this reason, parallel-resonant crystal oscillator circuits are used primarily at frequencies below 20 Mc for either fundamental or harmonic operation.

Series-resonant crystal oscillator circuits can be made to operate satisfactorily on overtone operation up to about 100 Mc. Operation is limited by shunt capacitances. Capacitance bridge circuits can be used up to 200 Mc on overtone operation.

6.5. UHF Oscillators. As the frequency of operation is increased above approximately 100 Mc, a number of high-frequency effects must be taken into consideration in the design of oscillators.

6.5a. Tube Limitations at High Frequencies. At frequencies above approximately 100 Mc, tubes become less efficient as oscillators or amplifiers because of a number of factors, the most important of which are outlined below:

1. Input Loading Due to Electron Transit Time. (See Sec. 7.4h.)

2. Input Loading Due to Cathode Lead Inductance. (See Sec. 7.4h.)

3. Interelectrode Capacitances. The minimum values of capacitance that the oscillator resonant circuit can operate with are the capacitances between the elements of the tube. At high frequencies, the necessary value of inductance becomes quite small and it becomes necessary to use a section of transmission line as the external resonant circuit rather than a lumped inductance to resonate with the tube and distributed capacitances (see Sec. 20.1).

4. Lead Inductances. In addition to cathode lead inductance which causes input loading, the inductances of the leads of the other electrodes within the tube form

minimum values of resonant circuit inductances when the external circuit connections are as direct as possible, and thus, together with the tube interelectrode capacitances, determine the maximum frequency of operation of the tube.

5. Distortion of Current Waveform within a Tube Due to Electron Transit Time. When the frequency of operation of a class C oscillator or amplifier is high enough
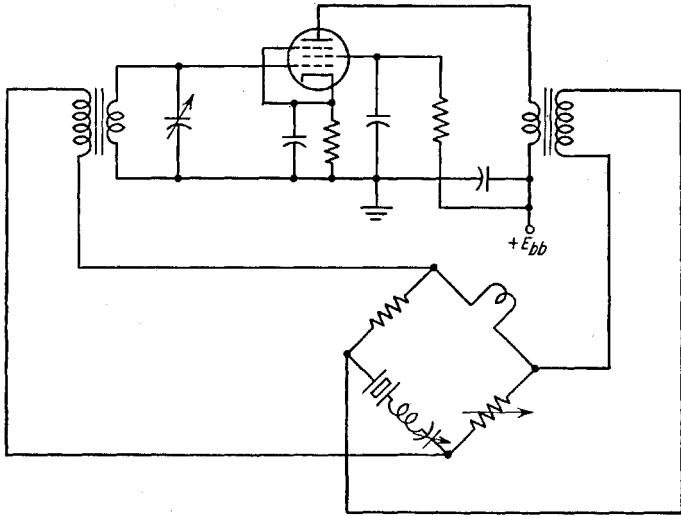
(*a*) MEACHAM RESISTANCE CRYSTAL BRIDGE
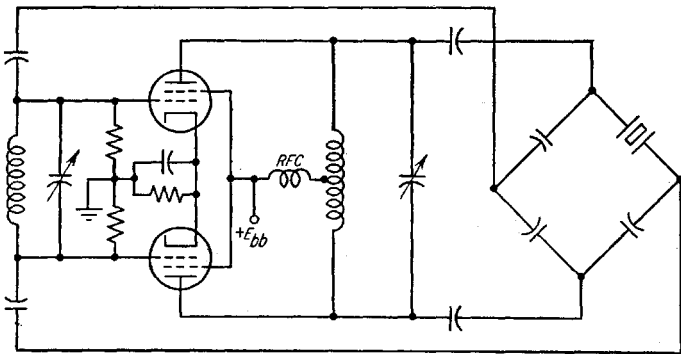
(*b*) BALANCED CAPACITIVE CRYSTAL BRIDGE

FIG. 6.27. Typical crystal bridge circuits.

that the time of transit of the electrons from cathode to plate is an appreciable fraction of one cycle of the operating frequency, the conventional pulse of current which flows at low frequencies (see Sec. 4.4*b*) becomes distorted because of the differences in the velocities of the electrons forming the current pulse due to the variation in plate voltage during the pulse. This distortion of the current pulse reduces the amplitude of the fundamental component of the current pulse and thereby reduces the power available in the plate circuit at the fundamental frequency.

6. *Increased Plate Dissipation Due to Electron Transit Time.* The effective genera-
tor voltage $-\mu e_{gk}$ in the plate circuit is shifted in phase more than 180° with respect
to e_{gk} by an amount

$$\beta = \omega_{osc}\tau \qquad \text{radians} \qquad\qquad (6.39)$$

where ω_{osc} = frequency of oscillation, radians/sec

τ = electron transit time from cathode to plate, seconds

At frequencies of about 100 Mc and higher, this additional phase shift may cause an
appreciable change in the plate current and plate dissipation of the oscillator tube.
Reference to Fig. 6.5 will show that a phase angle greater than 180° between e_{gk} and
$-\mu e_{gk}$ will have different effects on the Class I and Class II oscillators. In the Class
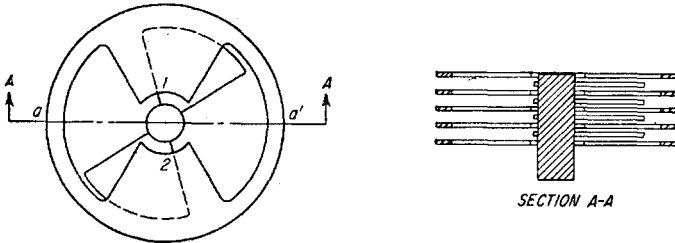
(*a*) BUTTERFLY RESONATOR
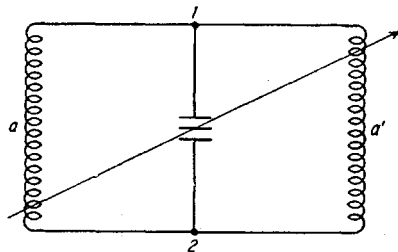
(*b*) EQUIVALENT CIRCUIT OF A BUTTERFLY RESONATOR

FIG. 6.28. Butterfly resonator.

I oscillator, the plate current and plate dissipation will be reduced by the additional
phase shift due to electron transit time. In the Class II oscillator, the plate current
and plate dissipation will increase with any additional phase shift in the tube due to
electron transit time. For this reason, Class I oscillators are preferred at high
frequencies.

7. *Increased Incidental Tube Losses.* At high frequencies skin effect causes the
depth of penetration of currents in the tube leads and elements to be reduced and I^2R
losses in these parts increase. Also, at high frequencies the tube insulators and glass
envelope become poorer dielectrics and the dielectric power losses increase.

The reduction of element spacing in tubes reduces electron transit time and extends
the maximum useful frequency range, but the smaller spacing makes tube cooling
more difficult and tends to limit the maximum plate dissipation. Alteration of the
electrode leads to bring them out separately to minimize interelectrode capacitance
and electrode configurations which bring the electrodes themselves through the enve-
lope of the tube to eliminate lead inductance altogether aid greatly in extending the
upper frequency limit of triode and tetrode vacuum tubes.

The Colpitts oscillator circuit (see Sec. 6.2e) is particularly suited to high-frequency use with conventional triode tubes since the effects of transit-time loading are less severe for this type of oscillator and because the cathode lead inductance is not a part of the resonant circuit, thereby reducing grid loading and eliminating the effects of cathode lead inductance on the feedback path.

6.5b. Butterfly Oscillators. At frequencies above 100 Mc the self-inductance and high loss due to current concentration at the terminals of conventional variable capacitors limits the usefulness of lumped *LC* resonant circuits. An alternate form of a variable resonant circuit is the butterfly resonator. It may be used in low-power oscillators at frequencies as high as 1,000 Mc and possesses the advantages over
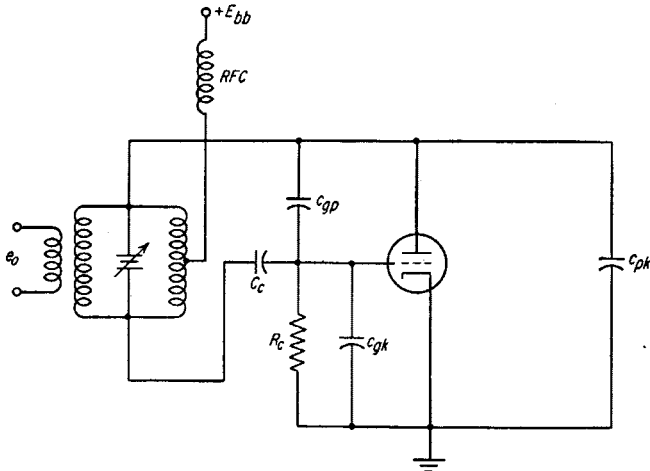
FIG. 6.29. Butterfly resonator in a Colpitts circuit.

lumped *LC* circuits of lower minimum inductance and lower losses at frequencies above about 100 Mc. The butterfly resonator is shown in Fig. 6.28a, and its equivalent circuit is shown in Fig. 6.28b. The two sets of rotor and stator plates form two variable capacitors in series which are in parallel with the two parallel inductances formed by the outer rings supporting the upper and lower stator plates. When the rotor plates are not fully meshed with the stator plates, they form eddy-current shields over parts of the two paralleled inductances and thereby lower the magnitude of the inductance. Tuning the rotor decreases capacitance and inductance simultaneously and vice versa. The maximum impedance of the resonator is across points 1 and 2, and when the rotor plates are either completely meshed with the stator plates or completely unmeshed, points *a* and *a'* are at r-f ground potential.[1] The circuit of a butterfly resonator in a Colpitts oscillator is shown in Fig. 6.29.

6.5c. Parallel-line Oscillators. Another type of oscillator frequently used between 100 and 1,000 Mc which is suitable for high-power oscillators is the parallel-line[2] oscillator. Typical circuits are shown in Fig. 6.30. In Fig. 6.30a the r-f short is achieved by a sliding r-f bypass capacitor. The shorted lines should have an electrical length something less than an odd multiple of a quarter wavelength at the frequency of operation, the λ/4 mode normally being used. The line is less than λ/4 long by an amount necessary to make the inductive reactance at the line terminals

[1] For greater detail in the design of butterfly resonators, see Radio Research Laboratory Staff, "Very High Frequency Techniques," vol. II, chaps. 28 and 30, McGraw-Hill Book Company, Inc., New York, 1947.
[2] See Sec. 20.2b.

equal to the capacitive reactance at the tube terminals. For a lossless line shorted at one end, the reactance at the open end of the line is given by

$$X = jZ_o \tan \frac{2\pi l}{\lambda} \tag{6.40}$$

where Z_o = characteristic impedance of line
  l = physical length of line
  λ = wavelength of operation

The characteristic impedances for several configurations of parallel lines with respect to a ground plane or chassis are given in Fig. 6.31. The choice of Z_o for

(a) GROUNDED-PLATE PARALLEL-LINE OSCILLATOR

(b) PUSH-PULL GROUNDED-PLATE PARALLEL-LINE OSCILLATOR

(c) COLPITTS OSCILLATOR WITH PARALLEL-LINE RESONANT CIRCUIT

(d) PUSH-PULL PARALLEL-LINE OSCILLATOR USING CATHODE CHOKES
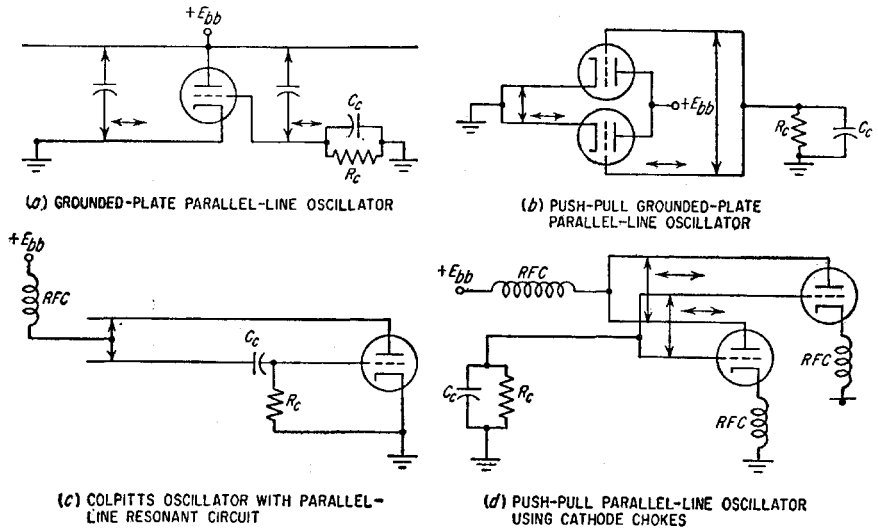
FIG. 6.30. Parallel-line oscillator circuits.

an oscillator is determined by two factors: mechanical layout and loaded Q of the oscillator. The size of tubes, pin spacing, and necessity of maintaining extremely short lead lengths usually restrict the spacing of the parallel lines. The diameter of the lines may be varied within the limits of mechanical construction, ruggedness, etc., to achieve the desired Q of the loaded line. For an unmodulated oscillator where good frequency stability is desired, the loaded Q should be as high as possible. The lower the characteristic impedance of the line, the higher the loaded Q and the longer the physical length of the line for a given reactance. In parallel-line oscillators Z_o is commonly between 100 and 1,000 ohms and the loaded Q is between 100 and 1,000.

The frequency range in which parallel-line oscillators are used is limited at low frequencies by the physical length of the lines and at high frequencies either by the electrode lead length within the tube envelope or by electron transit-time effects, depending upon which effect is more serious in a particular tube type.

Power may be coupled out of a parallel-line oscillator by a coupling loop or link which is parallel and adjacent to the oscillator line if the output line is balanced with respect to ground. A balanced to unbalanced transformer, such as the balun[1] must be used if a parallel-line oscillator is to feed an unbalanced transmission line. Direct

[1] See Sec. 20.4d.

coupling from a balanced transmission line to a point of proper impedance on the parallel line can also be used to couple power from the oscillator.

At high frequencies where the line separation d is an appreciable fraction of a wavelength, the parallel-line oscillator will radiate considerable power unless the lines are completely shielded. The shielding should have as low a surface resistivity as possible to keep the power loss low and the loaded Q as high as possible.

6.5d. Coaxial-line Oscillators. At very high frequencies the inductances of electrode leads within the tube are minimized by the construction of special tube types in which the electrodes are brought through the glass envelope for attachment to the external circuit. The lighthouse and pencil-triode tubes are typical. The coaxial-line resonant circuit lends itself to use with this type of tube structure because the tube elements can most easily be constructed to form a uniform continuation of an external coaxial line right to the center of the tube.

The circuit of a grid-separation or grounded-grid coaxial-line oscillator is shown in Fig. 6.32. This circuit is most frequently used in coaxial-line oscillators because the connection to the grid must necessarily come between the plate and cathode connections. Feedback from plate to cathode is achieved by capacitive coupling between the two resonant lines. For oscillation to take place the plate-grid circuit must be tuned to resonance at a frequency slightly higher than the desired frequency of oscillation, and the grid-cathode circuit must be tuned to resonance at a frequency slightly lower than the desired frequency of oscillation. The circuit is that of a grounded-grid class I oscillator.

$$Z_0 \approx \frac{276}{\sqrt{\epsilon}} LOG_{10} \frac{2D}{d}$$

(a) PARALLEL LINE REMOTE FROM GROUND

$$Z_0 = \frac{276}{\sqrt{\epsilon}} LOG_{10} \left[\frac{2D}{d\sqrt{1+(D/2h)^2}} \right]$$

FOR $d \ll D,h$

(b) PARALLEL LINE NEAR GROUND

$$Z_0 = \frac{276}{\sqrt{\epsilon}} LOG_{10} \left(\frac{4h \; TANh \; \frac{\pi D}{2h}}{\pi d} \right)$$

FOR $d \ll D,h$

(c) PARALLEL LINE BETWEEN PARALLEL GROUND PLANES

$$Z_0 = \frac{276}{\sqrt{\epsilon}} \left\{ LOG_{10} \left[\frac{4h \; TANh \; \frac{\pi D}{2h}}{\pi D} \right] - \sum_{m=1}^{\infty} LOG_{10} \left[\frac{1+U_m^2}{1-V_m^2} \right] \right\}$$

WHERE $U_m = \dfrac{SINh \frac{\pi D}{2h}}{COSh \frac{m\pi W}{2h}}$ $V_m = \dfrac{SINh \frac{\pi D}{2h}}{SINh \frac{m\pi W}{2h}}$

(d) TOTALLY ENCLOSED PARALLEL LINE

FIG. 6.31. Characteristic impedances of parallel-wire transmission lines.

At frequencies above 1,500 to 2,000 Mc a modified version of the grid-separation circuit known as the reentrant oscillator circuit is widely used. An oscillator of this type is shown in Fig. 6.33. The length of the grid cylinder and the position of the plunger are adjusted so that a portion of the grid-plate voltage is fed back to the grid-cathode terminals with a 180° phase shift.

The lower frequency limit of coaxial-line oscillators is restricted by the physical length of the line. The high-frequency limit is determined by electron transit time and other uhf effects (see Sec. 6.5a). Several tubes of the lighthouse and pencil-triode types may be used in coaxial-line oscillators at frequencies as high as 3,300 Mc.

The same considerations which applied in determining the characteristic impedance of the parallel line also apply for the coaxial line. The characteristic impedance

of a coaxial line, assuming a lossless line, is given by

$$Z_o = \frac{60}{\sqrt{\epsilon}} \ln \frac{d_2}{d_1} \quad \text{or} \quad \frac{138}{\sqrt{\epsilon}} \log_{10} \frac{d_2}{d_1} \qquad (6.41)$$

where ϵ = dielectric constant of line

d_1 = outside diameter of inner line

d_2 = inside diameter of outer line

Because of the tube structure, the ratio d_2/d_1 is not very large, e.g., $1.5:1$ to $4:1$, and
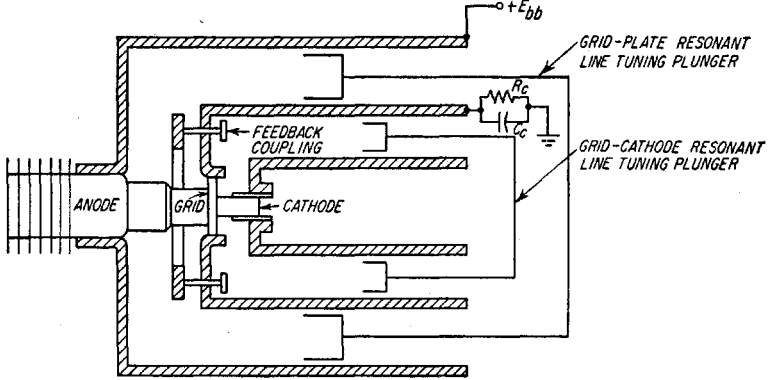
(a) CROSS-SECTION OF TYPICAL GRID-SEPARATION OSCILLATOR
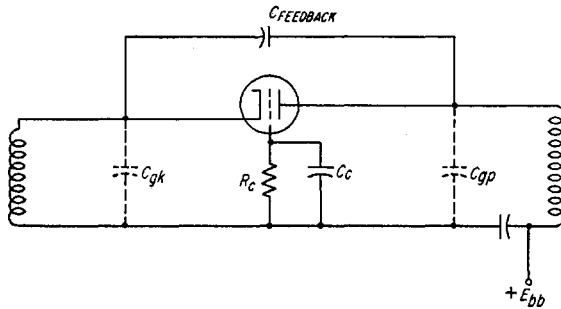
(b) CIRCUIT OF GRID-SEPARATION OSCILLATOR

FIG. 6.32. Typical grid-separation coaxial-line oscillator.

the characteristic impedances of typical coaxial lines in uhf oscillator circuits are between 20 and 90 ohms.

The physical length of shorted line required to have an inductive reactance at the open end which is equal to the capacitive reactance of the tube from grid to plate or grid to cathode can be determined from Eq. (6.40).

As in the case of parallel lines, the lower the characteristic impedance of a coaxial line, the higher the loaded Q and the longer the physical length of line required to present a given reactance at the line terminals.

Power may be coupled out of a coaxial-line oscillator by locating a probe (capacitive coupling) at a point of high r-f potential, i.e., near the tube end of a $\lambda/4$ line, or by locating a loop (inductive coupling) near a current maximum in the coaxial line, i.e.,

near the shorting plunger in a $\lambda/4$ line. Examples of loop coupling and probe coupling are shown in Fig. 6.34.

6.6. Microwave Oscillators. At frequencies where the operation of conventional space-charge triodes and tetrodes is limited by lead inductance, electron transit time, and other uhf effects, power can be generated by other types of oscillators which utilize electron transit time in their operation. The resonant circuits of these power generators are cavities which are either integral to the tubes themselves or surround the tube in such a manner that the internal portions of the tube effectively form part of the resonant circuit.

6.6a. Reflex Klystron Oscillators. The reflex klystron oscillator is a relatively low power oscillator which is generally used as the local oscillator in microwave receivers and as a low-power signal source. At present, these oscillators find application in the frequency range from approximately 500 to 30,000 Mc. Most reflex klystrons have

Fig. 6.33. Re-entrant oscillator circuit.

power outputs from 10 mw to 0.5 watt, although a few types are available which have power outputs as high as 10 watts.

The circuit of a reflex klystron is shown in Fig. 6.35. The resonator grids and associated cavity are at a positive potential with respect to the cathode. The electrons emitted from the cathode which pass through the control grid are accelerated toward the resonator grids. The electrons which pass through the resonator grids travel on toward the repeller. However, because the repeller is at a potential more

(a) CAPACITIVE-PROBE COUPLING   (b) INDUCTIVE-LOOP COUPLING

Fig. 6.34. Methods of coupling power from a coaxial-line oscillator.

negative than the cathode, the electrons do not reach the repeller. They are decelerated by the negative potential gradient between the last resonator grid and the repeller, and their forward motion is finally stopped and reversed. The electrons then travel back toward the resonator grids. The time required for the electrons to travel from the region between the resonator grids toward the repeller and return is a function of the various electrode voltages and the distance between the repeller and resonator grids. If an oscillation exists at the resonant frequency of the cavity, an r-f voltage e_c will exist between the two resonator grids. For the period that the electrons are between the two grids as they travel from the cathode toward the repeller, they will be accelerated or decelerated by the r-f voltage e_c by an amount depending upon the phase of e_c during that period. The period the electrons are between the two grids and subject to the accelerating or decelerating effects of the r-f voltage is

small compared to one cycle of the r-f voltage. Electrons which are decelerated by the r-f voltage will return to the resonator grid region after a shorter interval than those which pass through the grid region when the r-f voltage is zero and whose velocities are, therefore, unaffected by the r-f field. Those electrons which are accelerated by the r-f field will require a longer time before they return to the resonator grid space. The electrons ar· riving at the resonator grids from the cathode are approximately evenly distributed in time. However, those returning to the grids from the repeller region will be bunched together since the electrons which were accelerated by the grids return later than normal while the electrons which were decelerated return sooner than normal. This effect is illustrated in Fig. 6.36.

Since the distribution of electrons from the cathode is nearly constant over an r-f cycle, the average energy delivered to the electrons by the r-f field during their initial travel through the grid region is zero. However, if more electrons return to the resonator grid space during a positive half cycle of the r-f voltage than during a
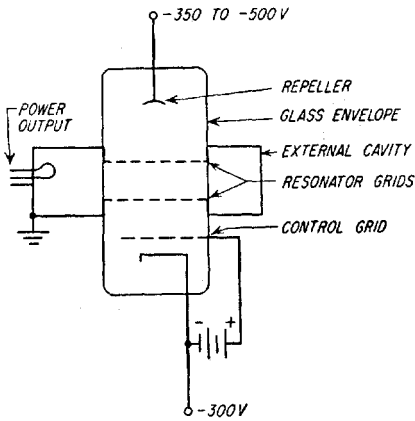
FIG. 6.35. Typical reflex klystron oscillator.
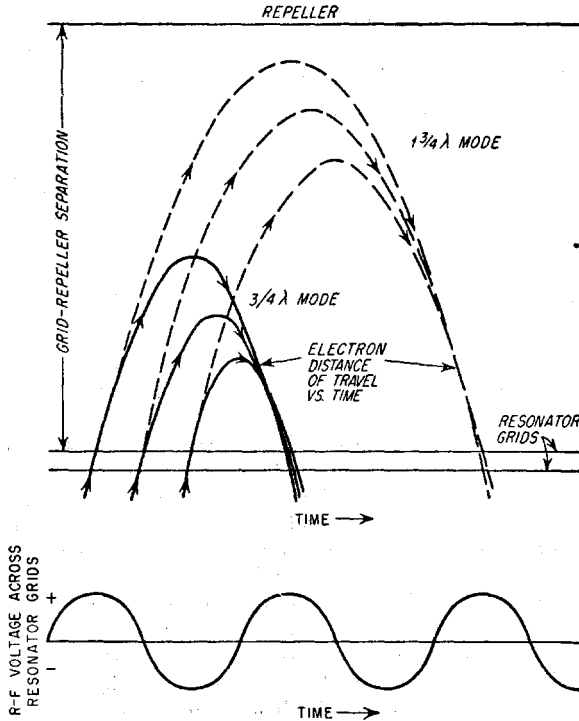
FIG. 6.36. Electron bunching in a reflex klystron oscillator.

negative half cycle, more returning electrons will be decelerated than will be accelerated and the electron beam will deliver energy to the r-f field. This transfer of energy from the electron beam to the r-f field will reinforce the original oscillation in the cavity, and oscillation will be sustained. The condition necessary for oscillation is that the repeller voltage be adjusted to such a value that a majority of the bunched electrons return during the positive half cycles of the r-f voltage.

Instead of the bunched electrons being returned on the next positive half cycle of the r-f voltage as indicated by the solid lines in Fig. 6.36, the repeller voltage may be made less negative so that the bunches form more slowly and the electrons require a longer time interval before they return to the grids. The bunches may thus be returned on later positive half cycles, sustaining oscillations at the same frequency as before. These conditions are known as the *voltage modes* of oscillation of the klystron. The lowest mode, corresponding to the shortest bunching time, is known as the ¾-cycle mode; the next mode is the 1¾-cycle mode, etc. Three modes of operation for a typical reflex klystron oscillator are shown in Fig. 6.37.
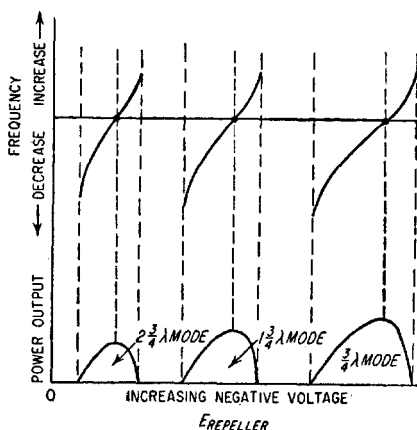
Fig. 6.37. Power output and frequency versus repeller voltage for a typical reflex klystron oscillator.

The frequency of oscillation can be varied over a small region by changing the repeller voltage and, thereby, changing the bunching time slightly. However, by returning the electrons before or after the optimum time, the bunches will be less perfectly formed and less energy will be delivered to the r-f field. This means less power output, as indicated in Fig. 6.37, and a lower magnitude of r-f voltage e_c across the resonator grids, which in turn means that the bunches will require more time to form. The bunches will be more poorly formed at the instant they arrive at the grid region if the repeller voltage is made slightly more negative than the optimum value for a particular cavity setting than if the repeller voltage had been made less negative by the same amount. This is because the electrons are returned sooner by the more negative repeller voltage, allowing less time in the drift space for bunching. If the repeller voltage is made less negative, the electrons spend a longer time in the drift space, partially compensating for the effects of reduced e_c in forming bunches. Thus, power output falls off more rapidly when frequency is increased than when it is decreased from the optimum value. The range of control is normally taken as the frequency variation that can be achieved before the power output has dropped more than 3 db. The rate of change of frequency with repeller voltage is normally between 0.1 and 2 Mc/volt.

The power output is less on the less negative repeller voltage modes because a smaller magnitude of r-f voltage is necessary to optimize bunching for the longer bunching periods, and the r-f voltage will automatically adjust itself to that value which provides the best bunching at the frequency of oscillation. The electronic tuning rate, i.e., megacycles change in frequency per volt change in repeller voltage, increases in the lower repeller voltage modes of operation. This is due to the fact that the percentage change in frequency is roughly proportional to the percentage change in repeller-to-grid potential, and this potential is decreased in the higher-order modes.

Variations in the resonator grid potential have less of an effect upon frequency than variations in repeller voltage changes because a change in the resonator grid potential produces two effects upon the electron beam which tend to compensate. An increase in resonator potential will cause the electrons to undergo greater acceleration, tending to take a longer interval after passing through the grids before being returned. However, this voltage increase causes the repeller-resonator grid potential to be more negative than before, tending to return the electrons more quickly to the grids. The rate of change of frequency with resonator voltage is about one-third as high as with repeller voltage changes.
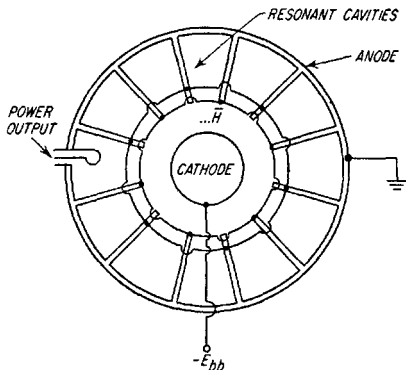
FIG. 6.38. A 12-cavity magnetron.

The electronic tuning of the klystron by variation of the repeller voltage is centered about the resonant frequency of the resonator-grid cavity, which can be tuned mechanically over a relatively wide frequency range.

6.6b. Magnetron Oscillators. The multiple-cavity electron-resonance magnetron is an oscillator which utilizes the transit time of electrons between tube elements as does the reflex klystron. Pulse-type magnetrons can provide peak power outputs ranging from about 5,000 kw at 10 cm to about 100 kw at 1 cm, with an average power of several hundred watts. Magnetron oscillators are used at frequencies as low as 600 Mc and as high as 35,000 Mc.

The end view of the cavity structure of a typical multicavity magnetron is shown in Fig. 6.38. It consists of a cylindrical diode with the cathode at the center and an anode composed of several resonant cavities around the outside. A constant magnetic field is applied in a direction parallel to the axis of the cylinder, and a d-c potential is applied between cathode and anode. The forces of the radial electrostatic field and the perpendicular magnetic field cause the electrons which are emitted from the cathode to travel cycloidal paths, the periods of which are determined by the magnetic-field intensity and the electric field gradient (see Sec. 2.1b). The direction of the cycloidal path is determined by the directions of the electric and magnetic fields existing in the region of the electron when it is emitted. The electric field is the resultant of the electrostatic field between cathode and anode and the r-f electric field across the cavity gaps. If the magnetron is in operation in the most fundamental (π) mode, the r-f fields across the cavity gaps at one instant of the r-f cycle will have the polarity and direction shown in Fig. 6.39. The mode of operation of a magnetron, i.e., the relative phases of the r-f field across adjacent cavity gaps, is established by strapping together anode segments having the same r-f potential. This is illustrated in Fig. 6.38 for a magnetron operated in the fundamental π mode (180° phase shift between adjacent cavity gaps). Depending upon the direction of the combined d-c and r-f electric fields at the region where the electrons are emitted, they will either spiral outward toward the anode, traveling in a cycloidal path, or they will spiral back to the cathode. The former case is shown in Fig. 6.40a. The electrons which travel toward the anode move at a constant average velocity v_e having a component in the direction opposite to the d-c electric-field and a component in the direction of the r-f electric-field. As a result, the average kinetic energy tends to be increased by E_{dc} and reduced by E_{rf}, thus providing a transformation of energy from the d-c field to the r-f field. At each cusp in the cycloidal path the electron kinetic energy is zero, and all of the energy imparted to the electron by the d-c field has been transferred to the r-f field. The electrons

emitted at point *a* in Fig. 6.40*b* with the r-f field as shown will progress as indicated by the solid line, transferring all of their energy to the r-f field at each cusp. After the electrons reach region *c* they will start to travel back to the cathode and absorb energy from the r-f field because of the reversed direction of the r-f electric field across the adjacent cavity gap unless the polarity of the r-f field is changed by 180°. The
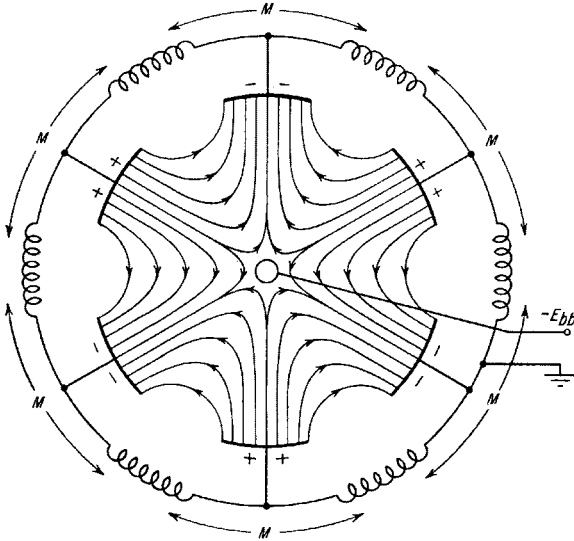
FIG. 6.39. Equivalent circuit of a multiple-cavity magnetron.

(*a*) MOTION OF ELECTRON HAVING ZERO INITIAL VELO-
CITY AT CATHODE SURFACE UNDER THE INFLUENCE
OF THE MAGNETIC AND ELECTRIC FIELD FORCES

(*b*) MOTION OF ELECTRONS EMITTED FROM MAGNETRON
CATHODE
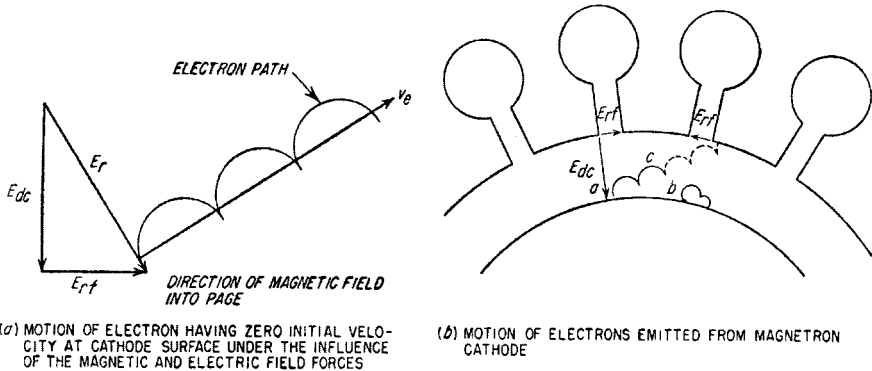
FIG. 6.40. Electron paths in a multiple-cavity magnetron.

r-f voltage across each cavity gap reverses polarity each half cycle, and if the frequency is correct, the polarity of the r-f voltage will be reversed as the electrons reach region *c* and they will continue on to the anode, transferring their energy to the r-f field. Proper adjustment of the magnetic field and d-c electric field will cause the electrons to progress laterally the width of one cavity each half cycle of the r-f voltage, and oscillation will be sustained in the cavities.

The kinetic energy of the electron when it strikes the anode is only that net energy it has received from the d-c field and not transferred to the r-f field between the last cusp of its cycloidal path and the anode. If the electron path contains several cusps

before reaching the anode, the energy of the electron when it strikes the anode will be small and the efficiency of energy transfer from the d-c field to the r-f field will be large. The electrons emitted from the cathode at the regions where the r-f field accelerates them and transfers energy to the electrons are returned to the cathode because the paths of these electrons are in a direction as shown at b in Fig. 6.40b. The average paths of these electrons will be much shorter than those of the electrons traveling to the anode, and consequently much more energy is transferred to the r-f
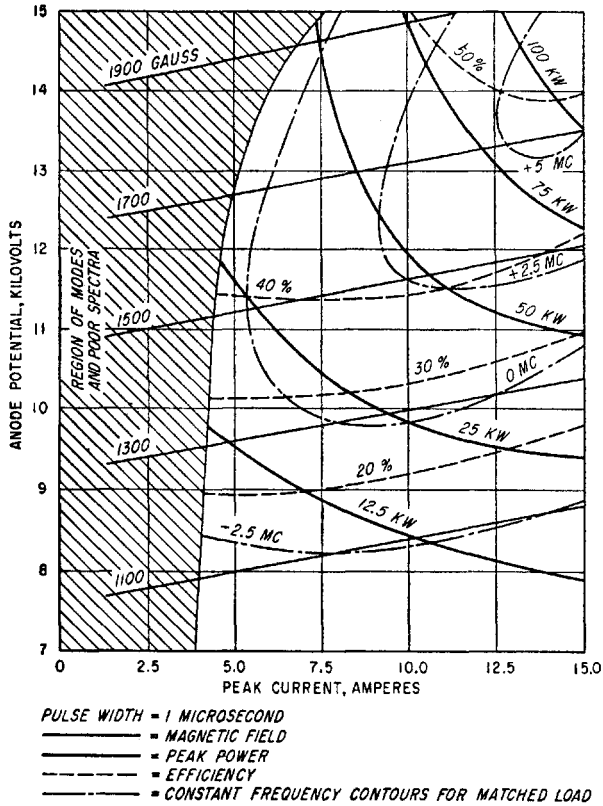
FIG. 6.41. Performance chart for a typical pulse magnetron.

field than is absorbed from it. The efficiency of high-power pulse magnetrons is in the order of 60 per cent at 1,000 Mc and 30 per cent at 30,000 Mc.

The electrons which are returned to the cathode strike it with a velocity sufficient to emit secondary electrons, and the kinetic energy of these electrons is high enough to increase considerably the temperature of the cathode surface. This effect is known as *cathode back-heating*. To maintain the cathode temperature at a satisfactory value, the filament power to magnetrons is reduced as the average magnetron power and the resulting back-heating power are increased. In certain pulse-type magnetrons, such as the 2J51, the external filament heating power is removed entirely when the tube is operating at full power.

The performance of magnetrons is specified by two plots of its characteristics: the performance chart and the Rieke diagram shown in Figs. 6.41 and 6.42, respectively. The performance chart relates applied voltage, magnetic-field strength, and current

to frequency, power output, and efficiency for a particular magnitude and phase of the magnetron load impedance. The Rieke diagram relates the power output and frequency of operation to the phase of the load impedance at the magnetron output connection and the standing wave ratio of the load. The *pushing* of a magnetron
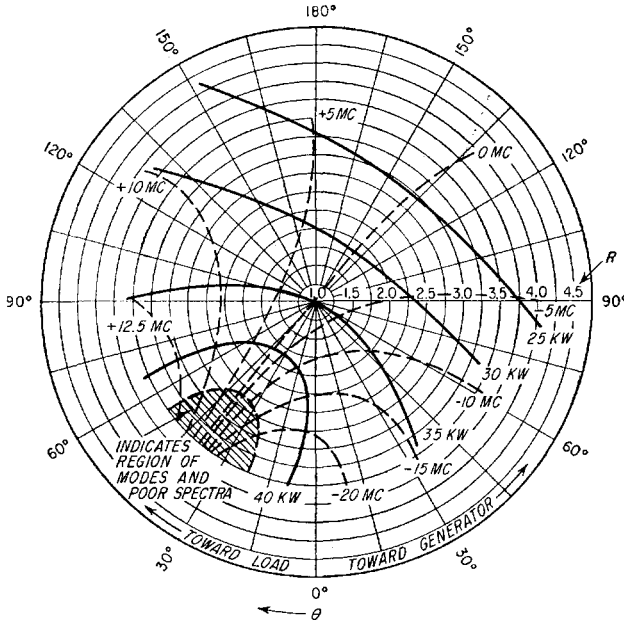
PRF = 2000 CPS
PULSE WIDTH = 1 MICROSECOND
MAGNETIC FIELD = 1300 GAUSS
PEAK PULSE CURRENT = 12.5 AMP

θ = DISTANCE OF VSWR MINIMUM FROM TUBE FLANGE
R = VSWR

Fig. 6.42. Rieke diagram of a typical pulse magnetron.

(a) TUNING OF MAGNETRON BY VARYING MAGNETIC FIELD IN CAVITIES

(b) TUNING OF MAGNETRON BY VARYING THE CAPACITY BETWEEN THE STRAPS CONNECTING ALTERNATE ANODE SEGMENTS OF A STRAPPED MAGNETRON
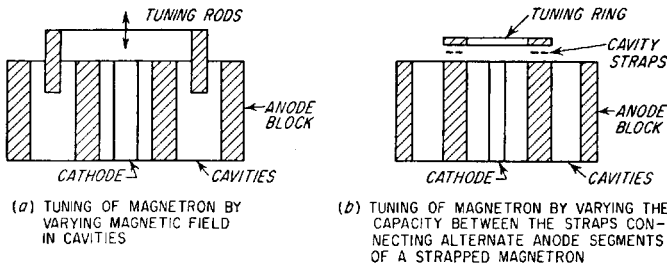
Fig. 6.43. Methods of tuning multicavity magnetrons.

is the change in frequency of oscillation for a specified variation in anode supply voltage for a fixed-load condition. It can be obtained from the performance chart of a magnetron. The *pulling* of a magnetron is the change in the frequency of oscillation for a specified change in the amplitude and phase of the standing wave at the magnetron output flange for a fixed supply condition. It can be obtained from the Rieke diagram.

Power is coupled out of magnetrons either by a loop inserted at a point of high r-f magnetic-field intensity within one of the cavities or by a slot located in the side wall of one cavity at a high current point.   The slot feeds a waveguide transmission line.

A magnetron may be tuned by varying the resonant frequency of the cavities in some manner.   Two methods of tuning are indicated in Fig. 6.43.

6.6c. Backward-wave Oscillators.   By transferring the energy of an electron beam to an electromagnetic field in a nonresonant structure it is possible to construct a microwave power oscillator which can be voltage-tuned over a wide frequency range. The exchange of energy between the electron beam and the field wave is conveniently achieved by causing the wave to propagate along a helix through the center of which the electron beam is passed as illustrated in Fig. 6.44.   As the wave propagates along the helix, the axial velocity of the wave is reduced to approximately that of the electron beam, and as a result of the continuous action of the electric field upon the electrons, the electrons in the beam are formed into bunches.   The bunched electron
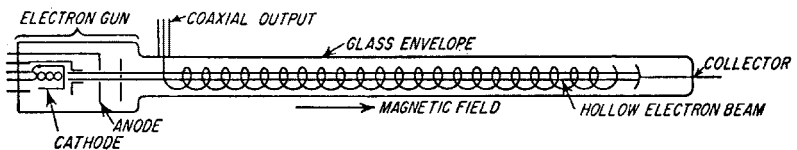
FIG. 6.44. Construction of a typical backward-wave oscillator.

beam in turn induces a field on the helix structure as it progresses down the helix. The bunching ideally increases as the square of the distance traveled by the electrons along the helix, and, consequently, the induced field increases as the square of the distance traveled by the electrons.   If the group velocity (see Sec. 20.1b) of the field wave is from the collector to the electron gun, so that power is transferred along the helix in this direction, the electron beam which is progressively bunched by the initially applied wave will induce a field wave propagating from the collector end of the helix toward the electron gun which, if the helix is of sufficient length, will be greater than the applied wave at the gun end.   This wave in turn increases the current wave of the succeeding electron bunches which in turn increase the induced field wave, etc. The result is an oscillation at the frequency where this condition of effective positive feedback between the electron beam and the field is satisfied.   This condition constitutes a wave having a phase velocity (see Sec. 20.1b) toward the collector and a group velocity toward the electron gun.

The tube will oscillate with minimum beam current at the frequency at which the successive induced field waves add with the minimum difference in phase at the gun end of the helix.   By varying the cathode-to-helix voltage, the velocity of the electron beam is changed and the wavelength of the induced wave on the helix which can progressively bunch the electrons traveling at this new velocity changes.   The helix is terminated in a matched load at the collector end, and power is coupled from the helix at the gun end.   The electron-beam current density and, consequently, power output are controlled by the anode-cathode potential in the gun structure.   Since only those electrons passing close to the helix interact with the wave propagated along the helix, the efficiency of a backward-wave oscillator is improved by employing a hollow electron beam.   The electron-beam path is made to approach the helix with a minimum of divergence by employing a strong axial magnetic field for focusing of the beam.

Tuning ranges of 2:1 are achievable in backward-wave oscillators having power outputs of approximately 1 watt or less.   A cathode-to-helix voltage change of approximately 8:1 is required to secure the 2:1 frequency change.   Power output

typically varies by a factor of 3:1 over a 2:1 frequency range, being highest at the high frequencies where the cathode-to-helix voltage and input power are highest.

Backward-wave oscillators can be frequency-modulated by varying the cathode-to-helix voltage and amplitude-modulated by varying the cathode-to-anode voltage. The minimum achievable pulse rise time is limited by the propagation of the wave along the helix. This is typically 0.05 to 0.1 μsec.

6.7. RC Oscillators. The condition necessary for oscillation, that is, $\mu B = 1$, can be obtained in vacuum tube circuits which contain only resistance and capacitance. RC oscillators are extensively used as oscillators in the frequency range of a few

(a) THREE-SECTION PHASE-LEAD PHASE-SHIFT OSCILLATOR

(c) THREE-SECTION PHASE-LAG PHASE-SHIFT OSCILLATOR

(b) FOUR-SECTION PHASE-LEAD PHASE-SHIFT OSCILLATOR

(d) FOUR-SECTION PHASE-LAG PHASE-SHIFT OSCILLATOR
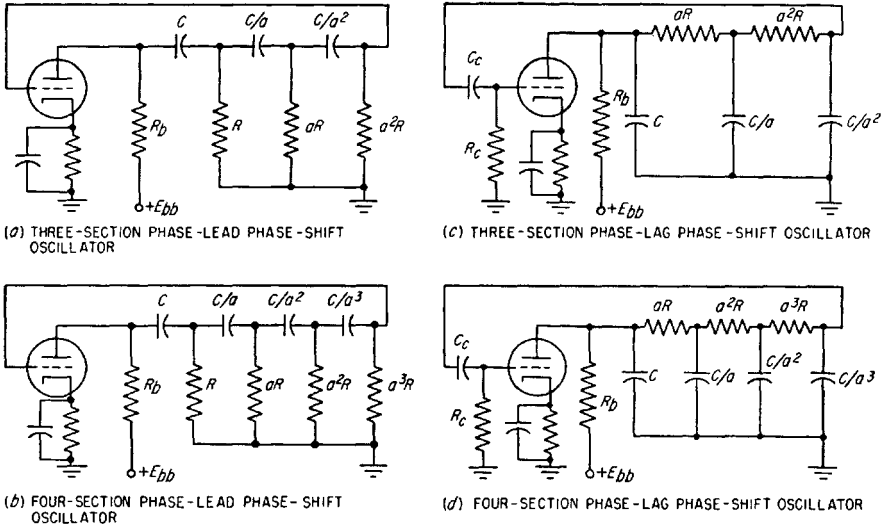
FIG. 6.45. Phase-shift oscillator circuits.

cycles to 100 kc. Their principal advantages over LC oscillators are the elimination of large and costly inductances and the property that frequency varies inversely with changes in R or C rather than inversely as the square root of L or C as in an LC oscillator.

6.7a. Phase-shift Oscillators. The phase-shift oscillator is a single-tube oscillator in which the output signal from the tube is shifted in phase by three or more cascaded RC networks to provide 180° phase shift at a specified frequency. The circuits of phase-shift oscillators containing three and four cascaded RC networks are shown in Fig. 6.45. The necessary phase shift can be obtained from either phase-lead networks or phase-lag networks. Three or more network sections are required to achieve the necessary 180° phase shift. The voltage gain of the tube necessary to provide a loop gain of unity is determined by the attenuation through the network, which varies as a function of the number of cascaded sections and the value of a (see Fig. 6.45). The frequency of oscillation is also a function of the number of cascaded sections and the value of a.

Phase-lead RC Oscillators. The equivalent circuits for the phase-lead oscillators are given in Fig. 6.46. The equivalent generator voltage is Ae_c, where

$$A = -\frac{\mu R_b}{r_p + R_b} \qquad (6.42)$$

The equivalent source impedance R_s is given by

$$R_s = \frac{r_p R_b}{r_p + R_b} \tag{6.43}$$

In order to sustain oscillation in the phase-lead oscillator, the minimum required gain

(a) EQUIVALENT CIRCUIT OF THREE SECTION
PHASE-LEAD OSCILLATOR CIRCUIT

(b) EQUIVALENT CIRCUIT OF FOUR SECTION PHASE-LEAD
OSCILLATOR CIRCUIT
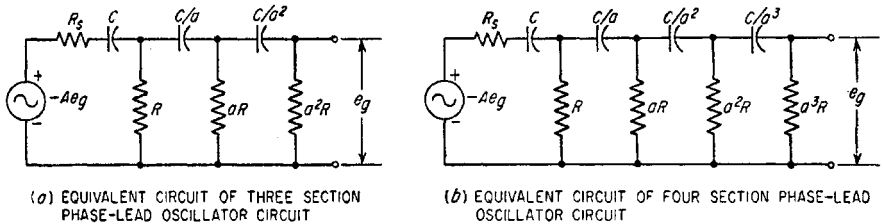
FIG. 6.46. Equivalent circuits of phase-lead RC oscillators.

A_r of the stage *in the absence of the loading of the phase-shift network* is given by Eqs.
(6.44) and (6.46). For the three-section phase-lead oscillator

$$A_r = -\left[8 + \frac{12}{a} + \frac{7}{a^2} + \frac{2}{a^3} + \frac{R_s}{R}\left(9 + \frac{11}{a} + \frac{4}{a^2}\right) + \left(\frac{R_s}{R}\right)^2 \left(2 + \frac{2}{a}\right) \right] \tag{6.44}$$

The frequency of oscillation for the three-section phase-lead oscillator is given by

$$\omega_{osc} = \frac{1}{RC\sqrt{3 + \dfrac{2}{a} + \dfrac{1}{a^2} + \dfrac{R_s}{R}\left(2 + \dfrac{2}{a}\right)}} \tag{6.45}$$

For the four-section phase-lead oscillator

$$A_r = -\left\{ \frac{3}{(\omega_{osc}RC)^2}\left[2 + \frac{2}{a} + \frac{1}{a^2} + \frac{R_s}{R}\left(1 + \frac{1}{a}\right) \right] - \frac{1}{(\omega_{osc}RC)^4} \right.$$
$$\left. - \frac{R_s}{R}\left(1 + \frac{1}{a} + \frac{1}{a^2} + \frac{1}{a^3}\right) - 1 \right\} \tag{6.46}$$

where $\omega_{osc}RC$ = normalized frequency of oscillation given by Eq. (6.47)
The frequency of oscillation for the four-section phase-lead oscillator is given by

$$\omega_{osc} = \frac{1}{RC\sqrt{\dfrac{4 + \dfrac{3}{a} + \dfrac{2}{a^2} + \dfrac{1}{a^3} + \dfrac{R_s}{R}\left(3 + \dfrac{4}{a} + \dfrac{3}{a^2}\right)}{4 + \dfrac{3}{a} + \dfrac{R_s}{R}}}} \tag{6.47}$$

The required values of gain A_r for the three-section and four-section phase-lead
networks are plotted as a function of a for various ratios of R_s/R in Figs. 6.47 and
6.48, respectively. The frequency of oscillation, assuming no additional phase shifts
due to tube capacitances or other causes, is given for the three-section and four-section
phase-lead oscillator circuits as a function of a for various ratios of R_s/R in Figs. 6.49
and 6.50, respectively. If

$$R_s \ll R + \frac{1}{j\omega_{osc}C}$$

the phase of the voltage at the plate of the oscillator tube will be essentially 180° out of phase with the grid voltage, resulting in an input capacitance to the tube of

$$C_i < c_{gk} + c_{gp}(1 + |A_r|) \qquad (6.48)$$

where A_r = minimum value of unloaded stage gain required for oscillation

This capacitance should have a reactance at least ten times the magnitude of the shunt resistance of the last network section if the additional phase shift caused
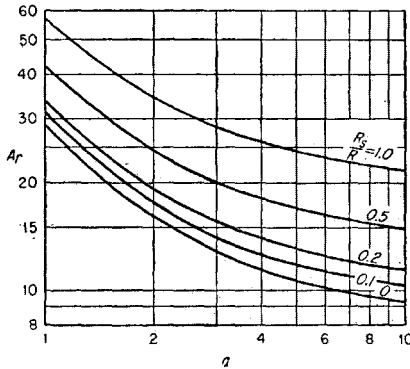
FIG. 6.47. Minimum values of unloaded stage gain for a three-section phase-lead RC oscillator.
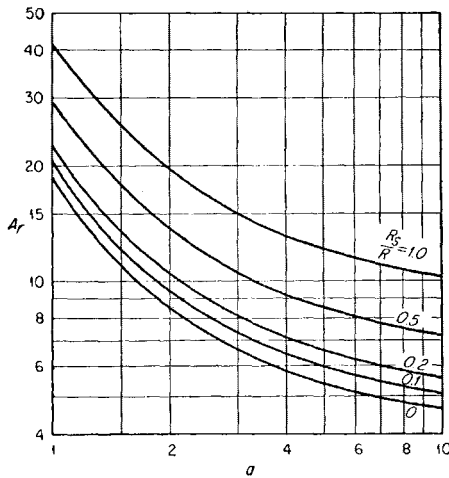
FIG. 6.48. Minimum values of unloaded stage gain for a four-section phase-lead RC oscillator.

by C_i is to be neglected. If R_s is not negligible compared to $R + \dfrac{1}{j\omega_{osc}C}$, the grid-plate capacitance will cause a resistive component to be present at the tube input. This may limit the maximum value of the effective shunt resistance of the last network section.

The steps to be followed in the design of a phase-lead RC oscillator are:

1. Select a tube with a high μ, low r_p, and low c_{gp}. Therefore, a tube with a high g_m is preferable. A pentode is suitable, but triodes are more often used. If a twin triode is used, the second half may be used as a cathode follower, thus providing a low source

impedance for the phase-shift network. This reduces the required gain A_r as indicated in Eqs. (6.44) and (6.46) and Figs. 6.47 and 6.48.

2. Select a value of load resistance R_b.

3. Determine R_s from Eq. (6.43).

4. Determine A from Eq. (6.42).

5. Determine the maximum value of the shunt resistance of the last network section.

6. Determine whether a three- or four-section network is to be used.
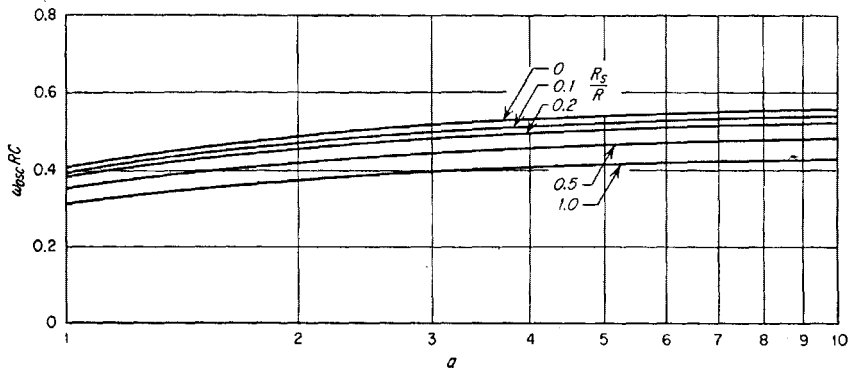
Fig. 6.49. Normalized frequency of oscillation of a three-section phase-lead RC oscillator.
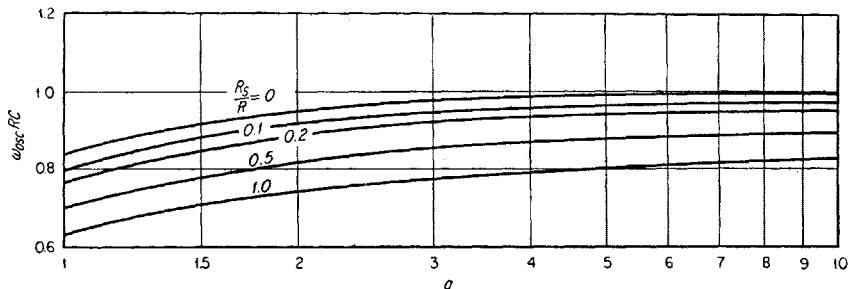
Fig. 6.50. Normalized frequency of oscillation of a four-section phase-lead RC oscillator.

7. Select a trial value of both a and R. Determine R_s/R and from Fig. 6.47 or 6.48 determine A_r, the required minimum value of A. If the value found in step 4 is larger than this minimum value, the values of a and R selected will be satisfactory. If the value of A in step 4 is not large enough, assume a larger value for R and/or a and redetermine R_s/R and A_r.

8. From Fig. 6.49 or 6.50, whichever is applicable, determine $\omega_{osc}RC$ from the selected values of a and R_s/R.

9. Determine C and the other resistances and capacitances in the remaining network sections.

10. Check the effect of c_{gp}.

The phase-lead oscillator circuit is most useful at frequencies of a few cycles per second and lower where the time constant of the grid blocking capacitor C_c and the grid resistor R_c required in a phase-lag oscillator must be several seconds or more in order not to introduce an undesired phase shift.

Phase-lag RC Oscillators. The circuits of three- and four-section phase-lag RC oscillators are given in Fig. 6.45c and d, respectively. In the phase-lag oscillator the parallel combination of r_p and R_b is made equal to R, the desired resistance of the first

section.   In order to sustain oscillation, the minimum required values of gain A_r *in the absence of the loading of the phase shift network* are given by Eqs. (6.49) and (6.50) for the three and four-section phase-lag circuits, respectively.

$$A_r = -\left(8 + \frac{12}{a} + \frac{7}{a^2} + \frac{2}{a^3}\right) \quad \text{(three-section)} \tag{6.49}$$

$$A_r = -\left[\left(4 + \frac{6}{a} + \frac{3}{a^2}\right)(1+B) - B^2\right] \quad \text{(four-section)} \tag{6.50}$$

where $B = \dfrac{\dfrac{2}{a^2} + \dfrac{1}{a^3}}{4 + \dfrac{3}{a}}$

$A = \dfrac{-\mu R_b}{r_p + R_b}$
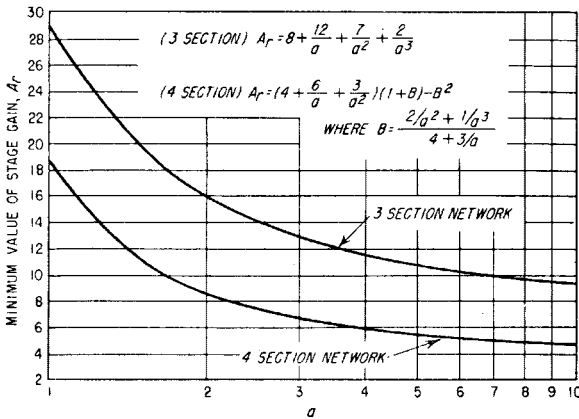
FIG. 6.51. Minimum values of unloaded stage gain for three- and four-section phase-lag RC oscillators.

The frequencies of oscillation for the three- and four-section phase-lag networks are given by Eqs. (6.51) and (6.52), respectively.

$$\omega_{osc} = \frac{\sqrt{3 + \frac{2}{a} + \frac{1}{a^2}}}{RC} \quad \text{(three-section)} \tag{6.51}$$

$$\omega_{osc} = \frac{\sqrt{\frac{4 + 3/a + 2/a^2 + 1/a^3}{4 + 3/a}}}{RC} \quad \text{(four-section)} \tag{6.52}$$

The minimum required unloaded gain A_r and the normalized frequency of oscillation $\omega_{osc}RC$ for the three- and four-section phase-lag oscillator circuits are given in Figs. 6.51 and 6.52, respectively.

The phase-lag oscillator circuit also possesses the advantage that the tube input capacitance can be made part of the last shunt capacitor C/a^n.   Since in triodes the input capacitance due to Miller effect may be large, this is a distinct advantage over the phase-lead oscillator circuit.   When approximately 60° phase shift is obtained per section in the three-section phase-lag circuit ($a > 4$), the input resistance and

capacitance of the amplifier including the effect of feedback through the grid-plate capacitance of the tube may be approximated by

$$C_i \simeq c_{gk} + \left(1 + \frac{|A_r|}{4}\right) c_{gp} \qquad (6.53)$$

$$R_x \simeq \frac{2.3}{\omega_{osc} c_{gp} |A_r|} \qquad (6.54)$$

When there is approximately 45° phase shift per section in the four-section phase-lag circuit ($a > 4$), the input resistance and capacitance of the amplifier due to feedback
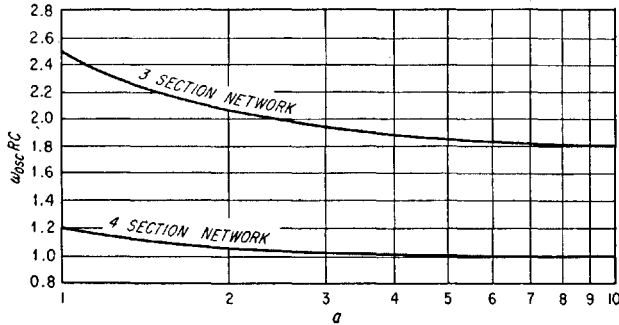
FIG. 6.52. Normalized frequency of oscillation of three- and four-section phase-lag oscillators as a function of a.

through the grid-plate capacitance of the tube may be approximated by

$$C_i \simeq c_{gk} + \left(1 + \frac{|A_r|}{2}\right) c_{gp} \qquad (6.55)$$

$$R_x \simeq \frac{2}{\omega_{osc} c_{gp} |A_r|} \qquad (6.56)$$

The steps to be followed in the design of a phase-lag oscillator are given below:

1. Select a tube with a high μ, low r_p, and low c_{gp}. Therefore a high g_m tube is preferable. A pentode is quite satisfactory, but triodes are more often used. If a twin triode is used, the second half may be used as a cathode follower, thus providing a low value of R which allows a high value of a.

2. Determine R_b. For the total grid circuit resistance, that is, R_c in parallel with R_x, not to introduce an appreciable phase shift by virtue of being in parallel with the shunt capacitance of the last network section, the total grid circuit resistance should be between five and ten times the reactance of the shunt capacitance at the frequency of oscillation. The input resistance R_x due to feedback will be at least five times the reactance of the shunt capacitance if the resistance R of the first section is selected to be

$$R = \frac{r_p R_b}{r_p + R_b} \leq \frac{1}{a \sqrt{1.26 \omega_{osc} c_{gp} g_m}} \qquad \text{(three-section)} \qquad (6.57)$$

$$R = \frac{r_p R_b}{r_p + R_b} \leq \frac{1}{\sqrt{2.5 a^3 \omega_{osc} c_{gp} g_m}} \qquad \text{(four-section)} \qquad (6.58)$$

Since the grid circuit resistor R_c is in parallel with R_x, R should be made somewhat less than the value given by Eq. (6.57) or (6.58). Assume a trial value of R_b and determine the maximum allowable value of a from Eq. (6.57) or (6.58).

3. Determine A_r from Fig. 6.51 using the value of a found in step 2. If it is less

than the value of A given by Eq. (6.42), the value of a found in step 2 is satisfactory. If the value of A_r from Fig. 6.51 is not less than that given by Eq. (6.42), a new value of R should be assumed and steps 2 and 3 repeated.

4. The value of C is determined from Fig. 6.52.

5. The resistances and capacitances of the remaining network sections are determined.

The phase-shift oscillator is normally used for fixed frequency operation because minimum tube gain is required with unequal values of R and C in successive sections. For variable frequency operation all sections can be made identical and a three- or four-gang variable resistor or capacitor used. Since frequency varies inversely with capacitance and resistance in the phase-shift oscillator instead of varying inversely as the square root of capacitance and inductance as in LC oscillators, a given range of capacitance variation results in an increased tuning range in the phase-shift oscillator.

The amplitude of oscillation in a phase-shift oscillator is determined by non-linearities in the tube, i.e., grid conduction and plate current cutoff. Grid-leak bias alone is usually not satisfactory if low distortion is desired because the high charging impedance when the grid conducts makes the rectification efficiency of the grid low, and the grid must conduct over a relatively large portion of the cycle to obtain adequate bias. Cathode bias helps greatly to improve the waveshape of the oscillator.

Example 6.1

Design a phase-shift oscillator to operate at 5,000 cycles utilizing one-half of a 12AT7 tube in a three-section phase-lag circuit. Assume a plate supply voltage of +250 volts.
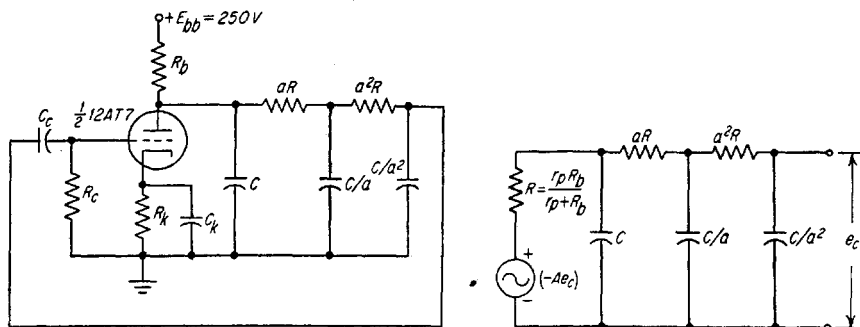
(a) CIRCUIT OF THE THREE-SECTION PHASE-LAG OSCILLATOR   (b) EQUIVALENT CIRCUIT

FIG. 6.53. Circuit of a phase-lag oscillator for Example 6.1.

Solution

1. Determine R_b.

See Figs. 6.53 and 6.54. R_b can be obtained from Eq. (6.57) once the r_p and g_m of the tube at the operating point have been found and c_{gp} has been determined. The grid-to-plate capacitance of the 12AT7 is approximately 1.5 $\mu\mu$f. To this should be added another 1.0 $\mu\mu$f for socket and wiring capacitances. Thus

$$c_{gp} \simeq 2.5 \ \mu\mu\text{f}$$

Assume a trial value of R_b of 25 kilohms and construct the load line as shown in **Fig. 6.54**. Assume a tentative bias of -1.6 volts and, therefore, a plate current of 4.0 ma. At this operating point

$$g_m = \frac{\Delta I_b}{\Delta E_c} = 5,300 \ \mu\text{mhos}$$

$$r_p = \frac{\Delta E_b}{\Delta I_b} = 16.0 \text{ kilohms}$$

Therefore, from Eq. (6.57)

$$a \leq \frac{r_p + R_b}{r_p R_b \sqrt{1.26\omega_{osc}C_{gp}g_m}}$$

$$\leq \frac{16.0 \times 10^3 + 25 \times 10^3}{16.0 \times 10^3 \times 25 \times 10^3 \sqrt{1.26 \times 2\pi \times 5,000 \times 2.5 \times 10^{-12} \times 5.3 \times 10^{-3}}}$$

$$\leq 4.48$$

Thus the ratio of R to input resistance due to feedback is low enough that an a of 4.48 could be used. However, since R_c in parallel with R_x will lower the total grid input resistance, a should be made less than 4.48 to minimize the reduction in phase shift due to R_c.
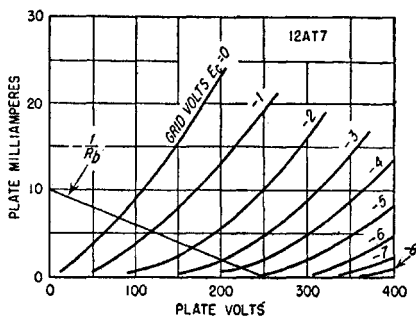
FIG. 6.54. Tube characteristic curves for Example 6.1.

Select a value of a of 4. The required amplifier gain A_r is then 11.5 (see Fig. 6.51). The amplifier gain A is given by

$$A = \frac{-\mu R_b}{r_p + R_b} = -g_m R = -5,300 \times 10^{-6} \times 9.76 \times 10^3$$

$$= -51.7$$

Thus, the gain margin for oscillation is ample.

2. Determine R_k and C_k

A bias of -1.6 volts is desired at the operating point. The necessary value of cathode resistor to obtain this value of cathode bias is

$$R_k = \frac{1.6}{4.0 \times 10^{-3}} = 400 \text{ ohms}$$

This cathode resistor should be bypassed by a capacitance having a reactance of $R_k/100$ or less in order that the phase shift due to the incompletely bypassed cathode resistor not appreciably affect the frequency of oscillation (see Sec. 3.10 and Fig. 3.28). A 10-μf capacitor has a reactance of 3.18 ohms at 5 kc. Therefore, let

$$C_k = 10 \ \mu f$$

3. Determine C.

From Fig. 6.52, the value of $\omega_{osc}RC$ for a three-section network having a value of a of 4 is seen to be 1.9. Thus

$$C = \frac{1.9}{\omega_{osc}R}$$

$$= \frac{1.9}{2\pi \times 5 \times 10^3 \times 9.76 \times 10^3}$$

$$= 6,200 \ \mu\mu f$$

4. Determine the resistance and capacitance values for the remaining network components ($R = 9.76 \times 10^3$ ohms from step 1).

$$aR = 4 \times 9.76 \times 10^3 = 39 \text{ kilohms}$$
$$a^2R = 156 \text{ kilohms}$$
$$\frac{C}{a} = 1,550 \ \mu\mu f$$
$$\frac{C}{a^2} = 388 \ \mu\mu f$$

From Eq. (6.53)

$$C_i \simeq c_{gk} + \left(1 + \frac{|A|}{4}\right) c_{gp}$$
$$\simeq 3.5 + \left(1 + \frac{51.7}{4}\right) 2.5$$
$$= 38.3 \ \mu\mu f$$

Therefore, C/a^2 should be reduced by 38.3 $\mu\mu$f, or

$$\frac{C}{a^2} \simeq 350 \ \mu\mu f$$

5. Determine R_c and C_c.

From Eq. (6.54), the input resistance due to feedback is given approximately by

$$R_x \simeq \frac{2.3}{\omega_{osc} c_{gp} |A|}$$
$$\simeq \frac{2.3}{2\pi \times 5 \times 10^3 \times 2.5 \times 10^{-12} \times 51.7}$$
$$\simeq 567,000 \text{ ohms}$$

The parallel combination of R_c and R_x should be equal to or greater than

$$\frac{5a^2}{\omega_{osc} C} = 411,000 \text{ ohms.}$$

Therefore, let $R_c = 2$ megohms. The reactance of C_c should be approximately $R_c/100$ or less in order that the phase shift of the coupling network not appreciably affect the frequency of oscillation. Therefore, let $C_c = 0.01 \ \mu$f.

6.7b. Other Types of RC Feedback Circuits. There are numerous other *RC* networks which can be used to provide either 180 or 0° phase shift at a specific frequency
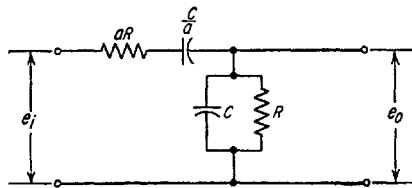
FIG. 6.55. Feedback network for *RC* oscillators.

and can, therefore, be used as feedback networks for an *RC* oscillator. A commonly used network is shown in Fig. 6.55. This network provides 0° phase shift at a frequency given by

$$f_{osc} = \frac{1}{2\pi RC} \tag{6.59}$$

It must be used in an oscillator where the required phase shift in the feedback path at the frequency of oscillation is 0°. The gain $1/A_r$ through the network at f_{osc} is plotted as a function of the ratio a in Fig. 6.56. The required amplifier no-load gain is A_r provided the amplifier output impedance is included as a part of aR in Fig. 6.55.

This network is frequently used where a relatively wide tuning range is desired.   The
ratio a is then made equal to unity, and a two-gang variable capacitor is used to tune
the oscillator.   The frequency varies inversely with C rather than as $\sqrt{C}$ as in LC

$$A = \frac{1}{1 + 2a}$$
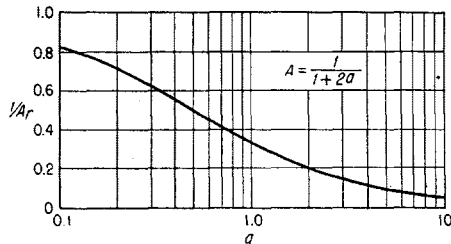
Fig. 6.56. Gain at $f = f_{osc}$ through network of Fig. 6.54 as a function of a.

$$\frac{1}{A_r} = \frac{1}{\sqrt{9 + \left(\omega - \frac{1}{\omega}\right)^2}}$$

$$\phi = -TAN^{-1} \frac{\omega - \frac{1}{\omega}}{3}$$
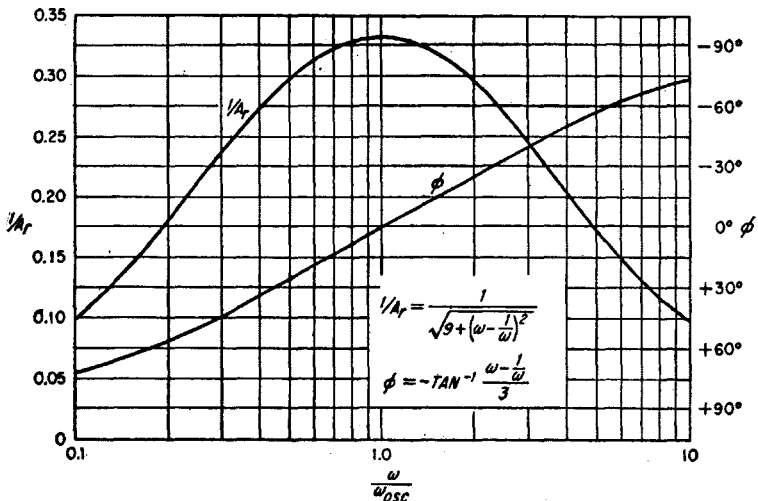
Fig. 6.57. Gain and phase shift ϕ of network of Fig. 6.54 for $a = 1$.

oscillators.   The attenuation and phase shift of the network as a function of frequency
for a equal to unity is shown in Fig. 6.57.

Another type of RC network which may be used to provide feedback paths having
either 180 or 0° phase shift at a selected frequency is the parallel-T network.[1]

[1] The parallel-T network having positive and negative transfer constants is discussed in
Sec. 16.2b.   Certain additional advantages may be achieved by the use of RLC Bridged-T
networks as discussed in Sec. 16.2a.

Receivers

The design requirements of a receiver depend upon the intended application. The suitability of a particular type to a specific application is measured by its ability to meet the performance requirements, i.e., sensitivity, noise figure, selectivity, gain, bandwidth, phase and amplitude linearity, image rejection, etc.

Design information on r-f and i-f amplifiers, mixers, local oscillators, detectors, superheterodyne receivers, tuned radio frequency (trf) receivers, crystal video receivers, superregenerative receivers, and f-m receivers is presented. A discussion of noise and its effects on receiver performance is also included.

7.1. Classification of Receivers. A receiver is any device which accepts and demodulates r-f signals to obtain the information or intelligence contained in the signal. In most applications it is necessary to amplify the demodulated signal to a level suitable for use. Since the amplitudes of the signals at the input to the receiver are ordinarily extremely small, a typical receiver must amplify the received signal by a factor of several thousand before the signal is of sufficient amplitude to be of use. This amplification may be accomplished before and/or after the signal has been demodulated.

The intelligence contained in the r-f signal may be in the form of continuous-wave (c-w) amplitude modulation, frequency modulation, pulse-amplitude modulation, pulse-length modulation, pulse-time modulation, etc.

Receivers are divided into the following types: *superheterodyne, tuned r-f, regenerative and superregenerative, crystal video,* and *f-m.*

In the superheterodyne the r-f signal is converted to an intermediate frequency where it is amplified and then demodulated.

The trf receiver is one in which the r-f signal is amplified to a relatively high level by tuned amplifiers resonant at the frequency of the incoming signal and then demodulated.

In regenerative and superregenerative receivers amplification is increased by the use of a controlled amount of positive feedback. In the regenerative receiver, the greatest amplification (sensitivity) is obtained when the amount of feedback is slightly less than that required for oscillation. In the superregenerative receiver an r-f amplifier or plate detector having sufficient positive feedback to cause oscillation is made nonoscillatory periodically by the application of a quench signal which may be developed from the oscillating stage or be obtained from an external source. Superregeneration provides considerably more amplification in a single stage than does regeneration.

The crystal video receiver is one in which the r-f signal is demodulated by a crystal diode detector at a low signal level prior to amplification. The demodulated signal is then amplified to a usable level by audio-frequency or video-frequency amplifiers.

Frequency-modulation receivers are capable of accepting a frequency-modulated signal and converting the frequency variations into amplitude variations and may be of the superheterodyne, trf, or superregenerative type.

7.2. Noise. Noise can be defined as any random or aperiodic signal which interferes with the extraction of intelligence from a desired signal. Other types of interference of a periodic nature, such as engine ignition radiation, vibration, hum, and other unwanted signals, are frequently considered to constitute noise inasmuch as they also interfere with the recognition of a desired signal. The effect of any type

of noise in a receiver is to limit the minimum signal amplitude which the receiver can demodulate satisfactorily.

The noise present at the receiver output terminals may have originated at a number of sources. Table 7.1 lists the major sources of noise and classifies the source according to whether it is external or internal to the receiver and its antenna.

TABLE 7.1. MAJOR SOURCES OF NOISE

Noise Sources External to Receiver and Antenna	Noise Sources Internal to Receiver and Antenna
Atmospheric noise	Thermal noise
Cosmic noise	Vacuum-tube noise
Man-made noise	Shot-effect noise
Precipitation static	Plate shot-effect noise
Antenna radiation noise	Grid-induced shot-effect noise
	Flicker-effect noise
	Secondary-emission noise
	Positive-ion noise
	Contact and breakdown noise
	Magnetic-fluctuation noise

7.2a. Atmospheric Noise. Atmospheric noise, often called static, is generated by disturbances such as electrostatic discharges between clouds. The noise thus produced varies in intensity with frequency, time of day, geographical location, and weather. The energy of this impulse type of noise is distributed primarily throughout a major portion of the spectrum below 50 Mc. Atmospheric noise is propagated in exactly the same manner as radio signals of the same frequency, and its intensity at any location is thus dependent upon propagation conditions. At frequencies above 50 Mc, less atmospheric noise is generated, and there is normally very little reflection from the ionosphere. Consequently, atmospheric noise above 50 Mc is lower in intensity and is local in origin. Figure 7.1 illustrates the typical variation in atmospheric noise as a function of frequency and time of day for a particular location.

Interference due to atmospheric noise can be minimized by utilizing the following techniques: (1) reduction of receiver bandwidth so that the sidebands of the highest modulating frequency desired fall just within the receiver passband, (2) amplitude limiting in the receiver in cases where the nature of the noise is such that the noise signals are of high amplitude and rather widely spaced in time, (3) utilization of a directive-receiving antenna when the major source of the noise is in a direction other than that from which the desired signal originates, (4) utilization of a high-gain, directive-transmitting antenna to increase the signal-power density at the receiving antenna, and (5) the use of high transmitter power.

7.2b. Cosmic Noise. Cosmic noise originates in the regions beyond the earth's atmosphere. By convention, cosmic noise is subdivided into two categories: solar noise and galaxy noise. Solar noise is radiated by the sun, and galaxy noise is all other noise arriving from interstellar space. The largest source of galaxy noise appears to be near the constellation Sagittarius. It is believed that cosmic noise is principally due to celestial-body thermal noise, i.e., noise radiated by celestial bodies because of their temperature. However, measurements of solar noise show that the black-body temperature of the sun is insufficient to cause thermal noise to be the entire source of solar noise.[1]

The intensity of interstellar noise is very low. In the absence of man-made noise and atmospheric noise, however, it may be the factor which limits a receiver's minimum detectable signal in the range of 10 to 300 Mc.

[1] H. V. Cottony and J. R. Johler, "Cosmic Radio Noise Intensities in the VHF Band," *Proc. IRE*, September, 1952.

7.2c. *Man-made Noise.* Man-made noise is produced by high-voltage power-line leakage, aircraft and automobile ignition systems, brush motors, diathermy, industrial heating generators, etc. Although all of these sources do not produce interference which is random, this interference is often considered to be noise. The intensity of this type of interference is greatest in cities and industrialized areas.

7.2d. *Precipitation Static.* Precipitation static often occurs in aircraft passing through snow, rain, or the electric fields of thunderclouds. Under these conditions, the aircraft may become charged to a sufficiently high potential with respect to the
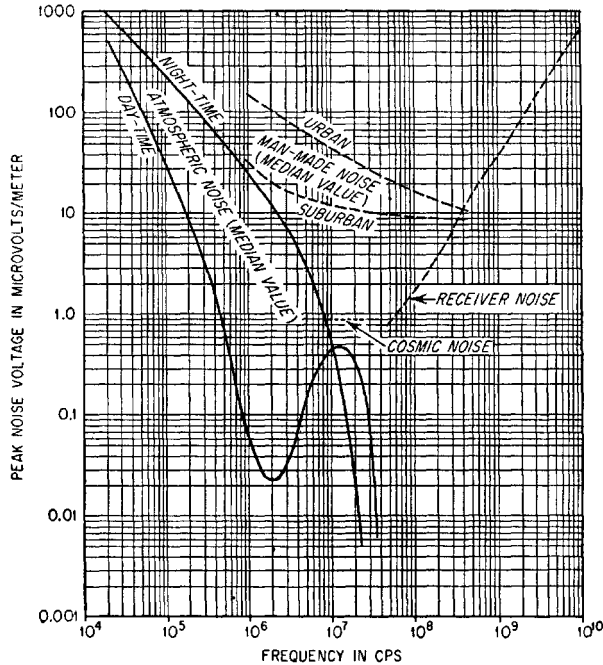
FIG. 7.1. Relative magnitudes of the major sources of r-f noise at various frequencies as measured by a receiver having a bandwidth of 10 kc and using a half-wave dipole antenna. ("Reference Data for Radio Engineers," 3d ed., Federal Telephone and Radio Corporation, 1951.)

surrounding space that a corona discharge takes place at sharp points on the surface of the plane. The disturbance created by the corona discharge consists of a strong induction field in which the electrostatic field is much stronger than the magnetic field.

Precipitation static can be minimized by using electrostatically shielded loop antennas, by equipping the aircraft with static dischargers which allow any accumulated charge to leak off slowly through a high resistance, and by locating the antennas as far as possible from the points of corona discharge.

7.2e. *Contact and Breakdown Noise.* Contact and breakdown noise is due to the breakdown of insulation or the interruption of minute paths within components. This type of noise exists in addition to thermal noise in some types of resistors. Contact and breakdown noise can be minimized by taking proper precautions in the manufacture of components.

7.2f. *Thermal Noise.* The free electrons in any conductor are in continuous motion. This motion is completely random and is the result of thermal agitation. The effect

of this electron motion is to cause minute voltages which vary in a random manner to be developed across the terminals of the conductor. This phenomenon was demonstrated by J. B. Johnson in 1928, and consequently thermal noise is sometimes referred to as Johnson noise. At the same time, H. Nyquist showed on the basis of the statistical theory of thermodynamics that the mean-square thermal-noise voltage generated in any impedance can be expressed as

$$E_n{}^2 = 4KT \int_{f_1}^{f_2} R(f) \, df \tag{7.1}$$

where E_n = rms value of thermal-noise voltage components within frequency band
f_1 to f_2, volts
K = Boltzmann's constant = 1.38×10^{-23} joules/°K
T = absolute temperature of R, °K
$R(f)$ = resistive component of impedance in ohms at frequency f
f = frequency, cps

If the resistive component of the impedance is independent of frequency, Eq. (7.1) reduces to

$$E_n{}^2 = 4KT \, \Delta f R \tag{7.2}$$

where Δf = bandwidth f_2-f_1, cps
Equations (7.1) and (7.2) indicate that the mean-square voltage in the frequency range from $f = 0$ to $f = \infty$ is infinite; however, this is not correct. At extremely high frequencies, classical statistical mechanics is no longer valid, and the equations must be revised on the basis of quantum theory. However, Eqs. (7.1) and (7.2) are valid for frequencies up to at least 6×10^{12} cps.[1]

(a) EQUIVALENT NOISE
VOLTAGE GENERATOR

(b) EQUIVALENT NOISE
CURRENT GENERATOR
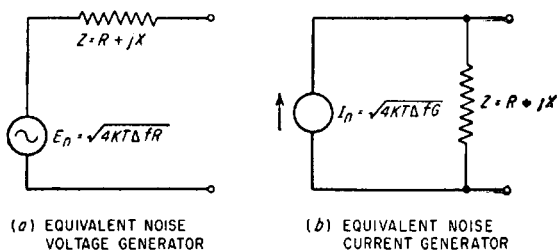
FIG. 7.2. Equivalent circuit of an impedance as a noise generator.

The equivalent circuit of any impedance as a source of noise voltage is shown in Fig. 7.2a. The thermal-noise voltage generated is dependent only on the resistive component of the impedance and is independent of the reactive component. Equations (7.1) and (7.2) show that the noise power per cycle is independent of the particular frequency band in which the noise is measured.

The equivalent voltage representation of a thermal-noise generator can be replaced by an equivalent current representation as shown in Fig. 7.2b. The magnitude of the equivalent noise current is given by Eq. (7.3).

$$I_n{}^2 = 4KT \, \Delta f G \tag{7.3}$$

where G = conductance of admittance Y (see Fig 7.2b)

$$G = \frac{R}{R^2 + X^2} \tag{7.4}$$

[1] S. Goldman, "Frequency Analysis, Modulation and Noise," p. 394, McGraw-Hill Book Company, Inc., New York, 1948.

Maximum noise power is transferred from a thermal-noise source when the load impedance is the conjugate of the source impedance. The power which could theoretically be transferred from a thermal-noise source under such conditions is

$$P_n = KT \, \Delta f \tag{7.5}$$

This power is termed the *available noise power*. However, since any resistor, regardless of the value of resistance, has the same available noise power at a given temperature, there is no net power transfer between two resistors which are connected together, regardless of their respective values of resistance, if they are at the same temperature.

If two or more independent random noise voltages are superimposed, the resultant noise voltage is equal to the square root of the sum of the squares of the noise voltages. If two resistors are placed in series as in Fig. 7.3, the resultant rms noise voltage E_n appearing across the two resistors is given by
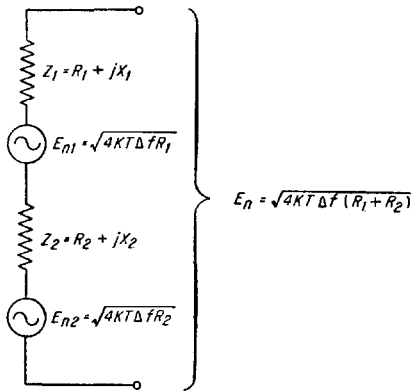
FIG. 7.3. Noise voltage output of two impedances in series.

$$E_n = \sqrt{E_{n1}{}^2 + E_{n2}{}^2} = \sqrt{4KT \, \Delta f(R_1 + R_2)} \tag{7.6}$$

Example 7.1

A signal generator having an internal resistance of 10,000 ohms is connected to the primary of an ideal transformer having a turns ratio of 1:10. The transformer has a passband from 10 cycles to 100 kc. The secondary of the transformer is connected to a 1-megohm load resistor R_L. Determine the rms noise voltage across the load resistor, assuming a temperature of 17°C.

1. Determine the resistance seen looking into the secondary of the ideal transformer.

$$R_2 = N^2 R_1$$

The primary circuit resistance is 10,000 ohms. The secondary impedance of the transformer is, therefore,

$$R_2 = 10^2 \times 10^4$$
$$= 1 \times 10^6 \text{ ohms}$$

2. Determine the parallel combination of R_s and R_L.

$$\frac{R_s R_L}{R_s + R_L} = 0.5 \times 10^6 \text{ ohms}$$

3. Determine the rms noise voltage. (0°C = 273°K)

$$E_n = \sqrt{4KT \, \Delta f R}$$
$$= \sqrt{4 \times 1.38 \times 10^{-23} \times 290 \times (10^5 - 10) \times 0.5 \times 10^6}$$
$$= \sqrt{8 \times 10^{-10}}$$
$$= 28.3 \, \mu\text{v}$$

Example 7.2

Determine the thermal-noise voltage developed within the 10-kc passband of a receiver by the resonant circuit illustrated in Fig. 7.4 when the circuit is resonant at the center frequency of the receiver. Assume that the temperature is 17°C.

1. Determine the resistive component of the impedance connected to the grid of the tube.

$$Z = \frac{(R + j\omega L)(1/j\omega C)}{R + j\omega L + 1/j\omega C}$$

The resistive component of Z is

$$R' = \frac{R}{(1 - \omega^2 LC)^2 + \omega^2 C^2 R^2}$$

At resonance

$$\omega_o{}^2 LC = 1$$

and R' reduces to

$$R' = \frac{1}{\omega_o{}^2 C^2 R}$$

or

$$R' = \frac{\omega_o{}^2 L^2}{R}$$

2. Determine $E_n{}^2$.

To determine rigorously $E_n{}^2$, Eq. (7.1) must be employed and R' evaluated at each frequency within the 10-kc passband. This exact solution is seldom employed in practice because of the complexity of the computation involved. Instead, the bandwidth of the noise developed in the resonant circuit is usually assumed to be equal to the bandwidth between the -3-db frequencies,[1] and Eq. (7.2) is used to determine $E_n{}^2$. If the latter method is used,

$L = 63.1\ \mu H$
$C = 400\ \mu\mu F$
$R = 3.96\ OHMS$
$f_{RESONANCE} = 1MC \qquad Q = 100$

FIG. 7.4. Circuit for Example 7.2.

$$E_n{}^2 = 4KT \ \Delta f R'$$

where $R' = \dfrac{\omega_o{}^2 L^2}{R} = Q^2 R$

$\begin{aligned}
E_n{}^2 &= 4KT \ \Delta f Q^2 R \\
&= 4 \times 1.38 \times 10^{-23} \times 290 \times 10^4 \times 100^2 \times 3.96 \\
&= 6.34 \times 10^{-12} \\
E_n &= 2.52 \ \mu v
\end{aligned}$

7.2g. Antenna Radiation Noise. The antenna to which a receiver is coupled may

(a) EQUIVALENT CIRCUIT OF AN ANTENNA (b) EQUIVALENT NOISE VOLTAGE CIRCUIT OF AN ANTENNA
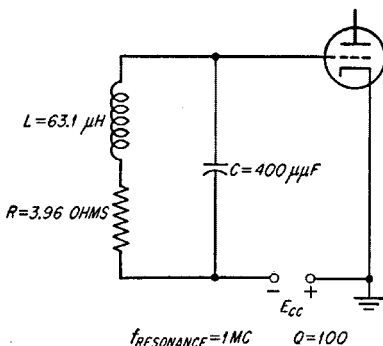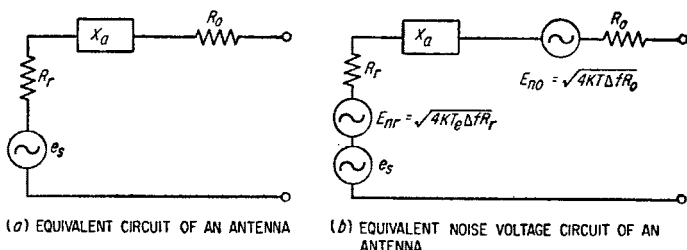
FIG. 7.5. Antenna equivalent circuit.

be represented by an impedance Z_a in series with a voltage generator e_s, representing the received signal, as shown in Fig. 7.5a.

$$Z_a = R_r + R_o + jX_a \qquad (7.7)$$

where Z_a = impedance of antenna circuit
R_r = radiation resistance of antenna
R_o = effective loss resistance of antenna circuit
X_a = reactance of antenna circuit

[1] The exact relationship between the noise bandwidth and the -3-db bandwidth of cascaded tuned stages may be found in "Vacuum Tube Amplifiers," p. 169, by G. E. Valley, Jr., and H. Wallman, McGraw-Hill Book Company, Inc., New York, 1948.

R_r is that resistance which, if connected in series with the antenna, would dissipate the same amount of power that is radiated by the antenna when used for transmitting. It represents the apparent resistance of an otherwise lossless antenna.

The loss resistance R_o of the antenna develops a thermal noise voltage in exactly the same manner as other ohmic resistances, and this voltage can be taken into account by inserting a noise voltage generator in series with R_o whose output is given by

$$E_{no} = \sqrt{4KT \, \Delta f R_o} \tag{7.8}$$

The radiation resistance of the antenna represents the resistance of the equivalent circuit voltage generator associated with the power received by the antenna. If an antenna having negligible loss were in exact thermal equilibrium with its entire surroundings, it would have an available thermal noise power at its terminals of

$$P_n = KT_s \, \Delta f \tag{7.9}$$

where T_s = temperature of antenna system, consisting of antenna and space into which it radiates

This noise power is received from the antenna's surroundings and is not generated within the antenna itself. However, the equivalent circuit noise voltage generator may be represented by

$$E_{nr} = \sqrt{4KT_s \, \Delta f R_r} \tag{7.10}$$

The entire surroundings of any actual antenna are not at a uniform temperature. Also, man-made noise, atmospheric noise, and cosmic noise are all absorbed from space by the antenna in addition to the thermal noise radiated from surrounding objects. Therefore the noise available from the antenna terminals is different from that indicated by Eqs. (7.9) and (7.10) by an amount that is dependent upon the location of the antenna, frequency, and the antenna's directional characteristics. Since the noise present at the antenna terminals cannot be separated as to the sources from which it originated, it is commonly lumped together mathematically by representing the effective noise-source voltage as

$$E_{nr} = \sqrt{4KT_e \, \Delta f R_r} \tag{7.11}$$

where T_e = effective noise temperature of radiation resistance, i.e., that temperature to which radiation resistance must be raised for noise present in antenna to be given by Eq. (7.11)

The square of the total noise voltage E_n present in an actual antenna having ohmic resistance is given by

$$E_n{}^2 = 4KT_e \, \Delta f R_r + 4KT \, \Delta f R_o \tag{7.12}$$
$$= 4KT \, \Delta f \left(\frac{T_e}{T} R_r + R_o \right)$$

The complete antenna equivalent circuit including noise sources is given in Fig. 7.5b.

7.2h. Vacuum-tube Noise. There are several important sources of noise in vacuum tubes (see Table 7.1). These noise sources together with thermal noise limit the maximum useful gain of an amplifier.

1. Shot-effect Noise. The most important type of tube noise is shot-effect noise. This noise arises from the fact that the electron emission from a hot cathode constitutes a current composed of a great many separate charges which are emitted in a random manner. As a result, the emission current exhibits minute fluctuations. This continuous fluctuation in tube current is called shot-effect noise. Noise due to shot effect is of two types: *plate shot-effect noise* and *grid-induced shot-effect noise.*

Plate Shot-effect Noise. Plate shot-effect noise refers to (1) the variations in plate current due to the randomness of electron emission from the cathode and (2) in multi-collector tubes, the additional effect due to the randomness of the space-current division between the collectors, e.g., between the plate and screen grid in a pentode. Space-current noise induces noise in the grid circuit which in turn increases the plate-current noise. This effect is referred to as grid-induced shot-effect noise and is treated separately from plate shot-effect noise.

PLATE SHOT-EFFECT NOISE IN DIODES AND NEGATIVE-GRID TRIODES. Plate shot-effect noise is largest when the tube current is limited by cathode temperature, i.e., when the plate voltage is high enough to attract all emitted electrons. Under this condition, the noise component of plate current in a diode, or negative-grid triode, or any tube in which the entire cathode current goes to one electrode is

$$I_{np} = \sqrt{2eI_b\,\Delta f} \qquad \text{(temperature-limited diode or triode)} \qquad (7.13)$$

where I_{np} = rms value of plate noise current in amperes measured in frequency band Δf
   e = charge of one electron = 1.60×10^{-19} coulomb
   I_b = d-c anode current, amp
   Δf = frequency band in cycles per second in which noise is measured

Equation (7.13) is valid up to frequencies where electron transit time becomes important. For calculating the effects of plate shot-effect noise in circuits external to the tube, the tube may be considered a current generator shunted by the plate resistance of the tube as shown in Fig. 7.6.

When a tube is operated so that the anode current is limited by plate voltage, the resulting electron space charge provides a reservoir of electrons which reduces the effect of the fluctuations in the cathode emission current on the plate current. Thus, plate shot-effect noise is greatly reduced if the plate current is limited by plate voltage. For the voltage-limited condition and provided that the frequencies are low enough so that

FIG. 7.6. Equivalent shot-effect noise generator representation of a vacuum tube.

electron transit-time effects can be ignored, the shot-effect noise current for a diode is given by

$$I_{np} = \sqrt{0.644(4KT_c\,\Delta f)g} \qquad \text{(voltage-limited diode)} \qquad (7.14)$$

where I_{np} = rms value of noise current, amp
   K = Boltzmann's constant
   T_c = absolute temperature of cathode, °K (usually about 1000°K for oxide-coated cathodes)
   Δf = frequency band in cycles per second in which noise is measured
   g = plate conductance of diode = $1/r_p$

For a negative-grid triode whose plate current is voltage-limited, the plate shot-effect noise current is given by

$$I_{np} = \sqrt{\frac{0.644(4KT_c\,\Delta f g_m)}{\sigma}} \qquad \text{(triode amplifier)} \qquad (7.15)$$

where g_m = grid-plate transconductance, mhos
   σ = parameter of tube (usually between 0.5 and 1.0)

The plate shot-effect noise current in triodes and pentodes can be referred to the grid circuit of the tube as an equivalent noise-voltage generator E_{ng} in series with the

grid by the relation

$$E_{ng} = \frac{I_{np}}{g_m} \qquad (7.16)$$

This should not be confused with grid-induced shot-effect noise. The plate shot-effect noise referred to the grid circuit may then be represented by

$$E_{ng} = \sqrt{\frac{0.644(4KT_c\,\Delta f)}{\sigma g_m}} \qquad \text{(triode amplifier)} \qquad (7.17)$$

The equivalent grid-noise voltage E_{ng} is equal to the noise that would be generated by a resistor R_{eq} in the grid circuit of value

$$R_{eq} = \frac{0.644 T_c}{\sigma T g_m} \qquad \text{(triode amplifier)} \qquad (7.18)$$

Once R_{eq} is determined for a particular tube, the plate shot-effect noise can be repre-

sented by an equivalent grid resistance in series with the grid of an ideal tube having no plate shot-effect noise. This method of representation is particularly convenient because it provides a measure of the noisiness of the tube which is independent of bandwidth. The equivalent resistance can be added directly to the resistive component of the external grid-circuit impedance to determine the over-all effect of thermal noise and plate shot-effect noise. It must be remembered, however (see Fig. 7.7) that the noise voltage generator in the equivalent circuit is in series with the tube grid and
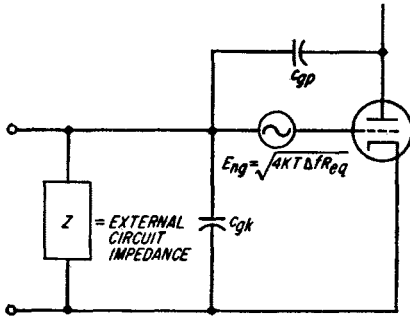
FIG. 7.7. Equivalent circuit for plate shot-effect noise referred to grid.

is between the grid itself and any interelectrode capacitance or externally connected impedance.

The equivalent grid noise resistance for a number of tubes is given in Table 7.2. An approximation to Eq. (7.18) can be made by assuming typical values of σ, T_c, and T. For $T = 293°K$, or 20°C, $T_c = 1000°K$, and $\sigma = 0.88$, Eq. (7.18) reduces to

$$R_{eq} = \frac{2.5}{g_m} \qquad \text{(triode amplifier)} \qquad (7.19)$$

Although Eq. (7.19) is not exact, it is usually sufficiently accurate for most purposes.

Before proceeding with a discussion of other types of tube noise, two examples are presented to illustrate the application of the preceding material.

Example 7.3

A diode whose emission current is temperature-limited is to be used as a noise generator for receiver measurements at 30 Mc. If the receiver bandwidth is 3 Mc and the maximum d-c plate current which the diode can conduct is 100 ma, determine the effective noise voltage which the diode will generate across the input terminals of the receiver. Assume that the receiver input resistance R_i is 300 ohms.

1. Determine the noise current developed in the diode.

The noise current I_{np} developed by the diode is determined by Eq. (7.13).

$$I_{np} = \sqrt{2eI_b\,\Delta f}$$

TABLE 7.2. EQUIVALENT GRID NOISE RESISTANCE FOR SOME COMMON TUBES

Tube type	E_b, volts	E_{c2}, volts	E_c, volts	I_b, ma	I_{c2}, ma	I_t, ma	g_m, μmhos	R_{eq} calculated
Triode amplifier:								
6AK5	120	120	−2.0	....	...	10.2	6,750	370
6AH6	150	150	−2.2	....	...	12.5	10,500	238
6CB6	150	150	−2.2	....	...	13.0	8,450	296
6AC7	150	150	−2.0	....	...	12.5	11,200	223
6AG7	150	150	−3.0	....	...	38.0	13,930	180
6AU6	150	150	−1.0	....	...	15.1	7,250	345
6BA6	150	150	−1.0	....	...	17.0	6,380	392
6C4	100	...	0	11.8	...	11.8	3,100	807
6J6	100	...	−0.4	8.5	...	8.5	5,300	472
12AU7	250	...	−8.5	11.0	...	11.0	2,250	1,110
12AT7	250	...	−2.0	10.0	...	10.0	5,500	455
12AX7	250	...	−2.0	1.2	...	1.2	1,600	1,560
6J4	150	...	−1.5	15.0	...	15.0	12,000	208
Pentode amplifier:								
6AK5	180	120	−2.0	7.7	2.5	10.2	5,100	1,820
6AH6	300	150	−2.2	10.0	2.5	12.5	9,000	717
6CB6	200	150	−2.2	10.0	3.0	13.0	6,500	1,390
6AC7	300	150	−2.0	10.0	2.5	12.5	9,000	717
6AG7	300	150	−3.0	30.0	8.0	38.0	11,000	1,220
6AU6	250	150	−1.0	10.8	4.3	15.1	5,200	2,605
6BA6	250	100	−1.0	12.0	5.0	17.0	4,500	3,820
6SJ7	300	125	−1.8	8.3	2.4	10.7	2,370	7,460
6SK7	250	100	−3.0	9.0	2.0	11.0	2,000	9,200
Triode mixer:								
6AK5	120	120	−1.0#	....	...	5.7	2,400*	1,330
6AH6	150	150	−1.0#	....	...	8.3	3,720*	857
6AC7	150	150	−1.0#	....	...	8.3	4,330*	735
6J4	150	...	−1.5#	4.8	...	4.8	3,000*	1,060
Pentode mixer:								
6AK5	180	120	−1.0	4.3	1.4	5.7	1,825*	7,570
6AH6	300	150	−1.0#	6.7	1.6	8.3	3,000*	3,170
6AC7	250	150	−1.0#	6.7	1.6	8.3	3,500*	2,830
Multigrid mixer:								
6BA7	250	100	−1.0	3.8	...	14.2	950*	61,700
6BE6	250	100	−1.5	3.0	...	10.6	470*	195,000
6SA7	250	100	−2.0	3.5	...	12.5	450*	249,000
6L7	250	150	−6.0	3.3	...	12.5	350*	397,000

NOTE: The average values of current in all cases are the d-c values.
# denotes the instantaneous grid voltage at the positive peak of local oscillator cycle.
* denotes the average value of conversion transconductance.
Values shown for all twin triode tubes are for one section only.

2. Determine the noise voltage developed across the receiver input due to I_{np}.

The equivalent circuit is shown in Fig. 7.8. The diode dynamic plate resistance is defined as

$$r_p = \frac{\Delta e_b}{\Delta i_b}\bigg|_{T_c = \text{const}}$$

Since the diode is operated in the temperature-limited region, the rate of change of plate current with plate voltage is zero because the emission current is being limited by the cathode temperature. Thus

$$r_p = \infty$$
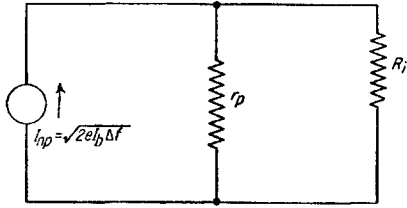
Fig. 7.8. Equivalent circuit for Example 7.3.

therefore,

$$
\begin{aligned}
E_{np} &= R_i \sqrt{2eI_b\,\Delta f} \\
&= 300 \sqrt{2 \times 1.60 \times 10^{-19} \times 0.1 \times 3 \times 10^6} \\
&= 93 \ \mu\text{v}
\end{aligned}
$$

Note that the total noise voltage developed across R_i must include the thermal noise voltage due to the input resistance itself.

Example 7.4

If it is assumed that neither $V1$ nor R_b contribute noise to the output of the circuit shown in Fig. 7.9a, determine the noise voltage required at the grid to produce the same amount

(a) CIRCUIT OF A ONE-STAGE AMPLIFIER

(b) EQUIVALENT GRID-NOISE CIRCUIT OF A ONE-STAGE AMPLIFIER
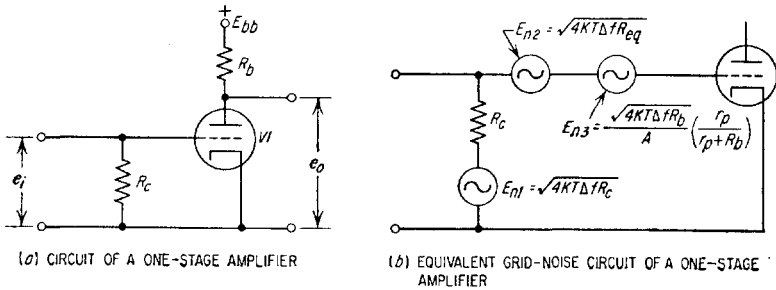
Fig. 7.9. Circuit for Example 7.4.

of output noise which would exist in an actual circuit where R_b contributes thermal noise and $V1$ contributes plate shot-effect noise only.

1. Determine the amplifier gain.
The equivalent noise circuit is shown in Fig. 7.9b.   The gain of the amplifier is

$$A = \frac{-\mu R_b}{r_p + R_b}$$

2. Determine the thermal-noise voltage of the load resistor referred to the tube grid. The noise voltage at the output due to R_b is reduced from $\sqrt{4KT\,\Delta fR_b}$ by the voltage divider action of r_p.
The equivalent grid noise source E_{n3} which would provide a noise voltage equal to $\sqrt{4KT\,\Delta fR_b}\left(\dfrac{r_p}{R_b + r_p}\right)$ is given by

$$E_{n3} = \frac{\sqrt{4KT\,\Delta fR_b}}{A}\left(\frac{r_p}{R_b + r_p}\right)$$

3. Determine the total effective noise voltage referred to the tube grid.
The total equivalent noise voltage E_n appearing at the grid of $V1$ is then

$$E_n = \sqrt{E_{n1}{}^2 + E_{n2}{}^2 + E_{n3}{}^2}$$

where E_{n1} = thermal noise voltage of R_c

E_{n2} = plate shot-effect noise referred to grid circuit

therefore

$$E_n = \sqrt{4KT \, \Delta f \left[R_c + R_{eq} + \frac{R_b}{A^2} \left(\frac{r_p}{R_b + r_p} \right)^2 \right]}$$

PLATE SHOT-EFFECT NOISE IN MULTICOLLECTOR TUBES. A tube in which the cathode current flows to more than one electrode develops more noise than a negative grid triode because of the minute fluctuations in the division of the total cathode current between the several electrodes. This increase in plate shot-effect noise is frequently referred to as *partition noise*. The equivalent grid-noise voltage representing the plate shot-effect noise in a tube operated as a pentode is greater than the equivalent grid-noise voltage for the same tube operated as a triode. The ratio of the equivalent grid noise voltages is given by

$$\frac{E_{ng(\text{pentode})}}{E_{ng(\text{triode})}} = \sqrt{1 + \frac{8.7 \times 10^3 I_{c2}\sigma}{T_c g_{mp}}} \qquad (7.20)$$

where I_{c2} = screen current, amp

g_{mp} = g_m of tube as a pentode

Equation (7.20) is valid provided that the relation between the transconductance of the tube operated as a pentode and as a triode is given by

$$g_{mp} = g_{mt} \frac{I_b}{I_t} \qquad (7.21)$$

where I_b = plate current

I_t = total cathode current

g_{mt} = g_m of tube as a triode

Equation (7.21) is usually sufficiently accurate so that Eq. (7.20) can be used to determine the relative noisiness of a tube operated as a pentode or as a triode. If σ is again assumed to be equal to 0.88 and T_c equal to 1000°K, Eq. (7.20) reduces to

$$\frac{E_{ng(\text{pentode})}}{E_{ng(\text{triode})}} = \sqrt{1 + \frac{7.66 I_{c2}}{g_{mp}}} \qquad (7.22)$$

Since the equivalent grid noise voltage can be considered as having been generated by an equivalent resistance given by

$$R_{eq} = \frac{E_{ng}^2}{4KT \, \Delta f} \qquad (7.23)$$

the equivalent grid noise resistance R_{eq} of a pentode can be found by substituting Eqs. (7.17), (7.21), and (7.22) in Eq. (7.23) and letting $T = 293°K$, $T_c = 1000°K$, and $\sigma = 0.88$.

$$R_{eq} = \frac{I_b}{I_b + I_{c2}} \left(\frac{2.5}{g_{mp}} + \frac{19 I_{c2}}{g_{mp}^2} \right) \qquad \text{(pentode amplifier)} \qquad (7.24)$$

where I_b = plate current of tube, amp

I_{c2} = screen current of tube, amp

To illustrate the increased noise in a tube operated as a pentode compared to the same tube triode operated, refer to the type 6AC7 tube in Table 7.2. The equivalent grid resistor is 717 ohms for the tube pentode-operated and 223 ohms for the tube triode-operated.

PLATE SHOT-EFFECT NOISE IN MIXERS. In triode and pentode mixers the plate shot-effect noise component of plate current differs from the noise current present in the same tube as an amplifier only because the transconductance of the tube changes

during the local oscillator cycle. In Eq. (7.15) it is therefore necessary to substitute $\bar{g}_m$ for g_m. The term $\bar{g}_m$ is defined as the average value of transconductance throughout one cycle of the local oscillator signal. For a triode mixer, the effective plate shot-effect noise current is given by

$$I_{np} = \sqrt{\frac{0.644(4KT_c \, \Delta f \bar{g}_{mt})}{\sigma}} \qquad \text{(triode mixer)} \qquad (7.25)$$

This is the same noise current which would be produced by an r-f noise voltage E_{ng} at the grid of a triode mixer having an amplitude of

$$E_{ng} = \sqrt{\frac{0.644 \bar{g}_{mt} 4KT_c \, \Delta f}{\sigma g_c^2}} \qquad \text{(triode mixer)} \qquad (7.26)$$

where g_c = conversion transconductance

The equivalent grid noise generator for a pentode mixer is given by

$$E_{ng} = \sqrt{\left(1 + \frac{8.7 I_{c2} \sigma \times 10^3}{\bar{g}_{mp} T_c}\right) \frac{0.644 \bar{g}_{mp} I_b 4KT_c \, \Delta f}{\sigma g_c^2 I_t}} \qquad \text{(pentode mixer)} \quad (7.27)$$

where I_b, I_{c2}, I_t = values of plate, screen, and cathode current, respectively, averaged over complete cycle of local oscillator

The equivalent grid noise resistance for the triode and pentode mixers can, therefore, be expressed as

$$R_{eq} = \frac{0.644 T_c \bar{g}_{mt}}{\sigma T g_c^2} \qquad \text{(triode mixer)} \qquad (7.28)$$

$$R_{eq} = \frac{0.644 T_c \bar{g}_{mp} I_b}{\sigma T g_c^2 I_t} \left(1 + \frac{8.7 I_{c2} \sigma \times 10^3}{\bar{g}_{mp} T_c}\right) \qquad \text{(pentode mixer)} \qquad (7.29)$$

By again substituting the approximate values $\sigma = 0.88$, $T_c = 1000°K$, $T = 293°K$, Eqs. (7.28) and (7.29) can be simplified to the following forms:

$$R_{eq} = \frac{2.5 \bar{g}_{mt}}{g_c^2} \qquad \text{(triode mixer)} \qquad (7.30)$$

$$R_{eq} = \frac{I_b}{I_b + I_{c2}} \left(\frac{2.5 \bar{g}_{mp}}{g_c^2} + \frac{19 I_{c2}}{g_c^2}\right) \qquad \text{(pentode mixer)} \qquad (7.31)$$

The plate shot-effect noise component of plate current in the multigrid type of mixer tube is much greater than in triode or pentode mixers because of the relatively small fraction of total tube current that goes to the plate. The equivalent grid noise resistance for a multigrid mixer is given approximately by

$$R_{eq} = \frac{19 I_b (I_t - I_b)}{I_t g_c^2} \qquad (7.32)$$

where I_t = total cathode current

The conversion conductance of a linear diode mixer is given by

$$g_c = \frac{g}{\pi} \qquad \text{(linear diode mixer)} \qquad (7.33)$$

where g = conductance of diode

The shot-effect noise component of current in a diode whose plate current is voltage-limited is given by Eq. (7.14). From this it can be seen that the equivalent noise resistance at the input to the diode is given by

$$R_{eq} = \frac{0.644 \pi T_c}{T g_c} \qquad \text{(linear diode mixer)} \qquad (7.34)$$

To evaluate the noisiness of triode and pentode tubes as mixers, it is necessary to know the conversion transconductance g_c and the average transconductance $\bar{g}_m$ over a complete local oscillator cycle, for the particular tube in question. These values are normally not specified in tube manuals for conventional triodes and pentodes. However, approximate values of g_c and $\bar{g}_m$ can be obtained for any tube for which tube characteristics are available. If g_m is assumed to increase linearly for grid voltages above cutoff and if it is assumed that the bias on the control grid is equal to cutoff, g_c and $\bar{g}_m$ are given by

$$g_c = \frac{g_m'}{4} \tag{7.35}$$

$$\bar{g}_m = \frac{g_m'}{\pi} \tag{7.36}$$

where g_m' = peak value of transconductance occurring at peak of local oscillator cycle
By substituting Eqs. (7.35) and (7.36) into Eqs. (7.30) and (7.31) the equivalent grid noise resistance for triode and pentode tubes operating as mixers may be determined approximately.

$$R_{eq} = \frac{40}{\pi g_m'} \quad \text{(triode mixer)} \tag{7.37}$$

$$R_{eq} = \frac{I_b}{I_b + I_{c2}} \left(\frac{40}{\pi g_m'} + \frac{304 I_{c2}}{(g_m')^2} \right) \quad \text{(pentode mixer)} \tag{7.38}$$

where $I_b \simeq I_b'/4$
  $I_{c2} \simeq I_{c2}'/4$
  I_b' = value of I_b at peak of local oscillator cycle
  I_{c2}' = value of I_{c2} at peak of local oscillator cycle
The values of R_{eq} in Table 7.2 have been calculated using Eqs. (7.37) and (7.38).
   Grid-induced Shot-effect Noise. When the electrons comprising the space current approach the grid plane, a positive current is induced in the grid circuit. As these electrons travel from the grid plane to the plate, a negative current is induced in the grid circuit. The shot-effect fluctuations in cathode current are thus induced into the grid circuit. Because of the finite transit time of electrons in the tube, the induced current in the grid circuit increases linearly with frequency. Since it is the flow of electrons to the plate which causes the induced control-grid current, the noise component of grid current I_{ng} is proportional to the noise component of plate current and is approximately

$$I_{ng} = \omega \sqrt{4 K T_c \, \Delta f \left(\frac{0.644 g_m k}{\sigma} \right)} \tag{7.39}$$

where k = a constant determined by tube structure
Since a tube exhibits an input conductance at high frequencies which is a function of the same electron transit time (see Sec. 7.4h) the grid-induced noise current may also be considered as having originated in the transit-time grid conductance in accordance with

$$I_{ng} = \sqrt{\frac{1.43(4 K T_c \, \Delta f)}{R_t}} \tag{7.40}$$

where R_t = grid input resistance in ohms due to electron transit-time effects (see Sec. 7.4h)
The equivalent noise circuit for a tube including the effects of both plate shot-effect

noise and grid-induced shot-effect noise is shown in Fig. 7.10.   In an actual circuit it is also necessary to consider the effects of thermal noise in the grid circuit impedance (see Fig. 7.11).   If the total grid circuit impedance including the grid resistance due to electron transit-time effects is expressed as Z_g, the corresponding grid noise voltage due to grid-induced shot-effect noise current is

$$E_{ng} = Z_g I_{ng} = Z_g \sqrt{\frac{1.43\,(4KT_c\,\Delta f)}{R_t}} \qquad \text{(amplifiers)} \qquad (7.41)$$

The plate noise current I_{np} by this grid voltage is then

$$I_{np} = g_m Z_g \sqrt{\frac{1.43\,(4KT_c\,\Delta f)}{R_t}} \qquad \text{(amplifiers)} \qquad (7.42)$$

Equations (7.41) and (7.42) are valid for tubes used as mixers provided that R_t is

$$\mathcal{E}_n = \sqrt{4KT\Delta f R_{eq}}$$

$$I_{ng} = \sqrt{\frac{1.43 \times 4KT_c \Delta f}{R_t}}$$

R_t

FIG. 7.10. Complete shot-effect noise equivalent circuit for a vacuum tube.

averaged over a complete local oscillator cycle and that the mixer conversion transconductance g_c is used instead of g_m in Eq. (7.42).

At low frequencies the current induced in the grid circuit is in quadrature with the applied signal voltage and the plate signal current.   This quadrature current presents an effective capacitance between the grid and cathode.   At higher frequencies where the time of transit of the electrons from the cathode to the grid requires an appreciable fraction of a cycle of the applied signal, the induced grid current contains an in-phase component as well as a quadrature component.   This in-phase component of current presents the effective resistive loading R_t to the applied signal.   If the external grid impedance is resistive, the plate-current noise due to grid-induced shot-effect noise is nearly 90° out of phase with the original fluctuations in tube current which produced it, and the total plate current noise is equal to the square root of the sum of the squares of the plate shot-effect noise current and the current due to grid-induced shot-effect noise.   At very high frequencies the phase of the induced grid current relative to the applied signal will differ from 90° by an amount great enough to cause an appreciable increase in the plate-current noise due to the direct addition of a portion of the grid-induced shot-effect noise with the plate shot-effect noise if Z_g is a pure resistance.   By proper adjustment of the phase angle of Z_g it is theoretically possible to reduce the plate-current shot noise by partial cancellation of the two noise components.

A comparison of the equations for plate shot-effect noise and grid-induced shot-effect noise [Eqs. (7.15) and (7.42)] shows that at high frequencies the effect of the grid-induced shot-effect noise may exceed the plate shot-effect noise itself.

2. Noise Due to Flicker Effect.   The electron emission from the surface of the cathode is continually varying over the entire cathode surface because of evaporation, diffusion, impurities, gas-ion bombardment, and similar causes.   These variations in emission give rise to a relatively low-frequency, e.g., 0 to 500 cps, type of noise known as *flicker-effect noise*.   At low audio frequencies the magnitude of this noise may exceed by several times the magnitude of shot-effect noise.   Flicker-effect noise is especially pronounced in tubes having oxide-coated cathodes.

Flicker-effect noise, in contrast to shot-effect noise, can be reduced by careful preparation of the cathode and, therefore, does not place a fundamental limitation upon amplifier sensitivity.

3. Secondary-emission Noise. When an electron or ion strikes a surface with sufficient velocity it will knock out one or more of the electrons contained in the surface. This phenomenon is known as secondary emission since the electrons are emitted by the surface only when it is struck with other electrons or ions. The ratio of secondary electrons to primary electrons is a function of the type of surface and the velocity of the primary electrons and may vary from less than one to about three for pure metal surfaces and as high as five to ten for surfaces composed of complex alkali-metal compounds. The ratio S of secondary to primary electrons can be expressed as

$$S = \frac{\text{average number of secondary electrons}}{\text{average number of primary electrons}}$$
$$= \text{current amplification of emitting electrode} \qquad (7.43)$$

Fluctuations in the emission of secondary electrons result in a noise component of tube current in a manner similar to thermionic emission of electrons from a hot cathode. If there is no space charge formed by these electrons, as is the case in electron photomultiplier tubes and other tubes using the emission of secondary electrons as a means of current amplification, the noise component of the current developed at the secondary emission electrode is

$$I_{ns} = \sqrt{2eI_s\,\Delta f} \qquad (7.44)$$

where I_{ns} = rms noise current due to randomness of secondary emission, amp
e = charge of one electron = 1.60×10^{-19} coulomb
I_s = average secondary emission current, amp
Δf = width of frequency band in cycles per second in which noise current is measured

Since the current amplification of the electrode is S, the noise component of the current striking the electrode is amplified by a factor of S. The total rms noise current I_n in the secondary electron current is given by

$$I_n = \sqrt{I_{pn}{}^2 S^2 + 2eI_a\,\Delta f S} \qquad (7.45)$$

where I_{pn} = rms noise current in primary electron current
I_a = average primary electron current

If the primary electron noise current is given by Eq. (7.44), as is the case for a photomultiplier tube, Eq. (7.45) becomes

$$I_n = \sqrt{2eI_a\,\Delta f(S^2 + S)} \qquad (7.46)$$

for the noise current from the first secondary-emission electrode. If n such secondary-emission electrodes are cascaded, the total noise current at the anode is

$$I_n = \sqrt{2eI_a\,\Delta f S^n(S^n + S^{n-1} + S^{n-2} + \cdots + S + 1)} \qquad (7.47)$$

Since the anode current is S^n times the original primary electron current, Eq. (7.47) may be rewritten as

$$I_n = \sqrt{2eI_b\,\Delta f \left(\frac{S^{n+1} - 1}{S - 1}\right)} \qquad (7.48)$$

where I_b = average anode current, amp

4. Positive-ion Noise. The electrons which pass from the cathode to the plate may strike any residual gas molecules with sufficient velocity to release any free electrons contained in the gas molecule. The released electrons travel on to the plate and add to the plate current. The positive ions formed by the gas molecules losing an electron will be attracted to the grid. The result is that the flow of electrons

through the external grid circuit to the grid to neutralize the gas ions will develop a positive voltage across the grid circuit impedance. Since the collision of gas molecules with electrons is of purely random occurrence, the grid and plate currents created by these collisions constitute sources of random noise. The amount of noise generated is a function of the number of positive ions formed, which in turn is dependent upon the tube space current and the number of gas molecules present in the tube. Since positive ions are collected by the grid, the amount of noise generated by positive-ion formation can be related to the average grid current caused by the positive ions. The equivalent grid resistor R'_{eq} which will develop an amount of noise in the plate circuit of a tube equal to the noise developed by positive-ion formation is given by[1]

$$R'_{eq} = \left(20R_c{}^2 + \frac{4I_b \times 10^4}{g_m{}^3}\right) I_c \qquad (7.49)$$

where R_c = grid circuit resistance, ohms
  I_b = average plate current, amp
  I_c = average grid current due to positive ions, amp
  g_m = grid-plate transconductance, mhos

To calculate the total equivalent grid noise voltage due to both plate shot-effect noise and positive-ion noise, the equivalent grid resistors R_{eq} and R'_{eq} must be placed in series. The noise component of plate current due to positive-ion noise is given by

$$I_{np} = g_m \sqrt{4KT\,\Delta f I_c \left(20R_c{}^2 + 4 \times 10^4 \frac{I_b}{g_m{}^3}\right)} \qquad (7.50)$$

7.2i. Magnetic-fluctuation Noise. Magnetic substances are not uniform and continuous in structure but rather are composed of many elementary magnetic particles. When such a material is subjected to a magnetizing force, the individual particles exhibit fluctuations in their rate of magnetization. These fluctuations are random in their occurrence, and therefore, when the material is being magnetized and demagnetized, there is a random fluctuation in the flux density produced in the material. If the flux from this magnetic material is linked by a coil, noise voltages will be induced in the coil corresponding to the random flux variations present in the magnetic material. The noise thus produced is known as *magnetic-fluctuation noise.* It is also referred to as the *Barkhausen effect.*

An example of magnetic fluctuation noise in a receiver might be noise voltages induced in a loop antenna by the core flux from a power transformer. The noise thus induced in an antenna may limit the sensitivity of a receiver.

7.2j. Signal-to-noise Ratio and Noise Figure. The relation of signal amplitude to noise amplitude is commonly referred to as the *signal-to-noise* ratio S/N. This ratio has not been well standardized and is used interchangeably to mean the ratio of rms signal voltage to rms noise voltage, the ratio of peak signal voltage to peak noise voltage, or, as in pulse radar systems, the ratio of peak signal power to average noise power. Therefore, when the term S/N is used, care must be exercised to determine exactly which ratio is being referred to.

The signal-to-noise ratio in itself does not specify the effect of the noisiness of a single stage, an amplifier, or a receiver upon a signal and, consequently, cannot be used alone as a measure of equipment performance. A measure of the equipment-noise quality which has received widespread acceptance in recent years is the *noise figure* of the equipment. The use of noise figure as a method of specifying network noisiness originated with H. T. Friis.[2] Noise figure compares directly the actual

[1] K. Spangenberg, "Vacuum Tubes," McGraw-Hill Book Company, Inc., New York, 1948.
[2] *Proc. IRE,* July, 1944, p. 419.

noise obtained from a circuit with the noise which would be obtained if the circuit were free of internal noise sources. The noise figure F of a network is defined as

$$F = \frac{S_i/N_i}{S_o/N_o} \qquad (7.51)$$

where S_i = available input signal power
N_i = ideal available input noise power
S_o = available output signal power
N_o = available output noise power
S_i is defined as the maximum available signal power which can be transferred from the source to a conjugate load. Thus, if a source has an open circuit terminal voltage of E_s and the impedance of the source is

$$Z_s = R_s + jX_s \qquad (7.52)$$

the available input signal power S_i to a load of impedance $Z_L = R_s - jX_s$ is

$$S_i = \frac{E_s^2}{4R_s} \qquad (7.53)$$

This does not necessarily represent the actual power transferred to the network whose noise figure is defined by Eq. (7.51) since the input impedance of the network is not necessarily the conjugate of the generator impedance.

N_i is defined as $KT \, \Delta f$. The temperature of the ideal generator is usually taken to be 293°K.

S_o is defined as the maximum available output signal power which can be transferred to a conjugate load under the particular matching conditions existing at the input to the network.

N_o is defined as the maximum available output noise power which can be transferred to a conjugate load under the particular matching conditions existing at the input to the network.

The noise figure of a network is independent of the load to which the network is connected in any particular circuit, but is dependent upon the conditions of matching present between the generator and the input of the network. The noise figure of a network can be written as

$$F = \frac{S_i/KT \, \Delta f}{S_o/N_o} = \frac{N_o}{GKT \, \Delta f} \qquad (7.54)$$

where $G = S_o/S_i$ = available power gain
The noise figure of a network is the ratio of the actual noise power available at the output to the noise power available at the output of a noise-free network having the same characteristics. Since the noise figure of a noiseless network is 1, the component of the network noise figure due to internally generated noise is $F - 1$.

The noise figure of two networks connected in cascade is determined from the individual network noise figures by

$$F_{ab} = F_a + \frac{F_b - 1}{G_a} \qquad (7.55)$$

where F_{ab} = over-all noise figure of networks a and b in cascade
F_a = noise figure of network a
F_b = noise figure of network b
G_a = available power gain of network a
The development of this equation is shown below. The available noise output power N_{ab} from networks a and b in cascade is given by

$$N_{ab} = F_{ab}KT \, \Delta f G_{ab} \qquad (7.56)$$

where $\Delta f = \Delta f_a = \Delta f_b$, that is, both networks have identical bandwidths

G_{ab} = available power gain of stages a and b in cascade = S_{ab}/S_g

S_{ab} = available output signal power

S_g = available input signal power

The available noise power at the output of network b has a component N_b' due to network a equal to

$$N_b' = (F_a G_a KT\ \Delta f)G_b \qquad (7.57)$$

where G_b = available power gain of network b

The available output noise power also has a component N_b'' due to the noise generated internally in network b equal to

$$N_b'' = (F_b - 1)G_b KT\ \Delta f \qquad (7.58)$$

$$N_{ab} = N_b' + N_b'' = F_a G_a G_b KT\ \Delta f + (F_b - 1)G_b KT\ \Delta f \qquad (7.59)$$

or

$$N_{ab} = \left(F_a + \frac{F_b - 1}{G_a}\right) G_a G_b KT\ \Delta f \qquad (7.60)$$

Substituting Eq. (7.60) into (7.56) and noting that

$$G_{ab} = G_a G_b$$

the over-all noise figure is seen to be

$$F_{ab} = F_a + \frac{F_b - 1}{G_a} \qquad (7.55)$$

Example 7.5

Determine the noise figure of a grounded cathode triode amplifier if it is connected to a signal source having an impedance R_s (see Fig. 7.11). Consider the signal frequency to be

E_{n1} = NOISE VOLTAGE DEVELOPED ACROSS SOURCE RESISTANCE

E_{n2} = EQUIVALENT GRID RESISTOR NOISE VOLTAGE

I_{n3} = INDUCED GRID SHOT-EFFECT NOISE CURRENT

(a) CIRCUIT OF AMPLIFIER

(b) EQUIVALENT CIRCUIT SHOWING SIGNAL AND NOISE SOURCES WHICH DETERMINE AMPLIFIER NOISE FIGURE
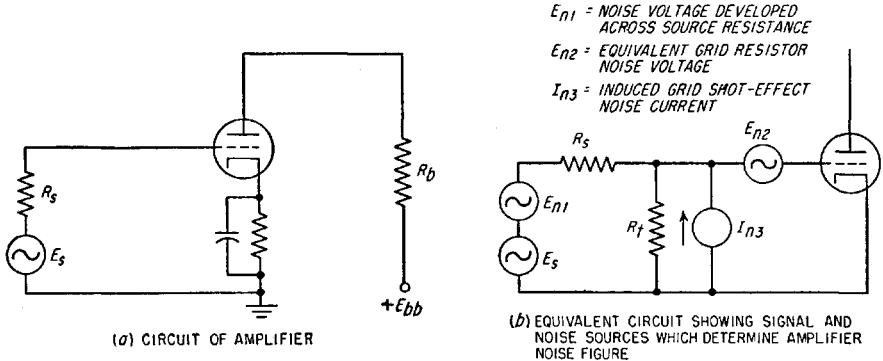
FIG. 7.11. Equivalent noise circuit of grounded-cathode amplifier.

high enough that the transit-time grid loading must be considered and assume that all circuit reactances are tuned out at the signal frequency. Assume that only plate shot-effect noise, grid-induced shot-effect noise, and thermal noise need be considered.

1. The available input signal power is given by

$$S_i = \frac{E_s^2}{4R_s}$$

2. The available output signal power S_o is that power delivered to a plate load resistance equal to r_p.

$$S_o = \frac{\mu^2 E_s^2}{4r_p}\left(\frac{R_t}{R_t + R_s}\right)^2$$

where R_t = resistive component of grid input impedance due to transit-time effects
  3. Determine the ideal available input noise power.
  The ideal available input noise power is equal to $KT\,\Delta f$.
  4. Determine the available output noise power.
  The sources of noise for the tube are shown in Fig. 7.11b. The square of the total noise current I_n flowing through the parallel combination of R_s and R_t is given by

$$I_n{}^2 = I_{n3}{}^2 + I_{n1}{}^2$$

where I_{n1} = value of current noise generator obtained by transforming voltage generator
         of voltage E_{n1} and source resistance R_s to a current generator
  I_{n3} = induced grid shot-effect noise current (See Eq. 7.40)

$$I_n{}^2 = \frac{1.43(4KT_c\,\Delta f)}{R_t} + \frac{4KT\,\Delta f}{R_s}$$

Let $1.43T_c$ be represented as T_c'.
Then,

$$I_n{}^2 = 4KT\,\Delta f\left(\frac{T_c'}{TR_t} + \frac{1}{R_s}\right)$$

The noise voltage squared developed across R_t and R_c in parallel is then given by

$$E_n{}^2 = 4KT\,\Delta f\left(\frac{T_c'}{TR_t} + \frac{1}{R_s}\right)\left(\frac{R_tR_s}{R_t + R_s}\right)^2$$

The total effective grid-circuit noise voltage squared is

$$E_{nt}{}^2 = 4KT\,\Delta f\left(\frac{T_c'}{TR_t} + \frac{1}{R_s}\right)\left(\frac{R_tR_s}{R_t + R_s}\right)^2 + 4KT\,\Delta fR_{eq}$$

The plate circuit of the tube may be represented as an equivalent noise generator E_{no} in series with r_p where the generator voltage squared is given by

$$E_{no}{}^2 = \mu^2\left[4KT\,\Delta f\left(\frac{T_c'}{TR_t} + \frac{1}{R_s}\right)\left(\frac{R_tR_s}{R_t + R_s}\right)^2 + 4KT\,\Delta fR_{eq}\right]$$

The available output noise power N_o is given by

$$N_o = \frac{E_{no}{}^2}{4r_p}$$

  5. Determine the noise figure from Eq. (7.54).

$$F = \frac{S_i/KT\,\Delta f}{S_o/N_o} = \frac{S_i}{S_o}\frac{N_o}{KT\,\Delta f}$$
$$= 1 + \frac{R_sT_c'}{R_tT} + \frac{R_{eq}}{R_s}\left(\frac{R_s}{R_t} + 1\right)^2$$

The determination of the noise figure of a receiver makes it possible to calculate the minimum detectable input signal $S_{\min}$ to the receiver provided the signal-to-noise ratio at the output associated with the minimum usable output signal has been determined. This relationship is given by

$$S_{\min} = F_rKT\,\Delta f\frac{S}{N} \tag{7.61}$$

where S/N = minimum usable output signal-to-noise ratio
  F_r = noise figure of over-all receiver
Equation (7.61) is valid provided that the receiver noise figure is determined for a generator having an internal impedance equal to that of the actual signal source.
  7.3. Superheterodyne Receivers. The superheterodyne receiver is the most commonly used type of radio or radar receiver. Although it is somewhat more com-

plicated than other types, it is, in general, capable of greater sensitivity and selectivity. A block diagram of a superheterodyne receiver is shown in Fig. 7.12. The basic principle is the conversion of the r-f signal to an intermediate frequency by heterodyning the r-f signal with a local oscillator whose frequency differs from that of the r-f signal by an amount equal to the desired intermediate frequency. Tuning the receiver consists of simultaneously tuning the local oscillator, mixer, and any r-f amplifier stages so that the intermediate frequency remains constant. The principal advantage of such a receiver is that most of the amplification of the signal can be accomplished in the i-f amplifier stages which do not have to be tuned and which usually operate at a lower frequency than the r-f amplifiers. This results in the sensitivity and selectivity being more nearly constant throughout the tuning range of the receiver,
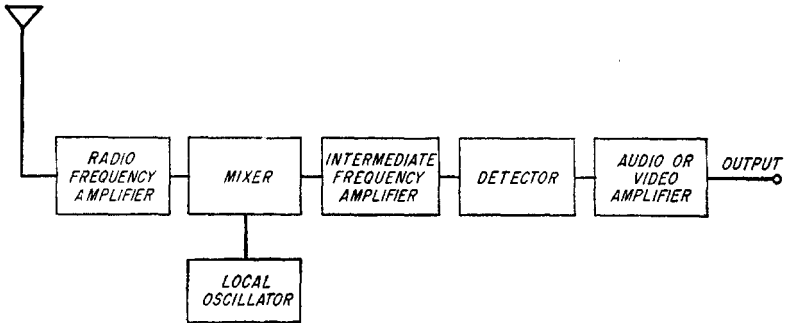
FIG. 7.12. Block diagram of a superheterodyne receiver.

and for practical reasons a greater gain per stage is realizable at intermediate frequencies than at radio frequencies.

7.4. R-F and I-F Amplifiers. The major portion of the amplification in a superheterodyne receiver is accomplished in tuned amplifiers resonant at the signal frequency and/or an intermediate frequency. Receiver sensitivity, selectivity, and image rejection are largely determined by the characteristics of these amplifiers.

The r-f amplifier must be capable of being tuned over the entire band of frequencies to be covered by the receiver. The i-f amplifier is fixed-tuned and usually tuned to a lower frequency than the r-f amplifier. The intermediate frequency of a particular receiver, however, may be much higher than the radio frequency of another receiver, e.g., microwave radar and television receivers compared to broadcast and short-wave receivers.

7.4a. General Requirements for R-F Amplifiers. A superheterodyne receiver does not necessarily contain an r-f amplifier. However, improvements in the image rejection can be achieved if the mixer is preceded by a frequency selective element such as one or more tuned stages of r-f amplification. In addition, the receiver noise figure can usually be improved if an r-f amplifier is employed. An exception to this is a receiver operating in the region of 4,000 Mc or higher where the noise figures of existing r-f amplifiers are higher than the noise figures obtainable in crystal mixers. Consequently, at these frequencies the over-all receiver noise figure will be increased by the addition of an r-f amplifier.

1. Image Rejection. The mixer in a superheterodyne will produce an output signal at the intermediate frequency f_{if} when the r-f input signal f_{rf} differs from the frequency of the local oscillator f_{lo} by the intermediate frequency. Thus, for a specific local-oscillator frequency, there are two frequencies at which an r-f signal will produce an i-f output.

$$f_{rf} = f_{lo} \pm f_{if} \tag{7.62}$$

One of these frequencies is the desired signal frequency and the other is the *image frequency*. When the receiver is tuned to a desired signal, a signal appearing at the image frequency will interfere with the desired signal unless it is sufficiently attenuated in some manner. The suppression of signals at the image frequency is accomplished by the use of filter networks. Networks of this type usually take the form of bandpass filters, e.g., one or more r-f stages or a resonant cavity tuned to the signal frequency. The bandwidths of these elements are made sufficiently narrow that the image frequency falls outside the passband and is appreciably attenuated.

If an inductively coupled circuit having a resonant secondary circuit is used to couple the antenna to the grid of the converter stage, the loading of the circuit by the
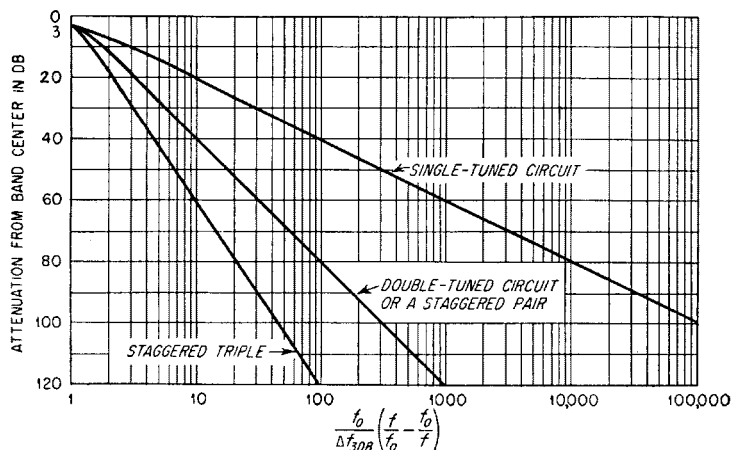
Fig. 7.13. Selectivity of bandpass coupling circuits.

antenna is usually so great that the desired narrow passband cannot be achieved unless a very small degree of coupling is used. A small degree of coupling, however, reduces the amplitude of the signal available at the converter grid. In fixed-tuned receivers, it is practical to use more complicated coupling circuits to achieve an impedance transformation providing a narrow passband and close coupling. In tuned receivers, however, it is usually necessary to provide one or more tuned r-f amplifier stages between the antenna and the converter input to obtain the desired close coupling and narrow passband. The degree of image rejection which can be achieved in a single r-f amplifier stage having a parallel resonant plate circuit is plotted in Fig. 7.13 as a function of the Q of the circuit and the ratio of the image frequency to the desired frequency. The increased selectivity which can be obtained by more complicated networks is also shown in Fig. 7.13. For values of relative attenuation at frequencies which are outside the circuit passband by approximately three or more octaves, the following equation can be used:

$$A_{db} = 20N \log_{10} \left[\left(\frac{f_o}{\Delta f_{3db}} \right) \left| \frac{f}{f_o} - \frac{f_o}{f} \right| \right] \qquad (7.63)$$

where f = signal frequency, cps

f_o = resonant frequency of tuned circuit, cps

Δf_{3db} = bandwidth of circuit between half-power frequencies

N = 1 for single-tuned circuits, 2 for double-tuned circuits or staggered pairs, and 3 for staggered triples

If two or more resonant circuits are used in cascade, the relative attenuation of the individual circuits can be added to yield the over-all relative attenuation.

2. Noise Figure. The equivalent grid-noise resistance R_{eq} of any tube operated as a mixer [see Eqs. (7.30) and (7.31)] is always larger than the equivalent grid-noise resistance of the same tube operated as an amplifier [see Eqs. (7.19) and (7.24)]. The reason is that the grid-plate transconductance of a tube is greater than the conversion transconductance of the same tube by a factor of approximately 4. The noise figures of pentagrid and other multielement mixer tubes are correspondingly greater than triode or pentode amplifiers because of the very low conversion transconductances of these tubes and the low ratios of plate current to total cathode current.

At frequencies low enough that the electron transit-time effects can be ignored, the noise figure for either a grounded-cathode amplifier (see step 5 of Example 7.5) or a grounded-cathode mixer is given by

$$F = 1 + \frac{R_{eq}}{R_s} \tag{7.64}$$

where F = noise figure expressed as a power ratio
R_s = source resistance

Equation (7.64) illustrates how the noise figure of a vacuum tube varies as a function of R_{eq}. The improvement in the over-all noise figure of a superheterodyne receiver by the addition of one or more stages of r-f amplification ahead of the mixer can be determined from Eq. (7.65).

$$F_{12} = F_1 + \frac{F_2 - 1}{G_1} \tag{7.65}$$

where F_1 = noise figure of r-f amplifier expressed as a power ratio
F_2 = noise figure of mixer expressed as a power ratio
G_1 = available power gain of r-f amplifier
F_{12} = over-all noise figure of r-f amplifier and mixer expressed as a power ratio

At sufficiently high frequencies, electron transit-time effects become so detrimental that crystal diode mixers surpass vacuum-tube amplifiers and mixers in regard to noise figure.

7.4b. General Requirements of I-F Amplifiers. The desired characteristics of the i-f amplifier are determined by the receiver application and the characteristics of the other sections of the receiver. The requirements which must be determined before the design of the i-f amplifier can be completed are:

1. Gain. The maximum voltage gain required must be determined in order that the number of stages in the i-f amplifier can be established. The maximum gain in the i-f amplifier is normally determined by the amplitude of the smallest mixer output signal which must be amplified and the signal level required at the second detector to provide linear detection. In an actual design, surplus gain should be provided to allow for tube aging, tube variations, and nonoptimum circuit alignment.

2. Bandwidth. Because the circuit element Q's are lower for a given bandwidth in a low-frequency amplifier than in a high-frequency amplifier, and because it is impractical to cascade several variable frequency amplifiers, the bandwidth and selectivity characteristics of the receiver are usually established by the i-f amplifier. The bandwidth of the i-f amplifier is ordinarily selected to make the signal-to-noise ratio as large as possible.

3. Noise Figure. The noise figure of two cascaded circuits is given by Eq. (7.65). If the over-all noise figure of the r-f amplifier and mixer is considered to be F_1 and the noise figure of the i-f amplifier to be F_2, the total receiver noise figure will ordinarily be very nearly equal to F_{12} since the gain of the i-f amplifier will usually be great enough that any succeeding portions of the receiver will have negligible effect on

the over-all receiver noise figure. Also, if the gain G_1 of the r-f amplifier and mixer is sufficiently large to make the second term in Eq. (7.65) negligible, the noise figure of the i-f amplifier will not contribute significantly to the over-all receiver noise figure.

4. Selectivity. If the receiver is required to have good selectivity, i.e., to reject signals close in frequency to the desired signal, the skirts of the i-f amplifier passband must have large slopes. In some instances this requirement may necessitate special interstage coupling networks.

5. Stability. The i-f amplifier must not be regenerative. If the amplifier is close to the point of oscillation, serious distortion of the amplifier passband can occur even though the amplifier does not actually oscillate. Also, if automatic gain control is applied to the i-f amplifier, the changes in gain of the individual stages can cause the input capacitance of these stages to change sufficiently that significant detuning will result unless compensating design techniques are employed (see Sec. 7.4h).

6. Limiting and Overloading. The signal-handling capabilities of the i-f amplifier must be adequate to handle the largest expected mixer output signal without limiting or overloading. The amplifier must also have enough gain to amplify the smallest signal to a usable level at the second detector. These requirements usually dictate the inclusion of some form of automatic gain control wherein the gain of the i-f amplifier is an inverse function of the input signal level.

7. Amplitude and Phase Response. In addition to the requirement that the important sidebands of the signal be amplified without undue changes in relative amplitude, there often exists the requirement that the relative phases of the sideband components be very nearly preserved. These factors must be considered in i-f amplifiers where the rise time and overshoot of a pulse are of importance.

8. Intermediate Frequency. There are numerous factors which affect the selection of the intermediate frequency. Some of these factors favor the selection of a high frequency and others favor the selection of a low frequency. The final selection of the frequency is always a compromise between these factors, and the relative merits of each must be examined for each receiver to determine the optimum.

Factors favoring a low intermediate frequency:

NOISE FIGURE. The noise figure of any type of tube increases with frequency. The increase in noise figure is caused by grid-induced shot-effect noise power which increases as the square of frequency.

INPUT LOADING. The input resistance of a tube varies inversely as the square of frequency due to both cathode lead inductance and electron transit time. At high frequencies this effect limits the available power gain of a tube.

NARROW BANDWIDTHS. The Q of a circuit is equal to the ratio of the resonant frequency to the bandwidth between the -3-db frequencies. For given coil Q's, narrower bandwidths are achievable at low frequencies.

Factors favoring a high intermediate frequency:

IMAGE REJECTION. The higher the intermediate frequency the larger will be the absolute frequency difference between the desired and image frequencies, which, in turn, increases the achievable rejection of the image signal in the r-f amplifier.

SECOND-DETECTOR FILTER DESIGN. If the detected modulation from the received signal has frequency components which approach the intermediate frequency, a sharp cutoff low-pass filter must be used at the output of the second detector in order to pass the high-modulation components and attenuate the ripple at the intermediate frequency. The use of a high intermediate frequency, therefore, eases the detector filter-design requirements. A considerable reduction in the amplitude of the i-f signal at the output of the second detector can be obtained by employing a full-wave detector since the frequency of the principal output ripple component will then be equal to twice the intermediate frequency.

In those instances where the image rejection of a receiver must be good and the

receiver passband very narrow, e.g., uhf receivers, the desired characteristics can be achieved by the use of a double superheterodyne receiver. This consists of a mixer and local oscillator followed by an i-f amplifier having a high intermediate frequency which is followed by a second mixer and local oscillator which feed a second i-f amplifier having a low intermediate frequency. Because of the high center frequency of the first i-f amplifier, good image rejection can be obtained in the r-f amplifier. The low center frequency of the second i-f amplifier makes it possible to obtain a narrow band-width in the over-all receiver. Another advantage of a double superheterodyne is that greater over-all i-f gain can be obtained without regeneration than in a conventional superheterodyne receiver because of the frequency isolation between the two i-f amplifiers.

Fig. 7.14. Tuned amplifier circuit.

7.4c. Selection of Tube Type. The type of tube to be used in an r-f or i-f amplifier is determined by a number of factors, the most important of which are discussed below:

1. Gain-bandwidth Product. The product of the voltage gain and bandwidth of an amplifier stage is a function of the type of tube used. Consider the amplifier shown in Fig. 7.14. If R_c is much smaller than both r_p and Q^2R_L, where R_L is the series resistance of the coil, the voltage gain A of the first stage at the frequency of resonance is essentially

$$A = -g_m R_c \qquad (7.66)$$

and the -3-db bandwidth Δf is

$$\Delta f = \frac{1}{2\pi R_c C_t} \qquad (7.67)$$

where $C_t =$ total shunt capacitance in interstage circuit including tube and stray capacitances

The product of the stage gain and bandwidth is given by Eq. (7.68).

$$A\,\Delta f = \frac{-g_m}{2\pi C_t} \qquad (7.68)$$

The gain-bandwidth product establishes a figure of merit for tubes used as amplifiers, and for the maximum gain-bandwidth product, g_m should be as large as possible and C_t as small as possible. In determining the gain-bandwidth product for any tube type, identical tubes are assumed to be in cascade to establish the tube input and output capacitances which are contained in the total shunt capacitance C_t. In Table 7.3, the gain-bandwidth product is given for several types of commonly used pentodes for the case where C_t is equal to $c_i + c_o$ and where C_t is equal to $c_i + c_o + 5\ \mu\mu f$. The quantity $5\ \mu\mu f$ very nearly represents the minimum possible value of the socket and wiring capacitances. In the calculation of the gain-bandwidth product from Eq. (7.68) for Table 7.3, the value of g_m has been expressed in micromhos and C_t in micromicrofarads. The gain-bandwidth values are, therefore, in megacycles.

In narrow-band receivers, such as broadcast and communications receivers, the gain-bandwidth product is of little significance. This is the case since the narrow-band high-selectivity response curves are obtained by paralleling the tube input and output capacitances with tuning capacitors which are many times larger than the tube capacitances themselves. For this particular application, the g_m of the tube has more significance than the values of c_i and c_o.

2. Stability. The grid-plate capacitance of a grounded-cathode amplifier stage provides a feedback path between the grid and plate circuits which may cause oscillation (see Sec. 7.4*i*). The maximum gain which can be obtained in one stage without regeneration is inversely proportional to the square root of the product of c_{gp} and frequency [see Eq. (7.124)]. Triodes are seldom used in grounded-cathode r-f or i-f amplifiers except as low-noise input stages since they usually must be neutralized to prevent oscillation and because of the appreciable variations in input capacitance due to Miller effect.

TABLE 7.3. GAIN-BANDWIDTH PRODUCTS FOR VARIOUS PENTODE TUBES

Tube type	g_m, μmhos		c_i, μμf		c_o, μμf		$\dfrac{g_m}{2\pi(c_i + c_o)}$ Mc	$\dfrac{g_m}{2\pi(c_i + c_o + 5)}$ Mc
	Min.	Max.	Min.	Max.	Min.	Max.		
6AC7	7000	11,000	8.8	13.2	3.5	6.5	89	76
6AG5	4000	6,000	5.2	7.8	1.3	2.3	96	60
6AG7	9200	14,200	11.5	14.5	6.5	8.5	91	73
6AH6	6000	11,000	8.0	12.0	2.5	4.7	99	73
6AK5	3500	6,500	3.4	4.4	2.5	3.3	118	68
6AN5	6000	10,000	6.0	12.0	3.0	6.0	94	69
6AU6	4150	6,250	4.4	6.6	3.5	6.5	79	53
6CB6		6,200*		6.3*		1.9*	120	75
6BA6	3600	5,200	4.4	6.4	3.5	6.5	74	49
5840	4200	5,800	3.5	4.9	2.9	3.9	105	59
6205	4200	5,800	3.5	4.9	2.9	3.9	105	59

* Nominal.

3. Noise Figure. If the noise figure of the r-f or i-f amplifier is important in determining the over-all receiver noise figure, special consideration should be given to the tube type used in the input stage of the amplifier. In general, the input tube should be a triode or triode-connected pentode having a high g_m and a low-input conductance due to electron transit time. Also, the output impedance of the network preceding the input tube should be adjusted for optimum noise figure (see Sec. 7.4*i*).

4. Circuit Loading. The input conductance of a tube limits the maximum gain which can be achieved in the preceding amplifier stage. The tube selected for use in an r-f or i-f amplifier must have an input resistance large enough to allow the desired gain in the preceding stage to be realized.

5. Signal Amplitude. An item which must be considered in the selection of tube types for an i-f amplifier is the maximum signal amplitude which the last stage of the amplifier must amplify without serious distortion. A receiver is normally required to operate satisfactorily with a large dynamic range of input signals, and in most cases some form of automatic gain control (AGC) is employed to maintain the output signal level relatively constant. If AGC is not used, a large input signal may result in a signal at the grid of the last i-f amplifier stage which will swing the grid both into the positive grid region and to cutoff. If AGC is used, the stages to which the compensating bias is applied should have transfer characteristics essentially linear for the largest signal amplitudes present. This requirement is usually met by the use of remote-cutoff pentodes (Fig. 7.15*b*).

6. Tube Cost. A tube having a high g_m and low interelectrode capacitances costs more to manufacture than a tube having a low g_m and high interelectrode capacitances because of the more critical spacing of tube elements. If the number of tubes in an

i-f or r-f amplifier can be reduced by the use of tubes having a large gain-bandwidth product, the increased tube cost will often be more than compensated for by the elimination of additional tubes and parts.

(a) DISTORTION WITH SHARP CUTOFF PENTODE

(b) ELIMINATION OF DISTORTION BY USE OF REMOTE CUTOFF PENTODE
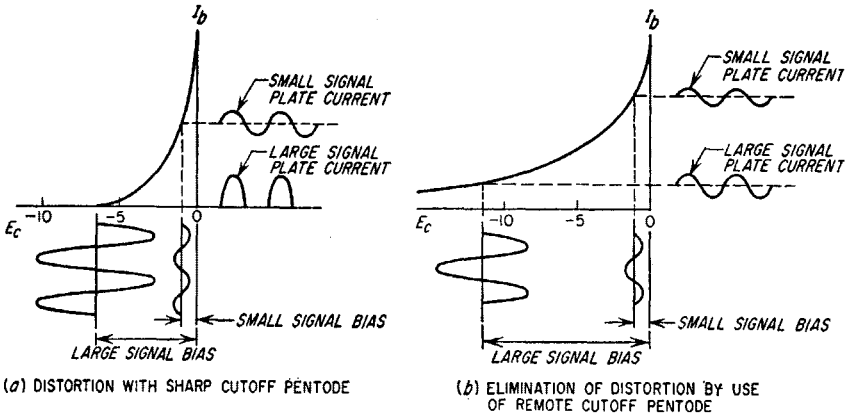
FIG. 7.15. Prevention of large signal distortion by use of AGC bias and a remote cutoff pentode.

7.4d. Synchronous Single-tuned Amplifier Stages. The circuit of a single-tuned amplifier is shown in Fig. 7.14, and its equivalent circuit is shown in Fig. 7.16. An amplifier consisting of several single-tuned stages in cascade and all tuned to the same frequency is said to be synchronously tuned. If a high-Q coil is used, the voltage gain of a single-tuned stage at resonance is given by Eq. (7.69).

$$A = \frac{-g_m Q_L \omega_o L}{1 + (\omega_o L Q_L / r_p) + (\omega_o L Q_L / R_c)} \qquad (7.69)$$

where $\omega_o = 1/\sqrt{LC_t}$

$Q_L = Q$ of inductance, that is, $Q_L = \omega_o L / R_L$

r_p = tube dynamic plate resistance

R_c = loading resistance in shunt with resonant circuit

C_t = total shunt capacitance across L

In practice, the amplifier tubes are usually pentodes and the term $\omega_o L Q / r_p$ can be neglected in Eq. (7.69).

1. Narrow-band Single-tuned Amplifiers. Pentodes are usually used in narrow-band amplifiers, and the grid resistor R_c of the following stage is made very large so that its effect upon the resonant circuit can be neglected. For these conditions and assuming that the Q of the inductance is very large, the voltage gain is given by Eq. (7.70).
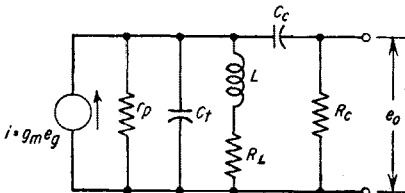
FIG. 7.16. Equivalent circuit of single-tuned amplifier.

$$A(f) = \frac{-g_m Q_o^2 R_L}{\sqrt{1 + Q_o^2 (f/f_o - f_o/f)^2}} \qquad (7.70)$$

where $Q_o = Q_L$

At midband, Eq. (7.70) reduces to

$$A = -g_m Q_o^2 R_L \qquad (7.71)$$

The bandwidth of a circuit is defined as the band included between the frequencies where the power has dropped to one-half (-3 db) of the value at resonance, or where

the voltage response is 0.707 of its value at resonance. The bandwidth Δf of a stage whose gain is given by Eq. (7.70) is given approximately by Eq. (7.72).

$$\Delta f \simeq \frac{f_o}{Q_o} = \frac{f_o}{Q_L} = \frac{R_L}{2\pi L} \tag{7.72}$$

The over-all bandwidth Δf_t of n identical high-Q synchronous single-tuned circuits in cascade is given by Eq. (7.78). The ratio $\Delta f_t/\Delta f$ for several values of n has been tabulated in Table 7.4.

TABLE 7.4. REDUCTION IN BANDWIDTH OF CASCADED IDENTICAL SYNCHRONOUS
SINGLE-TUNED STAGES

Number of Cascaded Stages, n	Reduction in Bandwidth, $\Delta f_t/\Delta f$
1	1.0
2	0.644
3	0.510
4	0.435
5	0.386
6	0.350
7	0.323
8	0.301
9	0.283
10	0.268
15	0.217
20	0.187

2. Wideband Single-tuned Amplifiers. In wideband amplifier design, R_c is made much smaller than the value of r_p in parallel with the resonant plate circuit. For this case the voltage gain at any frequency is given by Eq. (7.73).

$$A(f) = \frac{-g_m R_c}{\sqrt{1 + Q_o{}^2(f/f_o - f_o/f)^2}} \tag{7.73}$$

where

$$Q_o = \omega_o C_t R_c \tag{7.74}$$

The gain at resonance is given by Eq. (7.75), and the bandwidth Δf is given by Eq. (7.76).

$$A = -g_m R_c \tag{7.75}$$

$$\Delta f \simeq \frac{f_o}{Q_o} = \frac{1}{2\pi R_c C_t} \tag{7.76}$$

If n identical synchronous single-tuned stages are cascaded the over-all response function is

$$A_t(f) = \left[\frac{-g_m R_c}{\sqrt{1 + Q_o{}^2(f/f_o - f_o/f)^2}}\right]^n \tag{7.77}$$

The over-all -3-db bandwidth Δf_t for n cascaded stages can be determined from Eq. (7.78).

$$\Delta f_t = \Delta f \sqrt{2^{1/n} - 1} \tag{7.78}$$

The ratio of the bandwidth of n stages in cascade to the bandwidth of a single stage is tabulated as a function of n in Table 7.4. From Table 7.4 it is seen that as n increases, the ratio of the over-all bandwidth to single-stage bandwidth decreases rapidly.
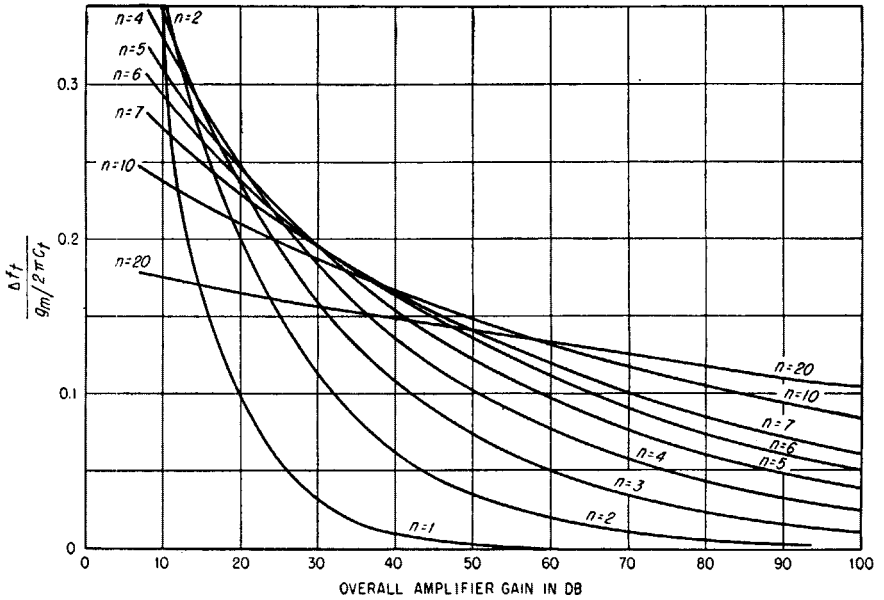
FIG. 7.17. Universal design curves for cascaded identical synchronous single-tuned amplifier stages. The curves indicate the maximum bandwidth that can be achieved with a given tube type, number of tubes, and over-all amplifier gain.
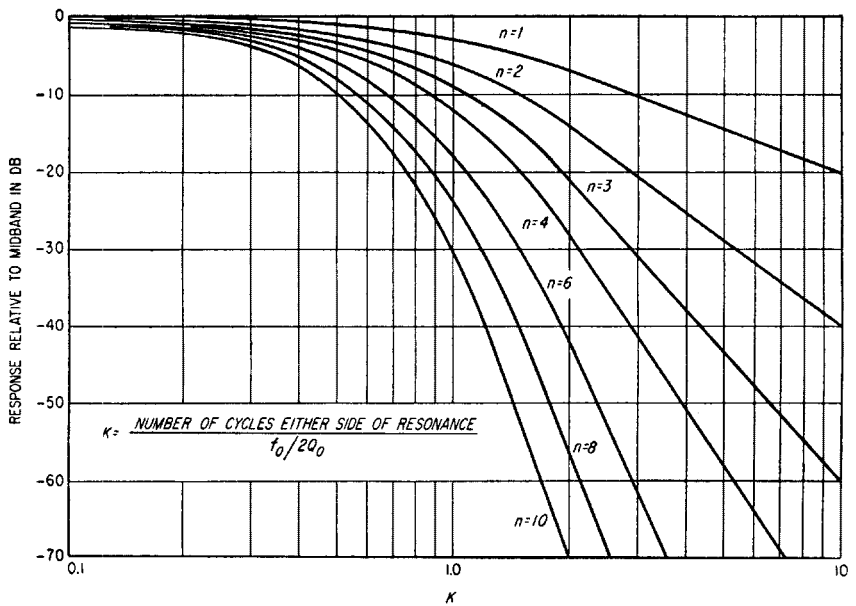
FIG. 7.18. Relative response of n cascaded synchronous single-tuned stages.

The relation between over-all gain and bandwidth for n cascaded stages is given in Eq. (7.79).

$$\Delta f_t = \frac{g_m \sqrt{2^{1/n} - 1}}{2\pi C_t A_t^{1/n}} \qquad (7.79)$$

where $\dfrac{g_m}{2\pi C_t}$ = gain-bandwidth product for particular tube used in amplifier

A_t = over-all gain of amplifier

A universal curve for synchronous single-tuned amplifier design is given in Fig. 7.17. From this graph the required number of stages for any desired over-all gain and bandwidth can be determined once the type of tube has been selected. Because of the manner in which the over-all bandwidth of an amplifier is reduced as additional identical stages are cascaded, there is a maximum bandwidth achievable for a given over-all gain and particular tube type. This maximum bandwidth occurs when the voltage gain in each stage is equal to $\sqrt{e}$, that is, a voltage gain of 1.65 (4.34 db). The design of an amplifier having a larger or smaller gain per stage, and consequently the use of a smaller or larger number of tubes to achieve the same gain, will result in a narrower over-all bandwidth.

As the number of cascaded stages is increased, the over-all response curve becomes narrower at the -3-db level and has steeper sides. This greatly increases the discrimination against adjacent-channel signals and increases the amplifier selectivity. A plot of the relative response of an amplifier having various numbers of cascaded single-tuned stages is given in Fig. 7.18.

Example 7.6

Design a synchronous single-tuned i-f amplifier having 85-db gain and an over-all bandwidth of 3.5 ± 0.2 Mc. Assume a center frequency of 15 Mc.

Solution

The center frequency of 15 Mc determines only the required coil inductance to resonate with the interstage shunt capacitance at the center frequency.

1. Determine the per-stage gain-bandwidth product.

To use a minimum number of tubes, the tube selected should have a high gain-bandwidth product. Assume that a 6AK5 tube is selected. A nominal value of g_m for the 6AK5 is 5,000 μmhos, and an average value of total interstage capacitance is 12 $\mu\mu$f. Under these conditions the gain-bandwidth product for the 6AK5 is, from Eq. (7.68),

$$
\begin{aligned}
|A|\ \Delta f &= \frac{g_m}{2\pi C_t} \\
&= \frac{5{,}000 \times 10^{-6}}{2\pi \times 12 \times 10^{-12}} \\
&= 66\ \text{Mc}
\end{aligned}
$$

2. Determine the ratio of over-all bandwidth to per-stage gain-bandwidth product.

The over-all bandwidth of the amplifier is to be 3.5 Mc. The ratio

$$\frac{\Delta f_t}{g_m/2\pi C_t} = \frac{3.5}{66} = 0.053$$

3. Determine the number of stages required.

From Fig. 7.17 it can be seen that five stages are required. The gain for five stages will be 86 db as shown in Fig. 7.17.

4. Determine the per-stage gain.

$$A = 86/5 = 17.2\ \text{db}$$

or, a voltage gain of 7.2.

5. Determine the required value of load resistance R_c

$$A = g_m R_c$$
$$R_c = \frac{7.2}{5,000 \times 10^{-6}}$$
$$= 1,440 \text{ ohms}$$

6. Determine the required value of inductance L to resonate with 12 $\mu\mu f$ at 15 Mc.

$$L = \frac{1}{(2\pi)^2 f_o{}^2 C_t} = \frac{1}{(6.28)^2 \times (15 \times 10^6)^2 \times 12 \times 10^{-12}}$$
$$= 9.39 \ \mu h$$

NOTE: As a check on the above design, the bandwidth per stage is

$$\Delta f = \frac{g_m/2\pi C_t}{A} = \frac{66}{7.2} = 9.17 \text{ Mc}$$

Table 7.4 shows that for five cascaded stages the over-all bandwidth is 0.386 times the per-stage bandwidth. Therefore,

$$\Delta f_t = 0.386 \times 9.17$$
$$= 3.54 \text{ Mc}$$

7.4e. Coupling Network with Untuned Primary and Tuned Secondary. In Fig. 7.19a the untuned primary inductance is made low and the coefficient of coupling between

(a) ACTUAL CIRCUIT   (b) EQUIVALENT CIRCUIT
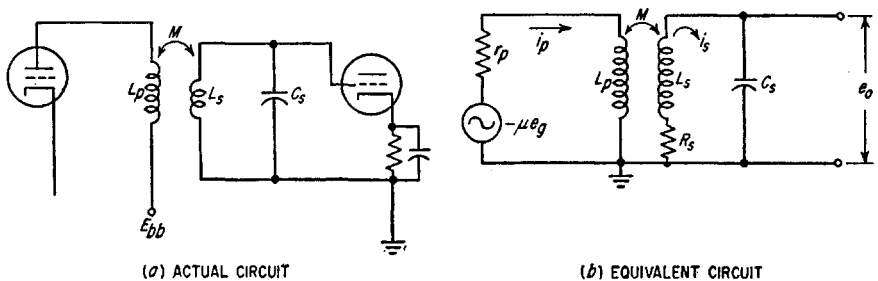
FIG. 7.19. Interstage coupling network which consists of a transformer having an untuned primary and a tuned secondary.

primary and secondary is made high. The result is that only the tube and stray capacitances on the secondary side of the transformer need be considered. The equivalent circuit is as shown in Fig. 7.19b. The voltage gain at resonance of one stage having this coupling network in its plate circuit is

$$A \simeq \frac{\mu\omega_o M Q_o}{r_p + (\omega_o{}^2 M^2/R_s)} \tag{7.80}$$

where ω_o = resonant frequency of secondary, radians/sec

M = mutual inductance between primary and secondary of transformer, henrys

R_s = a-c resistance of secondary inductance, ohms

Q_o = Q of secondary inductance at resonance

In general, r_p will be much greater than the coupled impedance of the resonant circuit and Eq. (7.80) reduces to

$$A \simeq g_m \omega_o M Q_o \tag{7.81}$$

Equation (7.80) has its maximum value when

$$M = \frac{\sqrt{R_s r_p}}{\omega_o} \qquad (7.82)$$

If the coupling is increased above the optimum value, both the bandwidth and the amplification will be reduced. If the coupling is reduced from the optimum value, the bandwidth will be increased and the amplification will be reduced. The optimum value of M usually cannot be achieved if a pentode is used because of its high value of plate resistance.

7.4f. Stagger-tuned Amplifiers. In cascading synchronous single-tuned stages, increasing the over-all amplifier gain by adding stages results in a lowering of the over-all bandwidth [see Eq. (7.78)]. If the figure of merit F_A is defined as the gain per stage times the over-all bandwidth, the value of F_A for a synchronous single-tuned amplifier is

$$F_A = \frac{g_m}{2\pi C_t} \sqrt{2^{1/n} - 1} \qquad (7.83)$$

It can be shown that by properly staggering the resonant frequencies of n single-tuned stages and by properly adjusting the Q's of these stages, the over-all amplifier figure of merit can be made to equal $g_m/2\pi C_t$, that is, the over-all bandwidth of n properly staggered stages having the same total gain as n synchronously tuned stages can be made equal to the bandwidth of a single stage of the synchronously tuned amplifier. The mathematical derivation of the tuning and damping relationships in the n stages is complex and is not treated here.[1] Staggering provides a passband which is maximally flat, i.e., the first $n - 1$ derivatives of the gain equation vanish at the center frequency.

Although any number of stages can be staggered as described above, only flat-staggered pairs and flat-staggered triples are widely used.

1. Flat-staggered Pairs. A flat-staggered pair consists of two cascaded stages having the following parameters:

f_o = geometric center frequency of over-all amplifier
f_1 = resonant frequency of stage 1
f_2 = resonant frequency of stage 2
Δf = over-all bandwidth of staggered pair
d = dissipation factor of individual stages, that is, $1/Q$
α = factor relating f_1 and f_2 to f_o
δ = dissipation factor of staggered pair

$$f_1 = \frac{f_o}{\alpha} \qquad (7.84)$$

$$f_2 = \alpha f_o \qquad (7.85)$$

$$\Delta f = \delta f_o \qquad (7.86)$$

$$\Delta f_1 = d f_1 \qquad (7.87)$$

$$\Delta f_2 = d f_2 \qquad (7.88)$$

The parameters of the tuned circuits are determined from Eqs. (7.89) and (7.90).

$$d = \sqrt{\frac{4 + \delta^2 - \sqrt{16 + \delta^4}}{2}} \qquad (7.89)$$

$$\alpha - \frac{1}{\alpha} = \sqrt{\delta^2 - d^2} \qquad (7.90)$$

---

[1] For greater detail see S. Butterworth, On the Theory of Filter Amplifiers, *Wireless Eng.*, 1930, pp. 536–541.

The damping factor d and the term α can also be determined from Fig. 7.20 where d and α have been plotted as a function of δ as determined from Eqs. (7.89) and (7.90).
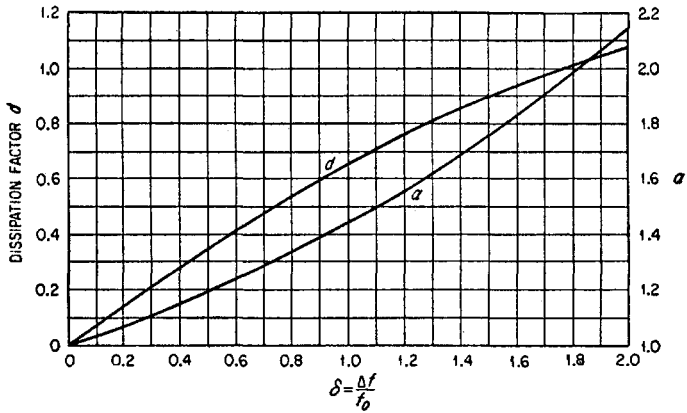
FIG. 7.20. Design curves for a flat-staggered pair.

The values of shunt damping resistances required to give the proper Q for each stage are given by

$$R_1 = \frac{1}{2\pi f_1 C_1 d} \qquad (7.91)$$

$$R_2 = \frac{1}{2\pi f_2 C_2 d} \qquad (7.92)$$

where C_1 and C_2 = total shunt capacitances across R_1 and R_2, respectively.

As an illustrative example of the relative responses of the individual stages which make up a staggered pair, refer to Fig. 7.21, which has been plotted for the case
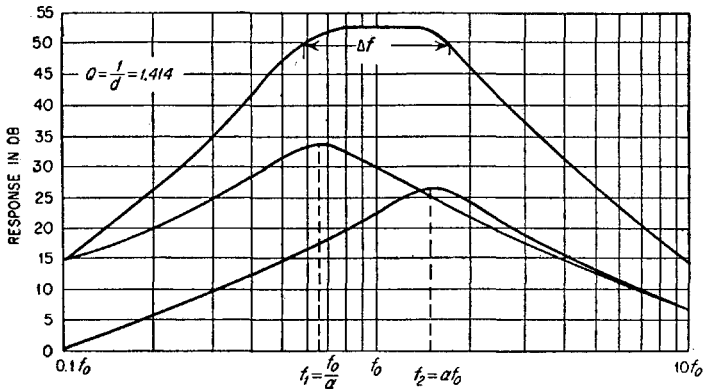
FIG. 7.21. Over-all response and individual stage response of a flat-staggered pair where $\delta = 1.1$, $A_1 = 33.5$ db, and $A_2 = 26.5$ db.

where $\delta = 1.1$. In Fig. 7.21, the gain of the individual stages at their center frequencies has been assumed to be 33.5 and 26.5 db, respectively.

The voltage gain as a function of frequency for a flat-staggered pair is

$$A(f) = \frac{g_{m1}g_{m2}}{\omega_o{}^2 C_1 C_2} \frac{1}{\sqrt{\delta^4 + (f/f_o - f_o/f)^4}} \qquad (7.93)$$

For a given bandwidth, the response curve of a flat-staggered pair has steeper sides

than the response curve of a single-tuned stage. Therefore the bandwidth reduction of n cascaded staggered pairs is less than the bandwidth reduction of n cascaded synchronous single-tuned stages. The over-all bandwidth of n cascaded pairs is given by

$$\Delta f_t = \Delta f \sqrt[4]{2^{1/n} - 1} \qquad (7.94)$$

Two identical cascaded single-tuned stages will have an over-all gain of

$$A_t = \left(\frac{g_m \sqrt{2^{1/2} - 1}}{2\pi \Delta f_t C_t} \right)^2$$

$$= 0.414 \left(\frac{g_m}{2\pi C_t \, \Delta f_t} \right)^2 \qquad (7.95)$$

The gain of a flat-staggered pair at band center is

$$A = \left(\frac{g_m}{2\pi C_t \, \Delta f_t} \right)^2 \qquad (7.96)$$

The ratio of the gain of a staggered pair to that of two identical single-tuned stages having the same over-all bandwidth is

$$\frac{A}{A_t} = \frac{1}{0.414} = 2.42 \qquad (7.97)$$

Thus, each stage in a staggered pair has an effective gain equal to 1.56 times that of the gain per stage of two single-tuned stages in cascade which have the same over-all
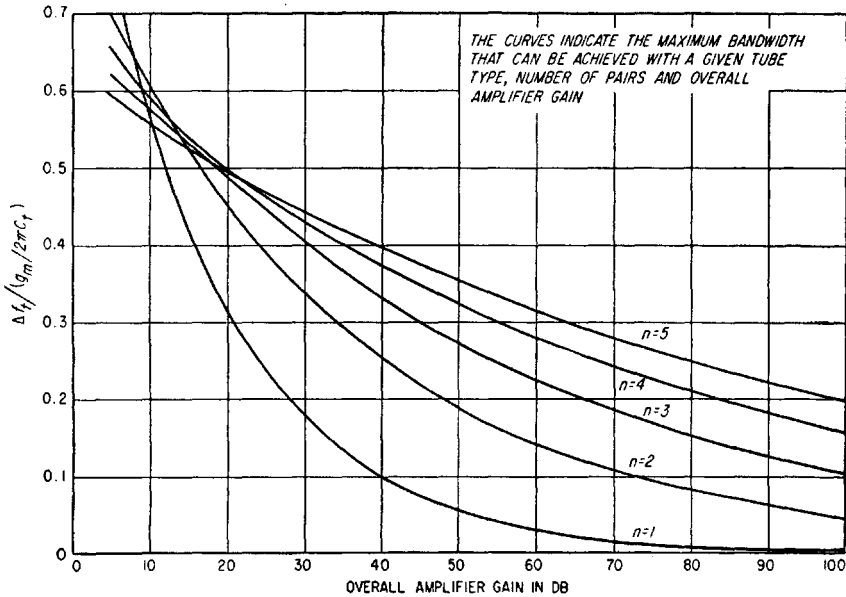
FIG. 7.22. Universal design curves for cascaded identical flat-staggered pairs.

bandwidth as the staggered pair. As the number of staggered pairs in cascade is increased, the effective gain per stage becomes increasingly greater as compared to synchronous single-tuned stages having the same over-all bandwidth and the same number of stages.

A universal design curve for cascaded identical flat-staggered pairs is given in Fig. 7.22. Once the tube type has been determined and the gain-bandwidth product

for the tube calculated, i.e., $g_m/2\pi C_t$, the number of staggered pairs required to achieve the desired over-all gain A_t and bandwidth Δf_t can be determined from Fig. 7.22. The bandwidth Δf of each pair is then determined from Eq. (7.94). The proper resonant frequency and Q of each stage in the pair are then determined from Fig. 7.20. The gain of each pair at midband is determined from Eq. (7.98).

$$A = \left(\frac{g_m}{2\pi C_t \, \Delta f}\right)^2 \tag{7.98}$$

where Δf = bandwidth of the staggered pair

Example 7.7

Design an i-f amplifier having 60 db gain and a bandwidth of 15 Mc using flat-staggered pairs. Assume that the center frequency is 30 Mc and that 6AG5 tubes are to be used.

1. Determine the gain-bandwidth product for the 6AG5 tube assuming the socket and wiring capacitance to be 5 $\mu\mu$f.

From Table 7.3 the gain-bandwidth product for the 6AG5 is

$$\frac{g_m}{2\pi C_t} = 60 \text{ Mc}$$

2. Determine the ratio of over-all bandwidth Δf_t to the per-stage gain-bandwidth product $g_m/2\pi C_t$.

Since the over-all bandwidth is to be 15 Mc,

$$\frac{\Delta f_t}{g_m/2\pi C_t} = \frac{15}{60} = 0.25$$

3. Determine the minimum number of pairs required from Fig. 7.22.

From Fig. 7.22 it is determined that for

$$\frac{\Delta f_t}{g_m/2\pi C_t} = 0.25$$

and for 60-db over-all gain, the minimum number of staggered pairs required is 4. The curve shows that 67.5 db can be achieved being four staggered pairs.

4. Determine the bandwidth of each pair.

From Eq. (7.94)

$$\Delta f = \frac{\Delta f_t}{\sqrt[4]{2^{1/4} - 1}}$$
$$= \frac{15 \text{ Mc}}{0.66}$$
$$= 22.7 \text{ Mc}$$

5. Determine the factor δ from Eq. (7.86).

$$\delta = \frac{\Delta f}{f_o} = \frac{22.7}{30} = 0.76$$

6. Determine α and d from Fig. 7.20.

$$\alpha = 1.32$$
$$d = 0.52$$

7. Determine the resonant frequencies of each stage in the pair from Eqs. (7.84) and (7.85).

$$f_1 = \frac{f_o}{\alpha} = \frac{30}{1.32} = 22.7 \text{ Mc}$$
$$f_2 = \alpha f_o = 1.32 \times 30 = 39.6 \text{ Mc}$$

8. Determine the value of the load resistor in each stage of one pair.

The total shunt capacitance C_t between stages is, using average values of c_i and c_o from Table 7.3,

$$C_t = 6.5 + 1.8 + 5 = 13.3 \ \mu\mu f$$

The correct value of damping resistor for each stage is determined from Eqs. (7.91) and (7.92).

$$R_1 = \frac{1}{2\pi f_1 C_1 d}$$

$$= \frac{1}{6.28 \times 22.7 \times 10^6 \times 13.3 \times 10^{-12} \times 0.52}$$

$$= 1{,}014 \text{ ohms}$$

$$R_2 = \frac{1}{2\pi f_2 C_2 d}$$

$$= \frac{1}{6.28 \times 39.6 \times 10^6 \times 13.3 \times 10^{-12} \times 0.52}$$

$$= 582 \text{ ohms}$$

9. Determine the correct values of inductance for each stage in one pair.

$$L_1 = \frac{1}{(2\pi f_1)^2 C_1}$$

$$= \frac{1}{(6.28 \times 22.7 \times 10^6)^2 \times 13.3 \times 10^{-12}}$$

$$= 3.70 \ \mu h$$

$$L_2 = \frac{1}{(2\pi f_2)^2 C_2}$$

$$= \frac{1}{(6.28 \times 39.6 \times 10^6)^2 \times 13.3 \times 10^{-12}}$$

$$= 1.22 \ \mu h$$

In Step 3, the over-all gain was found to be 67.5 db. If exactly 60 db is desired, the g_m of one or more tubes can be lowered by increasing the grid bias from the nominal value.

10. As a check on the design, determine the voltage gain of each pair as given by Eq. (7.98).

$$A = \left(\frac{g_m}{2\pi C_t \,\Delta f}\right)^2$$

$$= \frac{60^2}{22.7^2}$$

$$= 6.99, \text{ or } 16.88 \text{ db}$$

The over-all gain for four pairs is

$$A_t = 4 \times 16.88$$

$$= 67.5 \text{ db}$$

Thus, four staggered pairs of 6AG5's will provide slightly more than the required 60-db gain and 15-Mc bandwidth.

11. The circuit of one pair is shown in Fig. 7.23.
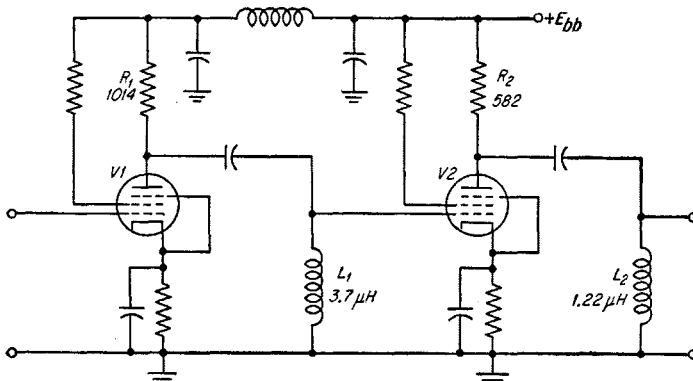
Fig. 7.23. Circuit of flat-staggered pair of Example 7.7.

2. Flat-staggered Triples. A flat-staggered triple consists of three cascaded stages. One stage is tuned to the center frequency f_o of the amplifier and has a bandwidth Δf equal to the over-all bandwidth of the triple. The other two stages are tuned to f_1 and f_2, where

$$f_1 = \frac{f_o}{\alpha} \tag{7.84}$$

and

$$f_2 = \alpha f_o \tag{7.85}$$

$$\Delta f = \delta f_o \tag{7.86}$$

where Δf = bandwidth of over-all staggered triple as well as one stage of triple

$$\Delta f_1 = df_1 \tag{7.87}$$

$$\Delta f_2 = df_2 \tag{7.88}$$

The parameters of the tuned circuits are determined by Eqs. (7.90) and (7.99).

$$d = \sqrt{\frac{4 + \delta^2 - \sqrt{16 + 4\delta^2 + \delta^4}}{2}} \tag{7.99}$$

The dissipation factor d and the term α can be obtained from Fig. 7.24 where d and α
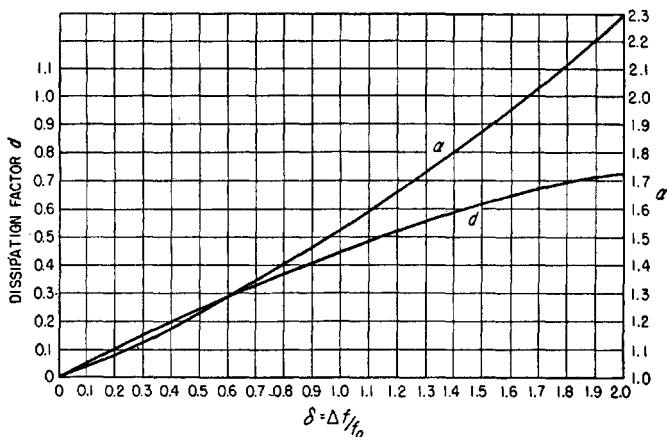
Fig. 7.24. Design curves for a flat-staggered triple.

have been plotted as a function of δ, the dissipation factor of the over-all triple. The correct values of stage load resistors are given by

$$R_o = \frac{1}{2\pi \delta f_o C_o} \tag{7.100}$$

$$R_1 = \frac{1}{2\pi f_1 C_1 d} \tag{7.101}$$

$$R_2 = \frac{1}{2\pi f_2 C_2 d} \tag{7.102}$$

As an illustrative example of the relative responses of the individual stages in a staggered triple, refer to Fig. 7.25 which has been plotted for the case where $\delta = 0.95$. In Fig. 7.25, the gain of the individual stages at their center frequencies has been assumed to be 22 db, 11.6 db, and 15 db, respectively.

RECEIVERS 7-39

The voltage gain as a function of frequency for a flat-staggered triple is given by

$$A = \frac{-g_{m1}g_{m2}g_{mo}}{\omega_o{}^3 C_1 C_2 C_o} \frac{1}{\sqrt{\delta^6 + \left(\dfrac{f}{f_o} - \dfrac{f_o}{f}\right)^6}} \qquad (7.103)$$

The amplitude response of a flat-staggered triple has steeper sides than that of a flat-staggered pair. Therefore the bandwidth reduction of cascaded staggered
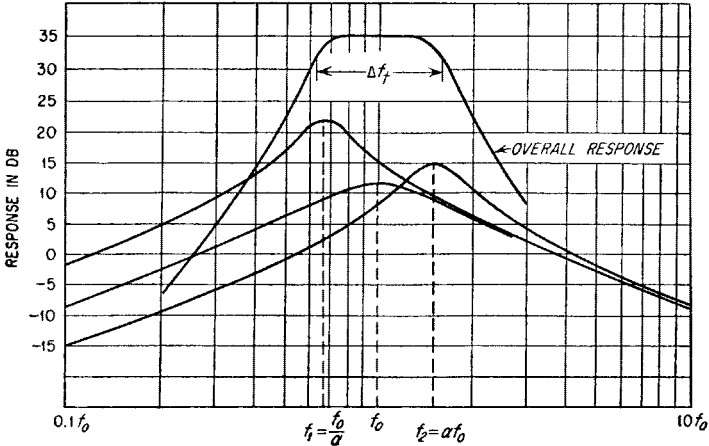
FIG. 7.25. Typical over-all response and individual stage responses of a flat-staggered triple in which $\delta = 0.95$, $d = 0.43$, and $\alpha = 1.5$.

triples is less than that of cascaded staggered pairs. The over-all bandwidth Δf_t of n cascaded triples is given by

$$\Delta f_t = \Delta f \sqrt[6]{2^{1/n} - 1} \qquad (7.104)$$

where Δf = bandwidth of one flat-staggered triple

Three identical cascaded single-tuned stages have an over-all voltage gain of

$$A_t = \left(\frac{g_m \sqrt{2^{1/3} - 1}}{2\pi C_t \Delta f_t}\right)^3$$

$$= \left(\frac{0.51 g_m}{2\pi C_t \Delta f_t}\right)^3$$

where Δf_t = over-all bandwidth of three synchronous single-tuned stages

The voltage gain at bandcenter for one flat-staggered triple is

$$A = \left(\frac{g_m}{2\pi C_t \Delta f}\right)^3 \qquad (7.105)$$

where Δf = over-all bandwidth of one flat-staggered triple

Thus, one staggered triple having the same over-all bandwidth as the bandwidth of three identical cascaded single-tuned stages will have a greater gain than the latter by the factor

$$\frac{A}{A_t} = \frac{1}{(0.51)^3} = 7.54 \qquad (7.106)$$

Therefore, the effective per-stage gain in a staggered triple is 1.96 times the gain per stage of three cascaded single-tuned stages having the same over-all bandwidth. A universal design curve for cascaded identical flat-staggered triples is given in Fig. 7.26. Once the type of tube has been determined and the desired over-all gain and band-width are known, the required number of staggered triples can be determined. The bandwidth Δf of each triple can then be determined from Eq. (7.104). After the bandwidth of each triple has been determined, the center frequency and dissipation
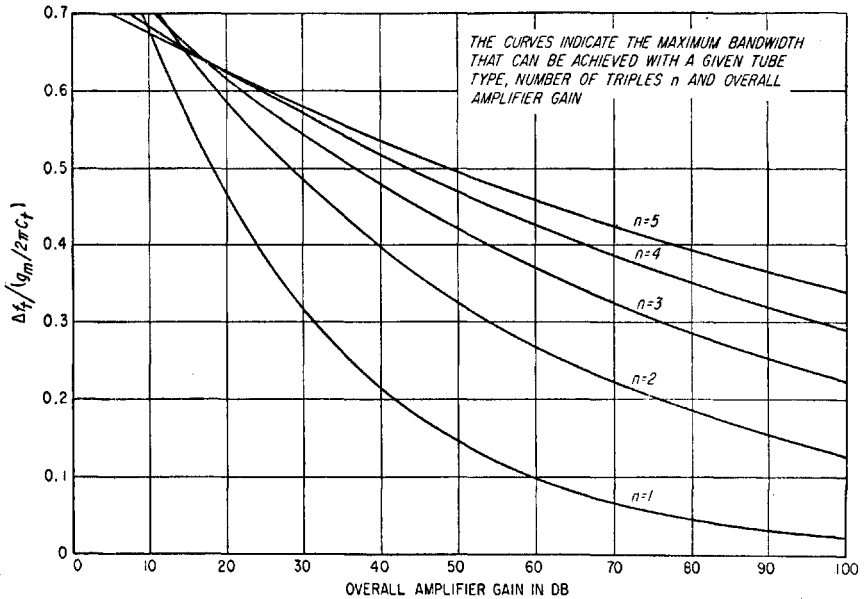
FIG. 7.26. Universal design curves for cascaded identical flat-staggered triples.

factor for each stage in the triple can be determined from Fig. 7.24 and the exact gain of each triple can be determined from Eq. (7.105).

Example 7.8

Design an i-f amplifier having 80-db gain and a bandwidth of 20 Mc at a center frequency of 60 Mc. Assume that 6AK5 tubes are to be used.

1. Determine the gain-bandwidth product of the 6AK5, assuming a nominal value of 5 $\mu\mu$f stray and socket capacitance.

From Table 7.3 the gain-bandwidth product of the 6AK5 under this condition is seen to be

$$\frac{g_m}{2\pi C_t} = 68 \text{ Mc}$$

2. Determine the ratio of over-all bandwidth to per-stage gain-bandwidth product.

$$\frac{\Delta f_t}{g_m/2\pi C_t} = \frac{20}{68} = 0.294$$

3. Determine from Fig. 7.26 the required number of triples to achieve 80-db gain.

$$n = 4$$

The curve shows that 99 db can be achieved with the desired bandwidth using four staggered triples.

4. Determine the over-all bandwidth Δf of each triple from Eq. (7.104).

$$\Delta f = \frac{\Delta f_t}{\sqrt[6]{2^{1/4} - 1}}$$

$$= \frac{20}{\sqrt[6]{1.189 - 1}} = 26.4 \text{ Mc}$$

5. Determine δ.

$$\delta = \frac{\Delta f}{f_o} = \frac{26.4}{60} = 0.44$$

6. Determine α and d from Fig. 7.24.

$$d = 0.218$$
$$\alpha = 1.195$$

7. Determine the resonant frequencies of each stage in one triple.

Stage 1 is tuned to $f_1 = f_o/\alpha = 50.2$ Mc
Stage 2 is tuned to $f_2 = \alpha f_o = 71.7$ Mc
Stage 3 is tuned to $f_o = 60$ Mc

8. Determine the correct value of load resistor for each stage in one triple from Eqs. (7.100), (7.101), and (7.102).
The approximate value of interstage capacitance can be obtained from Table 7.3.

$$C_t \simeq 11.8 \ \mu\mu\text{f}$$

Stage 1:

$$R_1 = \frac{1}{2\pi f_1 C_1 d}$$

$$= \frac{1}{6.28 \times 50.2 \times 10^6 \times 11.8 \times 10^{-12} \times 0.218}$$
$$= 1{,}230 \text{ ohms}$$

Stage 2:

$$R_2 = \frac{1}{2\pi f_2 C_2 d}$$

$$= \frac{1}{6.28 \times 71.7 \times 10^6 \times 11.8 \times 10^{-12} \times 0.218}$$
$$= 864 \text{ ohms}$$

Stage 3:

$$R_o = \frac{1}{2\pi f_o C_o \delta}$$

$$= \frac{1}{6.28 \times 60 \times 10^6 \times 11.8 \times 10^{-12} \times 0.44}$$
$$= 511 \text{ ohms}$$

9. Determine the correct values of inductance for each stage in one triple.

Stage 1:

$$L_1 = \frac{1}{(2\pi f_1)^2 C_1}$$

$$= \frac{1}{(6.28 \times 50.2 \times 10^6)^2 \times 11.8 \times 10^{-12}}$$
$$= 0.85 \ \mu\text{h}$$

Stage 2:

$$L_2 = \frac{1}{(2\pi f_2)^2 C_2}$$

$$= \frac{1}{(6.28 \times 71.7 \times 10^6)^2 \times 11.8 \times 10^{-12}}$$
$$= 0.42 \ \mu\text{h}$$

Stage 3:

$$L_o = \frac{1}{(2\pi f_o)^2 C_o}$$

$$= \frac{1}{(6.28 \times 60 \times 10^6)^2 \times 11.8 \times 10^{-12}}$$

$$= 0.60 \ \mu h$$

10. As stated in step 3, the over-all gain will be 99 db; however, exactly 80 db can be achieved by decreasing the g_m of one or more stages by increasing the operating bias.

11. As a check on the design, determine the voltage gain at bandcenter for one triple from Eq. (7.105).

$$A = \left(\frac{g_m}{2\pi C_t} \times \frac{1}{\Delta f} \right)^3$$

$$= \left(\frac{68}{26.4} \right)^3 = 17.1$$

$$= 24.7 \ db$$

12. Determine the exact over-all amplifier gain.

$$A_t = 24.7 \times 4$$

$$= 98.8 \ db$$

Thus, the design checks.

7.4g. Double-tuned Amplifiers. The circuit of a typical double-tuned amplifier stage is shown in Fig. 7.27. The primary and secondary circuits are both resonant and are inductively coupled together. The desired selectivity curve is achieved through proper adjustment of the primary and secondary resonant circuits and the degree of coupling between the primary and secondary.
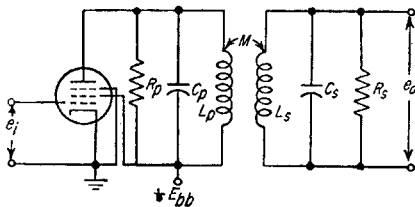
FIG. 7.27. An inductively coupled double-tuned circuit.

For purposes of design, double-tuned circuits are most conveniently divided into two parts: the high-Q case where the bandwidth is $0.1f_o$ or less and the low-Q case where the bandwidth is greater than $0.1f_o$.

1. High-Q Case. The high-Q case has been investigated in great detail by several persons,[1] and only applicable design information is given here. The general analysis of double-tuned circuits is covered in Sec. 13. In the high-Q case both the primary and secondary resonant circuits are tuned to the center frequency of the passband, i.e.,

$$f_o = \frac{1}{2\pi \sqrt{L_p C_p}} = \frac{1}{2\pi \sqrt{L_s C_s}} \tag{7.107}$$

The coefficient of coupling k between the primary and secondary is given by

$$k = \frac{M}{\sqrt{L_p L_s}} \tag{7.108}$$

where M = mutual inductance between windings
The midband gain A is given by

$$A = \frac{g_m k \sqrt{R_p R_s Q_p Q_s}}{1 + k^2 Q_p Q_s} \tag{7.109}$$

[1] For a complete account see C. B. Aiken, Two Mesh-tuned Coupled Circuit Filters, *Proc. IRE*, vol. 35, February, 1937.

	Critical coupling $Q_p/Q_s = a$ or $Q_s/Q_p = a$	Transitional coupling $Q_p/Q_s = a$ or $Q_s/Q_p = a$
Coefficient of coupling....................	$k_c = \dfrac{1}{\sqrt{Q_p Q_s}}$	$k_t = \dfrac{1}{aQ_s}\sqrt{\dfrac{1+a^2}{2}}$
−3-db bandwidth........................	$\Delta f = \dfrac{f_o}{Q_s\sqrt{2}}\sqrt{\left[\left(\dfrac{a-1}{a}\right)^4 + \dfrac{16}{a^2}\right]^{1/2} - \left(\dfrac{a-1}{a}\right)^2}$	$\Delta f = \dfrac{f_o(1+a)}{aQ_s\sqrt{2}}$
Midband gain...........................	$A = \dfrac{g_m\sqrt{R_p R_s}}{2}$	$A = g_m\sqrt{R_p R_s}\,\dfrac{\sqrt{2a(1+a^2)}}{(1+a)^2}$
Gain-bandwidth product..................	$A\,\Delta f = \dfrac{g_m\sqrt{a/2}}{4\pi\sqrt{C_p C_s}}\sqrt{\left[\left(\dfrac{a-1}{a}\right)^4 + \dfrac{16}{a^2}\right]^{1/2} - \left(\dfrac{a-1}{a}\right)^2}$	$A\,\Delta f = \dfrac{g_m\sqrt{1+a^2}}{2\pi\sqrt{C_p C_s}\,(1+a)}$
Over-all bandwidth of n stages.............	$\Delta f_t = \Delta f\sqrt[4]{2^{1/n}-1}$	$\Delta f_t = \Delta f\sqrt[4]{2^{1/n}-1}$
Midband input impedance at transformer input including C_p as part of transformer circuit	$R_{11} = R_p$	$R_{11} = \dfrac{2R_p}{a+1/a}$

where $Q_p = Q$ of primary circuit

$Q_s = Q$ of secondary circuit

For fixed values of Q_p and Q_s, there is a particular value of k that makes the midband gain A a maximum. This value of coupling is known as the *critical coefficient of coupling* k_c, and a circuit having this value of k is said to be critically coupled. The value of the critical coefficient of coupling is

$$k_c = \frac{1}{\sqrt{Q_p Q_s}} \qquad (7.110)$$

There is another value of the coupling coefficient k which provides the flattest passband. When k has this value, the circuit is said to be *transitionally coupled*. If the
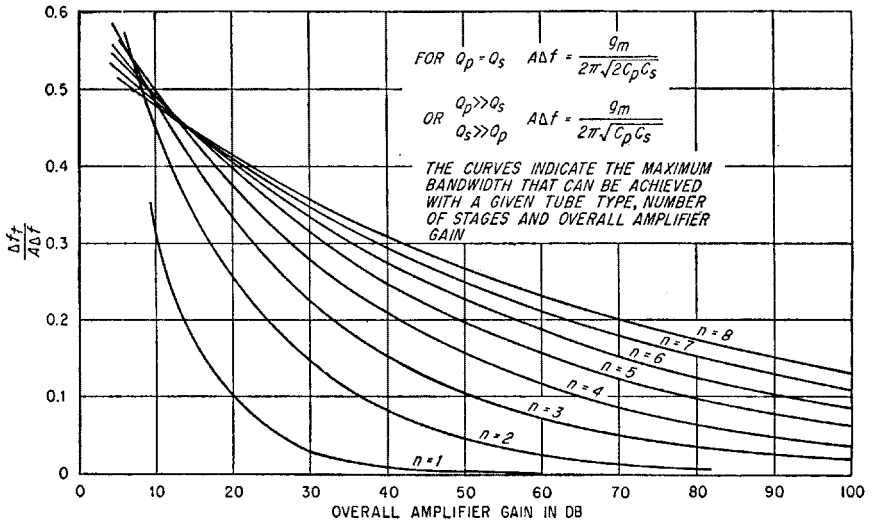
FIG. 7.28. Universal design curves for cascaded transitionally coupled double-tuned stages.

coupling is increased beyond the transitional value, the passband has two peaks. If the coupling is less than transitional, the passband has a single peak. This is illustrated in Fig. 13.4b. The transitional value of coupling coefficient is

$$k_t = \sqrt{\frac{1}{2}\left(\frac{1}{Q_p{}^2} + \frac{1}{Q_s{}^2}\right)} \qquad (7.111)$$

When the primary and secondary circuits have equal Q's, transitional and critical coupling occur at the same coefficient of coupling.

A summary of the general conditions for critical and transitional coupling for the high-Q case is given in Table 7.5.

The two conditions most often prevailing in the application of double-tuned circuits are transitional coupling with $Q_p = Q_s$ and transitional coupling with $Q_p \gg Q_s$ or $Q_s \gg Q_p$. A summary of the properties of double-tuned circuits for these conditions is given in Table 7.6. Universal design curves for cascaded transitionally coupled double-tuned stages having either equal primary and secondary Q's or $Q_p \gg Q_s$ or $Q_s \gg Q_p$ are given in Fig. 7.28. From these curves the number of stages required to achieve a particular gain and bandwidth can be determined once the gain-bandwidth product for the tube type to be used has been calculated.

TABLE 7.6. SUMMARY OF CONDITIONS FOR TRANSITIONALLY
COUPLED DOUBLE-TUNED STAGE, HIGH-Q CASE

	Equal Q $Q_p = Q_s = Q$	$Q_p \gg Q_s$
Coefficient of coupling.............	$k_t = \dfrac{1}{Q}$	$k_t = \dfrac{1}{Q_s \sqrt{2}}$
−3-db bandwidth..................	$\Delta f = \dfrac{\sqrt{2} f_o}{Q}$	$\Delta f = \dfrac{f_o}{\sqrt{2}\, Q_s}$
Midband gain......................	$A = \dfrac{g_m \sqrt{R_p R_s}}{2}$	$A = g_m R_s \sqrt{\dfrac{2C_s}{C_p}}$
Gain-bandwidth product............	$A\,\Delta f = \dfrac{g_m}{2\pi \sqrt{2C_p C_s}}$	$A\,\Delta f = \dfrac{g_m}{2 \sqrt{C_p C_s}}$
Over-all bandwidth of n stages.......	$\Delta f_t = \Delta f \sqrt[4]{2^{1/n} - 1}$	$\Delta f_t = \Delta f \sqrt[4]{2^{1/n} - 1}$

Example 7.9

Design an i-f amplifier having 100-db gain and 2-Mc bandwidth centered at 30 Mc. Assume double-tuned coupling with $Q_p = Q_s$ and 6AK5 tubes.

1. Determine the gain-bandwidth product for a 6AK5 tube in a double-tuned circuit.

$$A\,\Delta f = \frac{g_m}{2\pi \sqrt{2C_p C_s}} \qquad \text{for the double-tuned equal-}Q\text{ case (see Table 7.6)}$$

From Table 7.3 the average input and output capacitances of the 6AK5 are found to be 3.9 $\mu\mu$f and 2.9 $\mu\mu$f, respectively. If an additional 2.5 $\mu\mu$f input and output capacitance is assumed for the socket and wiring, the total values of C_p and C_s are

$$C_p = 2.9 + 2.5 = 5.4 \ \mu\mu\text{f}$$
$$C_s = 3.9 + 2.5 = 6.4 \ \mu\mu\text{f}$$

From Table 7.3 the average g_m is 5,000 μmhos. The gain-bandwidth product is

$$A\,\Delta f = \frac{5,000 \times 10^{-6}}{2\pi \sqrt{2 \times 5.4 \times 10^{-12} \times 6.4 \times 10^{-12}}}$$
$$= 95.6 \ \text{Mc}$$

2. Determine the ratio of $\Delta f_t / A\,\Delta f$.

$$\frac{\Delta f_t}{A\,\Delta f} = \frac{2}{95.6} = 0.0209$$

3. Determine the number of stages required for 100-db gain from Fig. 7.28. The gain with three stages is marginal, therefore

$$n = 4$$

4. Determine the necessary bandwidth of each stage from Table 7.6.

$$\Delta f_t = \Delta f \sqrt[4]{2^{1/n} - 1}$$
$$\Delta f = \frac{2}{\sqrt[4]{2^{1/4} - 1}} = \frac{2}{0.66}$$
$$= 3.03 \ \text{Mc}$$

5. Determine the desired primary and secondary circuit Q's and the coefficient of coupling k_t from Table 7.6.

$$Q_p = Q_s = Q = \frac{f_o \sqrt{2}}{\Delta f}$$

$$= \frac{30 \times 1.414}{3.03}$$

$$= 14.0$$

$$k_t = \frac{1}{Q} = 0.0714$$

6. Determine the necessary values of R_p, R_s, L_p, and L_s.

$$R_p = \frac{Q}{2\pi f_o C_p}$$

$$= \frac{14}{6.28 \times 30 \times 10^6 \times 5.4 \times 10^{-12}}$$

$$= 13,800 \text{ ohms}$$

$$R_s = \frac{Q}{2\pi f_o C_s}$$

$$= \frac{14}{6.28 \times 30 \times 10^6 \times 6.4 \times 10^{-12}}$$

$$= 11,600 \text{ ohms}$$

$$L_p = \frac{1}{(2\pi f_o)^2 C_p}$$

$$= \frac{1}{(6.28 \times 30 \times 10^6)^2 \times 5.4 \times 10^{-12}}$$

$$= 5.2 \ \mu\text{h}$$

$$L_s = \frac{1}{(2\pi f_o)^2 C_s}$$

$$= \frac{1}{(6.28 \times 30 \times 10^6)^2 \times 6.4 \times 10^{-12}}$$

$$= 4.4 \ \mu\text{h}$$

7. As a check, determine the gain at f_o from Table 7.6.

$$A = g_m \frac{\sqrt{R_p R_s}}{2}$$

$$= 5,000 \times 10^{-6} \times \frac{\sqrt{13,800 \times 11,600}}{2}$$

$$= 31.6, \text{ or } 30.0 \text{ db}$$

$$A_t = 4A = 120 \text{ db}$$

Exactly 100 db can be obtained by operating the tubes at increased bias.

2. Transitionally Coupled Low-Q Case. The analysis of high-Q double-tuned circuits involves the use of several simplifying approximations which cannot be made when the bandwidth is of the same order of magnitude as the center frequency. An exact analysis is extremely complicated and has been worked out fully only for the case of transitional coupling. Design information is given for transitional coupling with equal primary and secondary Q's and with $Q_p \gg Q_s$ or $Q_s \gg Q_p$.

EQUAL PRIMARY AND SECONDARY Q'S. The curves shown in Fig. 7.29 apply to the low-Q case and show the relationships between coupling coefficient, primary and secondary resonant frequencies, primary and secondary circuit Q's, and the ratio of the over-all bandwidth to the center frequency. For the low-Q case, the primary and secondary resonant frequencies are lower than the center frequency. The

primary resonant frequency f'_p with the secondary shorted and the secondary frequency f'_s with the primary shorted can also be determined from the figure. From the value of $Q/2\pi$ established from Fig. 7.29, the RC products for the primary and secondary can be determined.

$$R_p C_p = \frac{1}{f_o} \times \frac{Q}{2\pi} \qquad (7.112)$$

$$R_s C_s = \frac{1}{f_o} \times \frac{Q}{2\pi} \qquad (7.113)$$

where R_p and R_s are the resistances shunting the primary and secondary inductances. The midband voltage gain of the amplifier is

$$A = \frac{g_m}{2} \sqrt{R_p R_s} \qquad (7.114)$$

The gain-bandwidth product of one stage is

$$A \, \Delta f = \frac{g_m}{2\pi \sqrt{2 C_p C_s}} \qquad (7.115)$$

Fig. 7.29. Design curves for low-Q transitionally coupled double-tuned circuit having equal primary and secondary Q's. The value of Q applies to the center frequency f_o.

INFINITE PRIMARY Q AND LOW SECONDARY Q. Figure 7.30 gives the necessary design information for a transitionally coupled circuit having $Q_p \gg Q_s$. The proper coefficient of coupling can be determined for a given ratio of bandwidth to center frequency by curve a. The midband gain of the stage is given by Eq. (7.116) where α is determined from curve b of the figure.

$$A = \alpha g_m R_s \sqrt{C_s/C_p} \qquad (7.116)$$

The resonant frequency of the primary circuit is taken from curve d, and the resonant frequency of the secondary circuit is taken from curve c. The value of R_s is obtained by finding the value of $Q/2\pi$ from curve e and substituting this value into Eq. (7.113).

The gain-bandwidth product of one stage is

$$A \, \Delta f = \frac{g_m}{2\pi \sqrt{C_p C_s}} \qquad (7.117)$$

Fig. 7.30. Design curves for transitionally coupled double-tuned circuit having infinite primary Q and low-Q secondary.

The secondary Q may be made much larger than the primary Q. For this condition, the design equations and Fig. 7.30 still apply; however, the frequencies to which the primary and secondary circuits are resonated should be interchanged.

The bandwidth reduction factor for cascaded low-Q transitionally coupled double-tuned circuits is the same as for the high-Q circuits. This reduction factor is given in Table 7.6. The universal design curves for cascaded double-tuned circuits given in Fig. 7.28 apply directly to both the low-Q and high-Q cases.

7.4h. UHF Effects. The use of conventional tubes as amplifiers becomes increasingly difficult as the frequency is increased because of several effects which become pronounced in the uhf region. In general, these effects can be reduced by careful

circuit design.   However, the tube structure limits the improvement that may be achieved.

1. Input Loading.   The input conductance of most vacuum tubes is very low at low radio frequencies.[1]   As the frequency of operation of the amplifier is increased,

TABLE 7.7. Δc AND REQUIRED VALUE OF UNBYPASSED CATHODE RESISTOR
R_k FOR COMPENSATION OF SEVERAL AMPLIFIER TUBES

Tube type	Cold-tube input capacity c_i, $\mu\mu f$	Capacity increase (cold-to-hot) Δc, $\mu\mu f$	g_m, μmhos	I_p ma	I_k, ma	R_k, ohms
6AK5	2.8	1.2	5,100	7.7	10.2	63.5
6AH6	6.8	3.2	11,000	10.0	12.5	34.2
6AG5	4.2	2.3	5,100	7.2	9.3	83.2
6CB6	4.5	1.8	6,200	10.0	13.0	49.7
6J6	0.8	1.4	5,300	8.5	8.5	330
6C4	1.1	0.7	3,100	11.8	11.8	205
6BA6	3.0	2.5	4,400	11.0	15.2	137
6AU6	2.9	2.6	5,200	10.8	15.1	124

the input conductance increases above the low-frequency value because of electron transit time within the tube and cathode lead inductance.

INPUT LOADING CAUSED BY ELECTRON TRANSIT TIME.   At moderate radio frequencies, the time of transit of the individual electrons between the cathode and the grid and between the grid and the plate is an appreciable fraction of one cycle of the applied grid signal.   The electrons induce charges on the grid as they approach and recede from the grid structure in traveling from the cathode to the plate, and the induced charges on the grid cause current to flow in the external grid-circuit impedance.   At low frequencies this grid current is very nearly in quadrature with the applied signal voltage and appears at the grid input terminals as a capacitive reactance.   This apparent capacitance is in addition to the grid-cathode capacitance of the tube structure and represents the major increase in grid-cathode capacitance when voltages are applied to the tube.   There is an additional increase in grid-cathode capacitance as the transconductance of a tube is increased because of the movement of the space charge between the cathode and grid (these changes in input capacitance should not be confused with input-capacitance changes caused by Miller effect).   Values of the "cold-to-hot" incremental changes in input capacitance Δc, which include both space-charge and transit-time effects, are given for several tubes in Table 7.7.
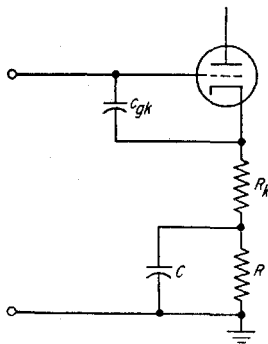
FIG. 7.31. Circuit for compensation of Δc by use of an unbypassed cathode resistor.

Because Δc is a function of the tube transconductance, a resonant circuit placed between grid and cathode will be detuned if the transconductance of the tube is varied.   If automatic volume control is applied to such a stage, the detuning effect

[1] See Sec. 3.8 for a discussion of the low-frequency input conductance of vacuum-tube amplifiers.

between strong signals and work signals may be appreciable. The detuning effects of Δc may be compensated by having a portion of the cathode bias resistor unbypassed as shown in Fig. 7.31. The decrease in the tube input capacitance by feedback as a result of the unbypassed cathode resistor will be equal to the increase in input capacitance because of Δc if R_k has the following value:

$$R_k = \frac{\Delta c I_b}{g_m c_{gk} I_k} \qquad (7.118)$$

where R_k = unbypassed portion of cathode resistor, ohms

Δc = "cold-to-hot" incremental capacitance of tube in $\mu\mu f$, i.e., the difference in c_{gk} when tube is cut off and when tube has transconductance of g_m

I_b = plate current at operating point

I_k = cathode current at operating point

g_m = grid-plate transconductance of tube in mhos at operating point

c_{gk} = grid-cathode capacitance of "cold" tube, $\mu\mu f$

At higher radio frequencies where the transit time of the electrons between cathode and grid is a significant fraction of one cycle of the applied signal, the induced grid current has a component which is in phase with the applied signal. This in-phase current represents a conductive component of grid-input admittance. A detailed analysis of this effect has been conducted by Llewellyn[1] and others. An approximate expression for input due to electron transit time is given by Eq. (7.119).

$$G_t = k g_m f^2 T^2 \qquad (7.119)$$

where G_t = tube input conductance, mhos

k = constant which is a function of tube structure

g_m = grid-plate transconductance, mhos

T = electron transit time from cathode to grid plane

f = frequency of operation, cps

Of special significance in Eq. (7.119) is the fact that the grid-input conductance increases as the square of frequency.

INPUT LOADING CAUSED BY CATHODE LEAD INDUCTANCE. At high frequencies, the grid, plate, and cathode leads internal to a tube represent inductances having appreciable reactance.

Of the three principal lead inductances, the inductance of the cathode lead is the most important because of the feedback it produces. The equivalent input circuit of a tube including cathode lead inductance is shown in Fig. 7.32. The voltage developed across the cathode lead inductance by the plate signal current is in series

FIG. 7.32. Input circuit of tube including cathode lead inductance.

with the input signal voltage. The effect of this feedback between plate and grid circuits is to cause a component of current to flow through the grid-cathode capacitance c_{gk} which is in phase with the applied signal e_s. The magnitude of this conductance is given approximately by

$$G_y = g_m \omega^2 L_k c_{gk} \qquad (7.120)$$

[1] F. B. Llewellyn, "Electron Inertia Effects," Cambridge University Press, London, 1941.

where G_y = grid conductance in mhos due to cathode lead inductance

g_m = tube grid-plate transconductance, mhos

ω = signal frequency, radians/sec

L_k = cathode lead inductance, henrys

c_{gk} = grid-cathode capacitance, farads

Equation (7.120) assumes that the reactance of L_k is small compared to the reactance of c_{gk} and that the voltage across L_k is small compared to the input signal e_s. Typical values of cathode lead inductance for miniature tubes fall between 0.01 and 0.02 μh. The effects of cathode lead inductance can be minimized in tubes having two cathode leads by the method illustrated in Fig. 7.33. One of the cathode leads contains the normal cathode bias resistor and bypass capacitor. The other lead has a small capacitor connected between the tube pin and ground of such a value as to be series-resonant with the inductance of that lead at the frequency of operation. The 6AK5 pentode, which has the suppressor and cathode internally connected and also has two base pins connected to these electrodes, is particularly suited to this type of compensation. The input conductance due to both electron transit time and cathode lead inductance is equal to
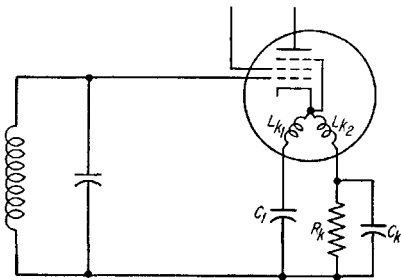
FIG. 7.33. Method of minimizing effects of cathode lead inductance.
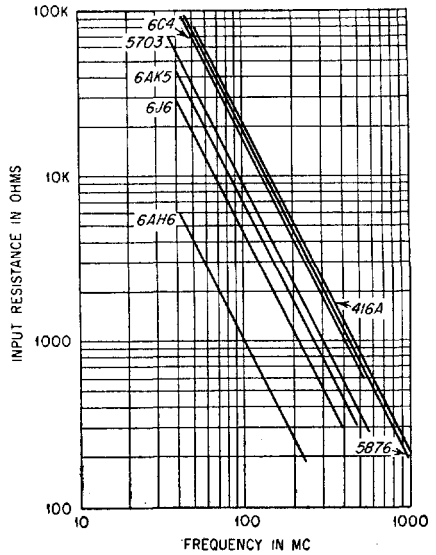
FIG. 7.34. Input resistance due to electron transit time and cathode lead inductance of several types of tubes as a function of frequency.

the sum of the individual conductances. The total input resistance due to electron transit time and cathode lead inductance effects is plotted as a function of frequency for several types of tubes used as r-f amplifiers in Fig. 7.34. The input conductance caused by cathode lead inductance is not a source of thermal noise since it is caused by feedback and not by a physical resistance. The detrimental effect of the conductance produced by cathode lead inductance is the reduction in the available power gain of the stage. Although this conductance contributes no noise itself, it is difficult to isolate from the input conductance caused by electron transit time, and the noise figure of an amplifier is normally calculated assuming that the total conductance due to both of these effects is caused by electron transit time.

2. Grid-induced Shot-effect Noise. In Sec. 7.2h it was shown that electron transit time causes a grid-induced shot-effect noise power which increases in direct proportion to the value of grid conductance caused by electron transit time. Since the grid conductance increases as the square of frequency, the grid-induced shot-effect noise power also increases as the square of frequency. This effect causes the noise figure of vacuum-tube amplifiers to increase with frequency and, together with grid con-

ductance, limits the usefulness of conventional vacuum tubes to frequencies below about 500 Mc.

3. Effect of Input Loading on Maximum Gain and Maximum Frequency of Operation. The maximum frequency at which a tube can be used successfully as an amplifier is normally limited to those frequencies at which the tube has a gain greater than unity. Since the input conductance of a tube increases as the square of frequency because of both transit time and cathode lead inductance effects, the maximum useful frequency at which the tube may be used will be limited by grid conductance.
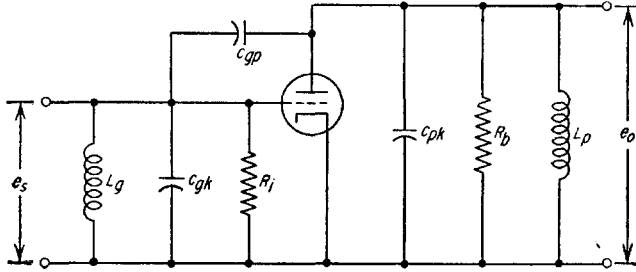
FIG. 7.35. Circuit of grounded-cathode tuned amplifier.

If the interstage coupling network in a tuned amplifier is assumed to be an ideal transformer and if there is no external load resistance on either primary or secondary, the maximum power gain possible from one stage is

$$G_{\max} = \frac{g_m{}^2 r_p R_i}{4} \tag{7.121}$$

where g_m = tube transconductance, mhos
R_i = input resistance of stage, ohms
r_p = output resistance of stage, ohms

4. Tube Resonance Effects at High Frequencies. The equivalent input circuit for a tube including cathode lead inductance is illustrated in Fig. 7.32. The grid-cathode capacitance is in series with the inductance of the cathode lead across the tube input. A maximum frequency limitation is placed upon tube operation at the frequency where these two reactances present a series-resonant circuit to the signal source. For the 6AK5 tube the resonant frequency of c_{gk} and L_k is approximately 500 Mc, and for the 6AC7 tube the resonant frequency is about 250 Mc.

7.4i. Circuit Configurations. The performance to be expected from a particular type of tube as an r-f or an i-f amplifier at the higher frequencies is largely determined by the specific circuit in which the tube is used. The three general classifications of circuits are the grounded cathode, grounded grid and grounded plate (cathode follower).

1. Grounded-cathode Amplifier. The grounded-cathode circuit for a tuned amplifier is shown in Fig. 7.35. The circuit applies to either a triode or a pentode with the screen and suppressor bypassed to the cathode. If ideal transformer coupling is assumed between stages, the maximum power gain is given by Eq. (7.121). If several identical stages are cascaded without benefit of transformer coupling and if the plate-load resistance R_L is much larger than the grid-input resistance of each stage, the power gain G per stage is

$$G = \frac{g_m{}^2}{(1/r_p + 1/R_i)^2} \tag{7.122}$$

where g_m = grid-plate transconductance, mhos

  r_p = dynamic plate resistance

  R_i = grid input resistance to each stage

The maximum power gain versus frequency is given for several commonly used tubes in Fig. 7.36 for the condition where the loading present in the interstage circuit is only that due to the effects of electron transit time and cathode lead inductance.
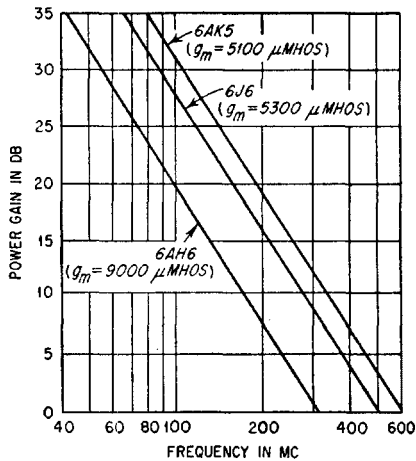
The grid-plate capacitance provides a feedback path between the plate and grid circuits which may cause an amplifier to oscillate. The conditions necessary for oscillation to occur in a tuned-plate tuned-grid stage are discussed in Sec. 6.2c. The frequency of oscillation will be lower than the resonant frequency of either the plate circuit or the grid circuit since the net reactance between the grid and cathode and between the plate and cathode must be inductive to achieve the required 180° phase shift across the grid-to-plate capacitive reactance. Oscillation is more likely to occur if the plate circuit is resonant at a higher frequency than the grid circuit since the tube gain required for oscillation is less in this case (see Sec. 6.2c). When the grid and plate circuit impedances have values such that conditions for oscillation are optimized, the maximum frequency of operation as an amplifier is limited to approximately

$$f = \frac{1}{\pi g_m c_{gp} R_g R_b} \tag{7.123}$$

where f = maximum frequency of operation, cps

  g_m = tube grid-plate transconductance, mhos

  c_{gp} = grid-plate capacitance of tube, farads

  R_g = parallel combination of tube-input resistance due to cathode lead inductance and electron transit time and any external resistance in grid circuit

  R_b = plate load resistance, ohms

Equation (7.123) assumes that the grid circuit is tuned to resonance at the lower −3-db frequency of the plate resonant circuit and that the reactance of c_{gp} is much larger than the net reactance of the plate circuit at the grid circuit resonant frequency. Equation (7.123) establishes the frequency of amplifier operation at which the gain around the feedback path provided by c_{gp} will be unity for the stated conditions. Since the phase shift through the feedback path is not 180° for these conditions, the amplifier will not oscillate at this exact frequency. The amplifier cannot oscillate unless the condition of unity gain around the closed loop formed by c_{gp}, the plate and grid resonant circuits, and the tube and the condition of 0° phase shift around this loop are simultaneously met at one frequency (see Sec. 6.2). It is usually desirable to keep the frequency of operation of a particular tube well below the value indicated in Eq. (7.123) to avoid distortion of the amplifier frequency response. In practice, the maximum frequency is usually limited to approximately 10 per cent of the value given by Eq. (7.123) unless precautions are taken to neutralize the grid-plate capacitance of the tube.[1]   For the special case where $R_g = R_b$, Eq. (7.123) can be rewritten

FIG. 7.36. Maximum power gain without transformer coupling for various tubes as limited by grid conductance.

[1] S. N. Van Voorhis, "Microwave Receivers," p. 126, Radiation Laboratory Series, vol. 23, McGraw-Hill Book Company, Inc., New York, 1948.

to express the maximum allowable voltage gain per stage as a function of c_{gp} and frequency.

$$A_{max} = \sqrt{\frac{g_m}{\pi f c_{gp}}} \qquad (7.124)$$

Equation (7.124) is plotted for several commonly used tubes in Fig. 7.37. Unless the grid-plate capacitance is neutralized, the maximum gain should be kept about

TUBE TYPE	c_{gp} $\mu\mu F$	g_m $\mu MHOS$
6AK5	0.02	5100
6AH6	0.02	9000
6AC7	0.015	9000
6J6	1.8	5300
6C4	2.0	3100
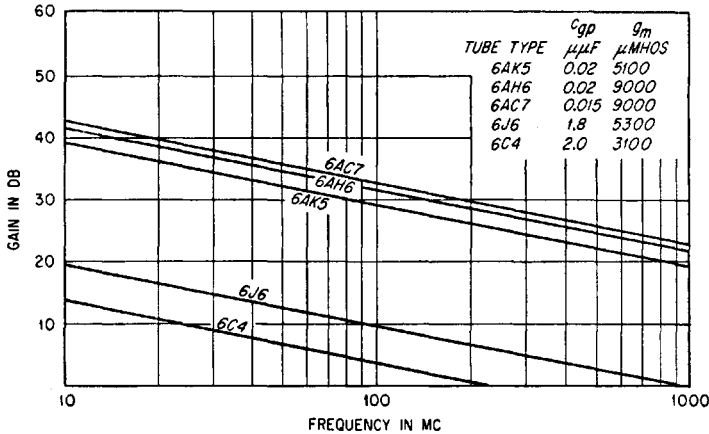
Fig. 7.37. Maximum voltage gain of various tubes as limited by c_{gp}.

10 db less than the values indicated in Fig. 7.37. This is equivalent to limiting the frequency of operation to approximately 10 per cent of the value given by Eq. (7.123).

A number of methods of neutralizing the grid-plate capacitance of a tube have been developed.[1] The principal method used in radio receivers at high frequencies is illustrated in Fig. 7.38. In this circuit the grid-plate capacitance of the tube is tuned out at the signal frequency by paralleling c_{gp} with an inductance and d-c blocking capacitor. The neutralizing inductance and c_{gp} are parallel resonant at the signal frequency. This

Fig. 7.38. Coil neutralization of grid-plate capacitance.

method of neutralization is adequate only over a relatively narrow band of frequencies.

The noise figure of a grounded-cathode amplifier is given by Eq. (7.125).

$$F = 1 + \frac{R_s}{R_1} + \frac{R_s T_c'}{R_t T} + \frac{R_{eq}}{R_s}\left(1 + \frac{R_s}{R_1} + \frac{R_s}{R_t}\right)^2 \qquad (7.125)$$

where F = noise figure of stage expressed as a power ratio

R_s = source resistance, ohms

R_t = grid resistance due to transit-time effects, ohms

R_1 = any external resistance shunting input circuit, including equivalent shunt resistance due to finite Q of input circuit inductance or transformer

R_{eq} = equivalent grid-noise resistance, ohms (see Sec. 7.2h)

T_c' = effective temperature of R_t, °K

T = room temperature, °K

[1] See F. E. Terman, "Radio Engineering," pp. 367–369, McGraw-Hill Book Company, Inc., New York, 1947.

At low frequencies where R_t is sufficiently large to be neglected, Eq. (7.125) reduces to

$$F = 1 + \frac{R_s}{R_1} + \frac{R_{eq}}{R_s}\left(1 + \frac{R_s}{R_1}\right)^2 \qquad (7.125a)$$

If R_t must be taken into account but R_1 is large enough to be neglected, Eq. (7.125) reduces to

$$F = 1 + \frac{R_s T_c'}{R_t T} + \frac{R_{eq}}{R_s}\left(1 + \frac{R_s}{R_t}\right)^2 \qquad (7.125b)$$

The value of R_s yielding minimum noise figure is found to be

$$R_s(\text{optimum}) = \frac{R_a}{\sqrt{1 + \dfrac{R_a^2}{R_1 R_{eq}}\left(1 + \dfrac{R_1 T_c'}{R_t T}\right)}} \qquad (7.126)$$

where $R_a = R_1 R_t/(R_1 + R_t)$

(a)

(b)
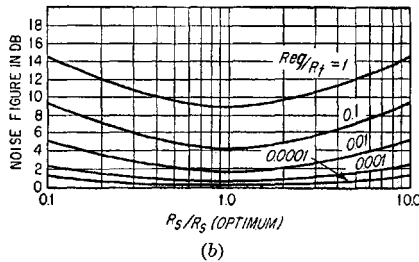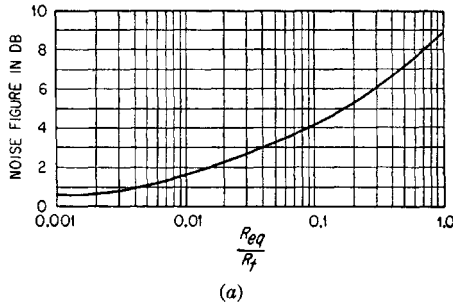
FIG. 7.39. Noise figure of a grounded-cathode amplifier. (a) Noise figure with optimum source resistance. (b) Variation in noise figure with nonoptimum source resistance.

When R_t can be neglected, differentiation of Eq. (7.125a) yields an optimum source resistance given by

$$R_s(\text{optimum}) = \frac{R_1}{\sqrt{1 + R_1/R_{eq}}} \qquad (7.126a)$$

When R_1 can be neglected, differentiation of Eq. (7.125b) yields an optimum source resistance given by

$$R_s(\text{optimum}) = \frac{R_t}{\sqrt{1 + T_c' R_t/T R_{eq}}} \qquad (7.126b)$$
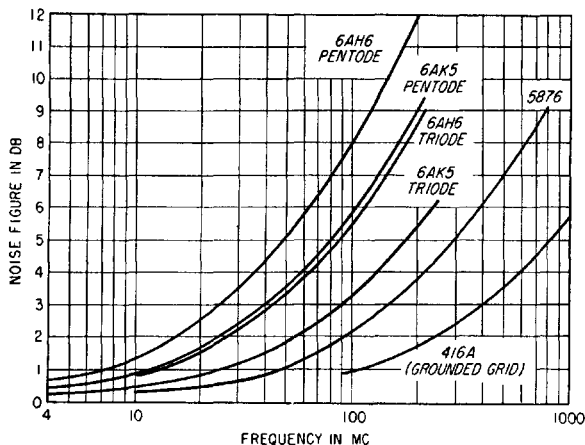
FIG. 7.40. Optimum noise figures of several tube types as grounded-cathode amplifiers.

If Eq. (7.126b) is substituted into Eq. (7.125b) to eliminate R_s, the noise figure becomes

$$F = 1 + \frac{2R_{eq}}{R_t} + 2\sqrt{\frac{T_c' R_{eq}}{T R_t} + \left(\frac{R_{eq}}{R_t}\right)^2}$$ (7.127)

For tubes having oxide-coated cathodes, the ratio T_c'/T is approximately 5. If this value is used in Eq. (7.127), the noise figure at any frequency can be expressed as a function of the ratio R_{eq}/R_t as plotted in Fig. 7.39a. The variation in noise figure as a function of source resistance is given in Fig. 7.39b.

The noise figures of several types of tubes as grounded-cathode amplifiers have been plotted in Fig. 7.40.

2. Grounded-grid Amplifier. The circuit of a grounded-grid amplifier is shown in Fig. 7.41. The input impedance of such an amplifier, assuming all reactances to be tuned out, is given by

$$R_i = \frac{R_t(r_p + R_b)}{R_t(\mu + 1) + r_p + R_b}$$ (7.128)
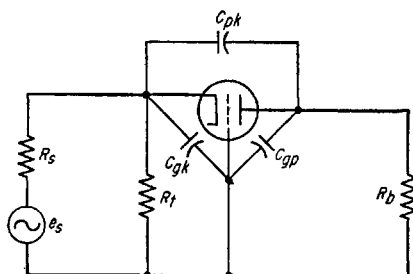
FIG. 7.41. Circuit of grounded-grid amplifier.

where R_t = input resistance in ohms due to transit time
  R_b = load resistance, ohms
  r_p = plate resistance, ohms
  μ = amplification factor

The tube input capacitance which must be resonated at the desired center frequency is

$$C_i = c_{gk} + c_{fk} + (1 - A \cos \phi)c_{pk}$$ (7.129)

where c_{gk} = grid-cathode capacitance of tube
  c_{fk} = filament-cathode capacitance of tube
  c_{pk} = plate-cathode capacitance of tube
  A = voltage gain of amplifier
  ϕ = phase angle of output voltage caused by reactive load, that is, $\phi = \phi_1 - \phi_2$
    where ϕ_1 is phase angle of load and ϕ_2 is phase angle of load impedance plus r_p

Note that the cathode-plate capacitance reduces the input capacitance if the plate impedance is resistive provided the voltage gain is greater than unity. The output impedance R_o of a grounded-grid amplifier, assuming all reactances to be tuned out, is given by

$$R_o = \frac{R_b[r_p(R_t + R_s) + (\mu + 1)(R_tR_s)]}{(R_t + R_s)(r_p + R_b) + R_tR_s(\mu + 1)} \qquad (7.130)$$

where R_b = plate-load resistance
R_t = tube input resistance due to electron transit time
R_s = signal-source resistance

The primary advantage of the grounded-grid amplifier is its stability at high frequencies. A triode can be operated at much higher frequencies without instability in a grounded-grid circuit than in a grounded cathode circuit having the same voltage gain.

The conditions necessary for oscillation in the grounded-grid amplifier are

$$X_{pk} + X_{pg} + X_{gk} = 0 \qquad (7.131)$$

and, assuming no circuit loading,

$$\mu \geq -\frac{X_{pk}}{X_{gk}} \qquad (7.132)$$

where μ = tube amplification factor
X_{pk} = plate-cathode reactance = $-1/\omega c_{pk}$
X_{pg} = net reactance of output circuit
X_{gk} = net reactance of input circuit

In the grounded-grid circuit, the plate-cathode reactance is capacitive (due to c_{pk}), and therefore the plate-grid reactance must be inductive and the grid-cathode reactance must be capacitive to obtain the required 180° phase shift between the plate and grid circuit for oscillation. This requires that the resonant frequency of the output circuit (between grid and plate) be higher than the resonant frequency of the input circuit (between grid and cathode) for the reactance between plate and grid to be inductive at frequencies where the grid-cathode reactance is capacitive. By tuning the plate circuit to a frequency slightly lower than the grid resonant frequency, the plate-to-grid reactance will become capacitive at a lower frequency than that at which the grid-to-cathode reactance becomes capacitive, and the condition necessary for oscillation cannot be satisfied. Instability of grounded-grid amplifiers at very high frequencies is usually due to grid-lead inductance between the grid structure within the tube and the point at which the grid lead is grounded. Grid lead inductance provides a common reactance between the input and output circuits which allows the phase shift required for oscillation to be obtained regardless of the relative tuning of the input and output circuits. For this reason, all tubes designed for operation as grounded-grid amplifiers at very high frequencies either employ a special tube structure to minimize the grid lead inductance, or, if conventional glass envelope construction is used, the internal grid structure is connected to three or more pins at the tube base, thus minimizing grid lead inductance by paralleling the multiple grid leads.

The voltage gain of a grounded-grid amplifier is given by

$$A = \frac{e_o}{e_s} = \frac{R_tR_b(\mu + 1)}{R_sR_t(\mu + 1) + (r_p + R_b)(R_s + R_t)} \qquad (7.133)$$

where e_o = output voltage
e_s = source open-circuit voltage

The source impedance R_s is ordinarily predetermined, and the load impedance R_b

is usually specified by the required bandwidth and the interstage capacitance. When the source impedance and the load impedance do not match the input and output impedances, respectively, of the grounded-grid stage, it is necessary to insert imped-ance transformers between the source and the stage and between the stage and the load to obtain maximum power gain. Since the input and output impedances of the stage are not independent, the equations for the impedance transformation ratios for the two transformers must be solved simultaneously. The resulting ratios are

$$N_1{}^2 = \frac{R_t}{R_s \sqrt{1 + R_t(\mu + 1)/r_p}} \tag{7.134}$$

$$N_2{}^2 = \frac{R_b}{r_p \sqrt{1 + R_t(\mu + 1)/r_p}} \tag{7.135}$$

where N_1, N_2 = required output-to-input voltage ratios which must be achieved with input and output matching transformers, respectively
   $N_1{}^2$ = ratio of tube input impedance to source impedance
   $N_2{}^2$ = ratio of load impedance to tube output impedance

The equivalent circuit of a grounded-grid amplifier including all the applicable noise sources is shown in **Fig. 7.42.**
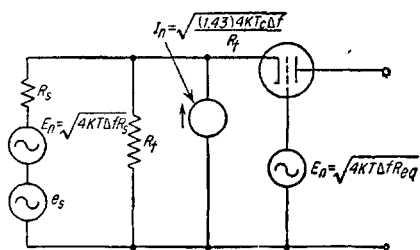
FIG. 7.42. Grounded-grid amplifier circuit including noise sources.
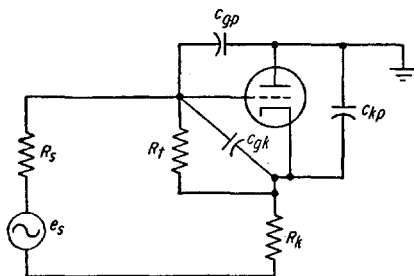
FIG. 7.43. Circuit of cathode-follower amplifier.

The noise figure of a grounded-grid amplifier including the effects of electron transit time is given by

$$F = 1 + \frac{R_s' T_c'}{R_t T} + \frac{R_{eq}}{R_s'} \left(\frac{\mu}{\mu + 1}\right)^2 \left(1 + \frac{R_s'}{R_t}\right)^2 \tag{7.136}$$

where T_c' = equivalent temperature, in degrees Kelvin of transit-time resistance R
   = 1.43 times cathode temperature $\simeq 5T$ for tubes with oxide-coated cathodes
   T = room temperature, °K
   R_t = grid-cathode resistance due to electron transit time
   R_s' = transformed source resistance, ohms = $N_1{}^2 R_s$
   R_{eq} = equivalent grid-noise resistance, ohms (see Sec. 7.2h)

The value of the transformed source resistance which will make the grounded-grid amplifier noise figure a minimum is given by

$$R_s'(\text{optimum}) = \frac{R_t}{\sqrt{1 + R_t T_c'(\mu + 1)^2/R_{eq} T \mu^2}} \tag{7.137}$$

3. *Grounded-plate Amplifier.* The circuit of a grounded-plate amplifier or cathode follower is shown in **Fig. 7.43.** The voltage gain of the cathode follower amplifier

at frequencies where grid loading due to transit time is important is given by

$$A = \frac{e_o}{e_s} = \frac{R_k(r_p + \mu R_t)}{r_p(R_k + R_s + R_t) + R_k[R_s + R_t(\mu + 1)]} \qquad (7.138)$$

where e_o = output voltage
      e_s = source voltage
      R_k = cathode load resistance
      R_t = grid resistance due to transit time
      R_s = resistance of signal source
      r_p = plate resistance

The input resistance R_i of the cathode follower amplifier at high frequencies is given by

$$R_i = \frac{R_t}{1 - A} \qquad (7.139)$$

where A = gain of cathode-follower amplifier as given by Eq. (7.138) when $R_s = 0$
The input capacitance of a cathode-follower amplifier is given by

$$C_i = c_{gp} + (1 - A)c_{gk} \qquad (7.140)$$

where c_{gp} = grid-plate capacitance
      c_{gk} = grid-cathode capacitance
      A = gain of cathode-follower amplifier as given by Eq. (7.138) when $R_s = 0$
The output resistance R_o of the cathode-follower amplifier at high frequencies, assuming circuit and tube reactances to be tuned out is

$$R_o = \frac{r_p R_k'}{r_p + R_k'[1 + \mu R_t/(R_s + R_t)]} \qquad (7.141)$$

where $R_k' = \dfrac{(R_t + R_s)R_k}{R_t + R_s + R_k}$

The equivalent circuit of the cathode-follower amplifier including noise sources is

FIG. 7.44. Circuit of cathode-follower amplifier including noise sources.

shown in Fig. 7.44. The noise figure of the cathode-follower amplifier at frequencies where R_t is much larger than either R_s or R_k is given by
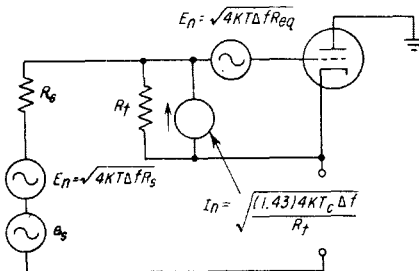
$$F = 1 + \frac{R_{eq}}{R_s} \qquad (7.142)$$

where R_{eq} = equivalent grid noise resistance
At frequencies where R_t is of the same order of magnitude as R_k and R_s, the expression for the noise figure of a cathode follower is complex and is not treated here.[1]

7.4j. Transient Response in Tuned Amplifiers. The response of tuned amplifiers to pulses is very important in the design of an r-f or i-f amplifier for television or radar use. The requirements of pulse-rise time and overshoot may determine the type of interstage coupling networks which must be used in such an amplifier.

If the fractional bandwidth $\Delta f/f_0$ of a tuned amplifier is small, the response characteristic exhibits essentially arithmetic symmetry about the center frequency f_0.

[1] For information on this subject see G. E. Valley, Jr., and H. Wallman, "Vacuum Tube Amplifiers," Radiation Laboratory Series, vol. 18, McGraw-Hill Book Company, Inc., New York, 1948.

For this case, the effect of the amplifier passband upon the envelope of a modulated carrier at the center frequency f_0 will be very nearly the same as the effect of a low-pass equivalent of the amplifier passband on a video pulse having the same amplitude versus time response as the envelope of the modulated carrier. For small fractional bandwidths, a low-pass equivalent of the amplifier passband is a low-pass filter having a response defined by the upper half of the amplifier-passband characteristic when shifted down to zero frequency.

The response of an amplifier consisting of n synchronous single-tuned stages and having a small fractional bandwidth to an input carrier f_0 modulated with a step function will have no overshoot. The response is equivalent to the response of n cascaded RC networks having one-half the bandwidth of the synchronous single-tuned stages to a video step-function input. The relation between rise time and bandwidth for any number of cascaded synchronously tuned stages is given with an accuracy of ± 10 per cent by

$$T \, \Delta f_t = 0.70 \qquad (7.143)$$

where T = pulse-rise time in microseconds between the 10 and 90 per cent points

Δf_t = over-all bandwidth in megacycles of the n cascaded stages

The determination of the rise-time response of amplifiers having poles (see Sec. 18.4b) at more than one frequency, e.g., staggered pairs, staggered triples, double-tuned circuits, etc., is considerably more complicated than for synchronous single-tuned circuits even assuming the fractional bandwidth to be small. The response of a multipole network to a carrier modulated by a step function will have an overshoot. The amount of the overshoot is a function of the ratio of the separation of the circuit resonant frequencies to the individual circuit bandwidths for staggered-tuned circuits and is a function of the coefficient of coupling and circuit Q's for double-tuned circuits. For overstaggered stages or overcoupled double-tuned circuits, the overshoot is also a function of the peak-to-dip ratio of the frequency-response curve.[1]

7.4k. Time Delay in Tuned Amplifiers. Tuned circuits introduce a phase shift to an applied signal which varies as a function of the signal frequency relative to the resonant frequency of the circuit. When a modulated carrier is impressed on a parallel resonant circuit, the phase shift of the carrier and the sidebands will be a function of the circuits Q and the ratio of the carrier frequency to the resonant frequency of the circuit. If the circuit is tuned to the carrier frequency, the phase shift of the carrier will be zero and the upper and lower sidebands will have lagging and leading phase shifts, respectively. Because the sideband signals are shifted in phase with respect to the carrier, the modulation envelope will also be shifted in phase. This phase shift of the envelope is equivalent to a time delay of the envelope in passing through the circuit. If the upper and lower sidebands incur phase shifts relative to the carrier of equal magnitude and opposite sign, the time delay of the modulation is given by

$$T_m = \frac{\Delta \beta}{\omega_m} \qquad (7.144)$$

where T_m = delay time, sec

$\Delta \beta$ = phase shift in radians of lower sideband relative to carrier. $\Delta \beta$ is positive when lower sideband incurs a leading phase shift relative to carrier

ω_m = modulating frequency, radians/sec.

The envelope delay of a number of modulating frequencies will be constant only if the phase shift of the tuned amplifier is a linear function of the frequency displace-

[1] For greater detail concerning transient response of bandpass amplifiers see Valley and Wallman, "Vacuum Tube Amplifiers," vol. 18, Radiation Laboratory Series, McGraw-Hill Book Company, Inc., New York, 1948.

ment from the carrier frequency. With the exception of the sign of the phase shift, the phase shift must be the same for the upper and lower sidebands.

When the input signals to a tuned amplifier consist of rectangular pulses, the modulation envelope is composed of an infinite number of sidebands having a $(\sin x)/x$ amplitude distribution. The input pulse will be reproduced exactly at the output with some time delay only if $\Delta\beta/\Delta\omega$ is constant throughout the amplifier passband and if the amplifier bandwidth is great enough to pass all the modulation sidebands. In practice, neither of these conditions can be met, and as a result the output pulse is distorted because of both amplitude and phase distortion. The time by which a pulse is delayed in passing through a tuned amplifier is given within an accuracy of 15 per cent by the following equations where the time delay T_d is defined as the delay of the pulse at the 50 per cent amplitude point.

$$T_d = \frac{n}{4\,\Delta f} \qquad \mu\text{sec} \qquad \text{(cascaded synchronously tuned stages)} \qquad (7.145)$$

where n = number of cascaded stages
Δf = bandwidth of each stage, Mc

$$T_d = \frac{n}{1.88\,\Delta f} \qquad \mu\text{sec} \qquad \text{(cascaded flat-staggered pairs or transitionally} \qquad (7.146)$$
coupled double-tuned stages)

where n = number of cascaded pairs or double-tuned stages
Δf = bandwidth of each pair or double-tuned stage, Mc

$$T_d = \frac{n}{1.35\,\Delta f} \qquad \text{(cascaded flat-staggered triples)} \qquad (7.147)$$

where n = number of cascaded triples
Δf = bandwidth of each triple, Mc

Equations (7.145) through (7.147) are derived on the basis that $\Delta\beta/\Delta\omega$ is constant throughout the amplifier passband and equal to $\Delta\beta/\Delta\omega$ as measured between the -3-db frequencies of the amplifier passband.

7.5. Mixers. The fundamental principle of the superheterodyne receiver is the conversion of the received signal at some radio frequency to a signal at an intermediate frequency where it is amplified and then demodulated. The process of frequency conversion is known as heterodyne action. In the mixer circuit, a local oscillator signal is modulated by the r-f signal to produce modulation frequencies which are the sum and difference of the frequencies of the r-f signal and the local oscillator. Signals will also be produced at the sum and difference frequencies of the r-f signal and all harmonics of the local oscillator. One of the modulation signals is selected and amplified and is referred to as the intermediate frequency signal. If the functions of mixing and generating the local oscillator signal occur within the same tube, the tube is ordinarily referred to as a converter.

Intermodulation between two signals can occur only in a nonlinear element (see Sec. 5.1); therefore mixers and converters are necessarily nonlinear devices.

7.5a. Triode and Pentode Mixers. Triodes and pentodes can be used as mixers under certain operating conditions. The important factors to be considered in their use as mixers are (1) conversion transconductance, (2) method of local oscillator injection, (3) input and output impedances, (4) conversion gain, (5) noise figure, and (6) special bias considerations.

1. Conversion Transconductance. The fundamental property of a mixer tube is its conversion transconductance g_c.

$$g_c = \frac{I_{if}}{E_{rf}} \qquad (7.148)$$

where I_{if} = peak output current of mixer at intermediate frequency

E_{rf} = peak signal input voltage at signal frequency

In conventional mixer operation, the peak local-oscillator voltage is approximately equal to the tube cutoff voltage and is large compared to the r-f signal. Under these conditions, g_m will be a function only of the applied local-oscillator voltage. An example of how g_m may vary with local-oscillator signal is shown in Fig. 7.45. The variations in g_m with time as a function of local-oscillator signal can be expressed as a Fourier series.

$$g_m = a_o + a_1 \cos \omega_{lo}t$$
$$+ a_2 \cos 2\omega_{lo}t + \ldots \quad (7.149)$$

where ω_{lo} = local-oscillator frequency, radians/sec

The constants in Eq. (7.149) can be evaluated by the Fourier coefficient equations

$$a_o = \frac{1}{2\pi} \int_0^{2\pi} g_m(t)\, d\,(\omega_{lo}t) \quad (7.150)$$

$$a_n = \frac{1}{\pi} \int_0^{2\pi} g_m(t)\, \cos\,(n\omega_{lo}t)\, d\,(\omega_{lo}t)$$
$$(7.151)$$
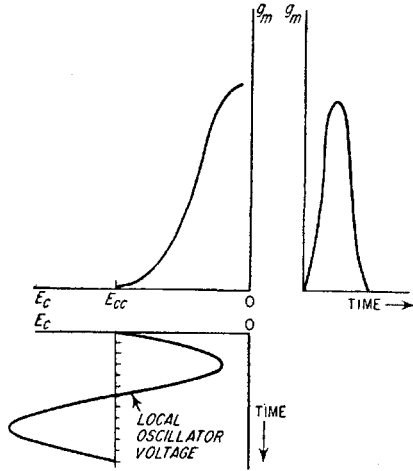
FIG. 7.45. Mixer transconductance versus local-oscillator signal.

When an r-f signal is applied to the tube grid in the presence of a local-oscillator signal, the resultant instantaneous plate current is given by

$$i_p = g_m E_{rf} \sin\,(\omega_{rf}t) \quad (7.152)$$

$$= a_o E_{rf} \sin\,(\omega_{rf}t) + E_{rf} \sum_{n=1}^{\infty} a_n \sin\,(\omega_{rf}t)\, \cos\,(n\omega_{lo}t)$$

$$= a_o E_{rf} \sin\,(\omega_{rf}t) + \tfrac{1}{2} E_{rf} \sum_{n=1}^{\infty} a_n \sin\,(\omega_{rf} + n\omega_{lo})t$$

$$+ \tfrac{1}{2} E_{rf} \sum_{n=1}^{\infty} a_n \sin\,(\omega_{rf} - n\omega_{lo})t \quad (7.153)$$

If the intermediate frequency selected is $\omega_{rf} \pm n\omega_{lo}$, the conversion transconductance of the tube is given by

$$g_{cn} = \frac{I_{(\omega_{rf} \pm n\omega_{lo})}}{E_{rf}} \quad (7.154)$$

where $I_{(\omega_{rf} \pm n\omega_{lo})}$ = peak value of i-f current out of mixer

g_{cn} = conversion transconductance as established by r-f signal and local-oscillator nth harmonic

Thus, the only term of interest in Eq. (7.153) is the term containing $I_{(\omega_{rf} \pm n\omega_{lo})}$. The peak value of plate current at this frequency is $a_n E_{rf}/2$, and Eq. (7.154) becomes

$$g_{cn} = \frac{a_n}{2}$$

or, substituting for a_n the value of the Fourier coefficient given by Eq. (7.151),

$$g_{cn} = \frac{1}{2\pi} \int_0^{2\pi} g_m(t) \cos(n\omega_{lo}t) \, d(\omega_{lo}t) \tag{7.155}$$

In most applications the intermediate frequency is established as the difference between the local-oscillator fundamental frequency and the signal frequency. For this condition n equals 1 and the conversion transconductance g_c reduces to

$$g_c = \frac{1}{2\pi} \int_0^{2\pi} g_m(t) \cos(\omega_{lo}t) \, d(\omega_{lo}t) \tag{7.156}$$

It should be noted that even though the applied local-oscillator signal may contain no harmonics, all the terms in Eq. (7.153) will be present in the tube plate current since the harmonics of the local-oscillator frequency are generated within the tube itself.

The conversion transconductance of a mixer tube can be calculated graphically from the g_m versus E_c characteristic of the tube for an assumed local-oscillator voltage and tube bias.[1] g_m is plotted as a function of the applied local-oscillator voltage as shown in Fig. 7.45. The value of g_m is measured for d-c bias voltages equal to the amplitude of the local-oscillator signal at 30° intervals starting at the negative peak of the applied local-oscillator voltage waveform and ending at the positive peak. Seven values of g_m are obtained by this procedure. The approximate conversion transconductance for the case where the i-f frequency is established by the local-oscillator fundamental, that is, $\omega_{if} = \omega_{rf} - \omega_{lo}$, is given by

$$g_{c1} = \frac{1}{12}[(g_{m7} - g_{m1}) + (g_{m5} - g_{m3}) + 1.73(g_{m6} - g_{m2})] \tag{7.157}$$

The approximate conversion transconductances for the intermediate frequencies of $\omega_{if} = \omega_{rf} - 2\omega_{lo}$ and $\omega_{if} = \omega_{rf} - 3\omega_{lo}$ are given, respectively, by Eqs. (7.158) and (7.159).

$$g_{c2} = \frac{1}{12}[2g_{m4} + \tfrac{3}{4}(g_{m3} + g_{m5} - g_{m6} - g_{m2}) - (g_{m7} + g_{m1})] \tag{7.158}$$

$$g_{c3} = \frac{1}{12}[(g_{m7} - g_{m1}) - 2(g_{m5} - g_{m3})] \tag{7.159}$$

The maximum value of g_{c1} occurs when the terms $(g_{m7} - g_{m1})$, $(g_{m5} - g_{m3})$, and $(g_{m6} - g_{m2})$ are maximum. This occurs when the amplitudes of the applied local-oscillator voltage and the grid bias are adjusted so that the g_m is 0 for points g_{m1}, g_{m2}, and g_{m3} and maximum for points g_{m5}, g_{m6}, and g_{m7}. This corresponds to cutting the plate current off for slightly less than one-half cycle of the local-oscillator signal. If the amplitude of the applied local-oscillator signal is insufficient to meet this condition, maximum conversion transconductance is achieved when the tube grid bias is adjusted to the point of maximum slope of the g_m versus E_c characteristic for the tube. The maximum conversion transconductance at the local-oscillator fundamental is equal to approximately one-quarter of the maximum grid-plate transconductance of the tube.[2] This approximation is usually within 15 per cent and is based on the assumptions that g_m increases linearly with E_c as the grid voltage is increased from the cutoff value toward zero bias and that the grid bias and the peak local-oscillator voltage are equal to the cutoff value for the tube.

Example 7.10

Determine the maximum conversion transconductance at the local-oscillator fundamental frequency for the 6AH6 tube as a pentode with a screen-grid voltage of 150 volts and

---

[1] E. W. Herold, The Operation of Frequency Converters and Mixers for Superheterodyne Receivers, *Proc. IRE*, February, 1942.

[2] For more complicated mixer techniques in which this value can theoretically be more than doubled, see E. W. Herold, Superheterodyne Frequency Conversion Using Phase Reversal Modulation, *Proc. IRE*, April, 1946.

plate voltage of 300 volts. Find the associated values of grid bias and applied local-oscillator voltage. Assume that the grid cannot be allowed to go positive during any portion of the local-oscillator cycle. The g_m versus E_c characteristic of the 6AH6 tube operated as a pentode is shown in Fig. 7.46.

6AH6

$E_{bb} = 300$ VOLTS
$E_{c2} = 150$ VOLTS
$E_{c3} = 0$ VOLTS
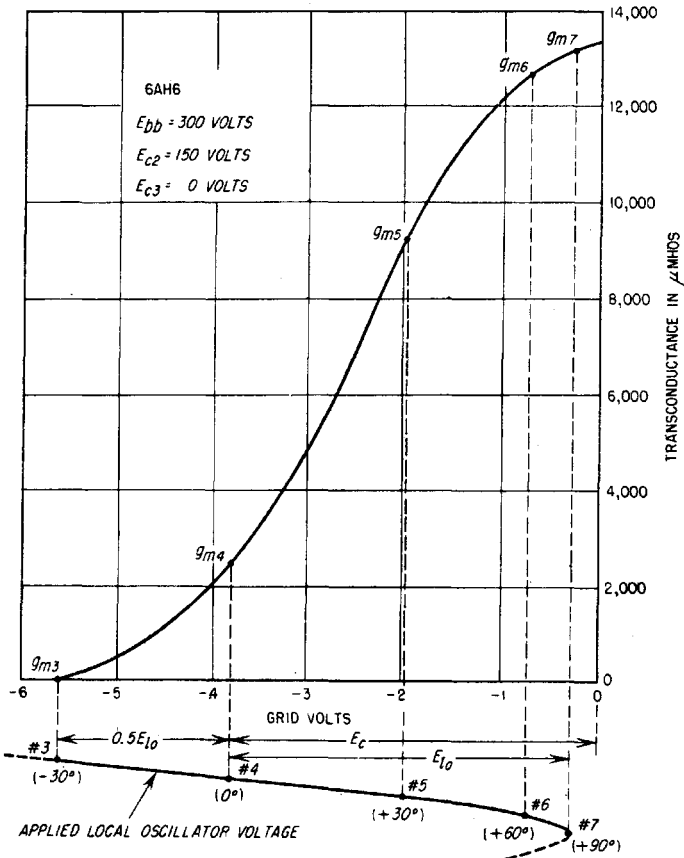
FIG. 7.46. Graphical determination of conversion transconductance.

Solution

1. Determine the peak local-oscillator voltage E_{lo}.
From Eq. (7.157) the values of g_m at points 1, 2, and 3 should be 0 if unlimited local-oscillator signal is available. Also, the values of g_m at points 5, 6, and 7 should be as large as possible. This means that point 7 should be very nearly at $E_c = 0$, for example, -0.25 volt, and that point 3 should be very nearly at cutoff. Since point 3 occurs at $-0.5E_{lo}$, that is, 60° from the negative peak of E_{lo}, the difference in voltage between the cutoff bias for the tube and -0.25 volt is equal to $1.5E_{lo}$. Cutoff occurs at -5.6 volts, therefore

$$E_{lo} = \frac{5.6 - 0.25}{1.5} = 3.57 \text{ volts}$$

2. Determine the bias voltage E_c.
The d-c bias is equal to

$$E_c = -3.57 + (-0.25) = -3.82 \text{ volts}$$

3. Determine the tube transconductance at each 30° interval of the local-oscillator voltage.
The values of g_m at the 30° intervals are

$$g_{m1} = 0, \ g_{m2} = 0, \ g_{m3} = 0, \ g_{m4} = 2,500 \ \mu\text{mhos}$$
$$g_{m5} = 9,300 \ \mu\text{mhos}, \ g_{m6} = 12,700 \ \mu\text{mhos}, \ g_{m7} = 13,500 \ \mu\text{mhos}$$

4. Determine the conversion transconductance at the local-oscillator fundamental. Substituting these values into Eq. (7.157),

$$g_{c1} = \tfrac{1}{12}[(g_{m7} - g_{m1}) + (g_{m5} - g_{m3}) + 1.73(g_{m6} - g_{m2})]$$
$$= \tfrac{1}{12}[13,500 + 9,300 + 1.73(12,700)]$$
$$= 3,730 \ \mu\text{mhos}$$

5. The value of conversion conductance calculated by dividing the peak g_m by 4 is

$$g_{c1} \simeq \frac{g_m}{4} = \frac{13,500}{4}$$
$$\simeq 3,380 \ \mu\text{mhos}$$

The approximate value is within 10 per cent of the actual value.

2. Local-oscillator Injection. In triode mixers, the local-oscillator voltage can be introduced into the grid of the mixer tube or it can be injected into the cathode circuit. In pentodes the local-oscillator signal can also be coupled into the suppressor-grid circuit or the screen-grid circuit. The r-f signal is ordinarily applied between the control grid and cathode.

(a) CONDUCTIVE COUPLING OF L-O VOLTAGE INTO MIXER
    GRID CIRCUIT

(b) CAPACITIVE COUPLING OF L-O VOLTAGE INTO
    MIXER GRID CIRCUIT

(c) INDUCTIVE COUPLING OF L-O VOLTAGE INTO
    MIXER GRID CIRCUIT

(d) BALANCED MIXER WITH INDUCTIVE COUPLING OF L-O VOLT-
    AGE INTO MIXER CIRCUIT
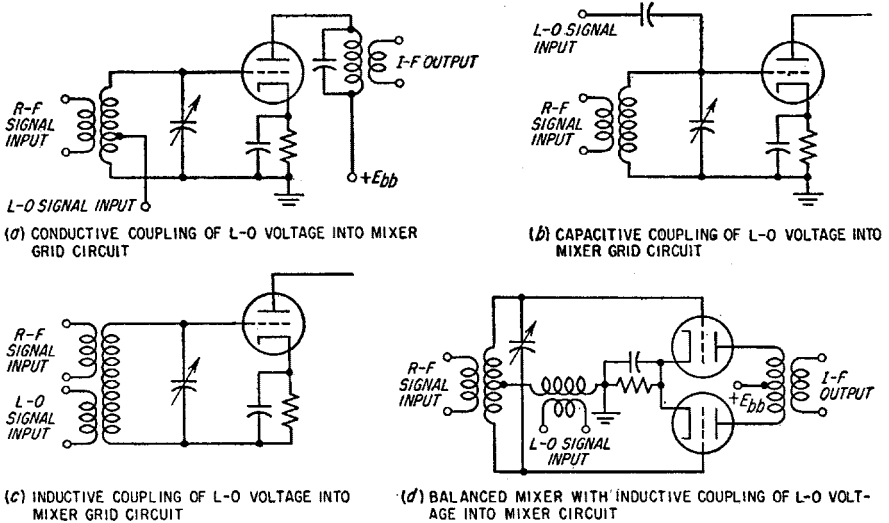
FIG. 7.47. Methods of grid circuit local-oscillator injection.

CONTROL-GRID INJECTION. The local-oscillator signal can be coupled to the mixer tube capacitively, inductively, or conductively and either in parallel or in series with the r-f signal voltage. Circuits illustrating some of the various methods of grid injection are shown in Fig. 7.47. In general, local-oscillator signal injection into the grid circuit is objectionable because of the interaction between the local-oscillator output and the r-f signal source. This practice can also be objectionable if the input to the mixer is connected directly to the receiver antenna since the radiation of the local-oscillator signal by the antenna can cause interference in other nearby receivers.

Since the local oscillator is displaced in frequency from the r-f signal by an amount equal to the i-f frequency, the mixer resonant circuit, tuned to the r-f signal, presents a reactance at the local-oscillator frequency. This reactance in conjunction with the local-oscillator output impedance and any coupling impedances forms a voltage divider for the local-oscillator signal. Although a constant-frequency difference is maintained between the mixer circuit resonant frequency and the local-oscillator frequency, the local-oscillator voltage applied to the mixer may vary with tuning because of changes in the relative impedances in the referenced voltage divider. This results in a variation in conversion transconductance with receiver tuning. The reactances which the local oscillator and r-f signal circuits present to each other cause detuning effects in both circuits and may make tracking of the local oscillator and r-f signal difficult.

Use of the balanced mixer circuit shown in Fig. 7.47d minimizes local-oscillator radiation and interaction between the signal and oscillator circuits.

(a) INDUCTIVE COUPLING OF LO SIGNAL TO MIXER CATHODE

(b) CAPACITIVE COUPLING OF LO SIGNAL TO MIXER CATHODE
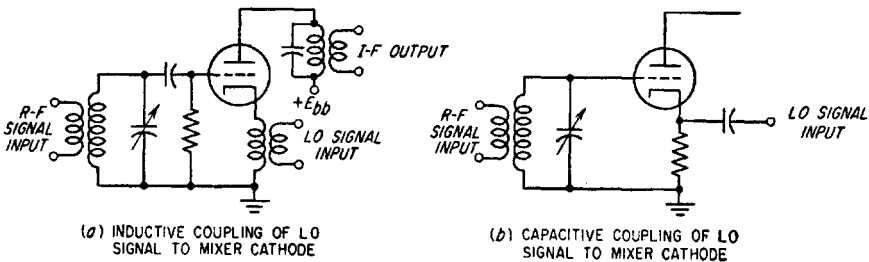
FIG. 7.48. Cathode injection of local-oscillator signal.

CATHODE INJECTION. The local-oscillator voltage can be introduced into the cathode circuit of the mixer tube as illustrated in Fig. 7.48. The use of cathode injection of the local-oscillator signal minimizes the local-oscillator voltage appearing across the r-f signal source and thus minimizes interaction between these two circuits. In addition, it reduces radiation of the local-oscillator signal if the input to the mixer stage is not isolated from the antenna.

One disadvantage of cathode injection is that the cathode input impedance of the mixer is relatively low, viz., approximately equal to the average value of $1/g_m$. If a straight-line g_m versus E_c characteristic is assumed, the average cathode input impedance $\bar{Z}_k$ for a complete cycle of the local-oscillator signal is given approximately by

$$\bar{Z}_k \simeq \frac{\pi}{g_m'} \qquad (7.160)$$

where g_m' = grid-plate transconductance at peak of local-oscillator cycle

If the local oscillator is tuned higher than the r-f signal, the local-oscillator circuit will appear inductive at the signal frequency. If this inductive reactance is coupled into the cathode circuit of the mixer tube, the feedback through the grid-cathode capacitance may severely load the tube input circuit (see Sec. 7.4h).

Additional loading of the r-f signal circuit can be caused by feedback through the grid-plate capacitance of the mixer if this capacitance is very large. If the i-f circuit shunt tuning capacitance is small, e.g., wideband television and radar i-f circuits, the mixer plate load may present a relatively high capacitive reactance at the signal radio frequency. The resulting feedback through c_{pg} may result in appreciable resistive loading of the mixer grid circuit. In a pentode mixer, the grid-plate capacitance is usually small enough to make this effect negligible. In triode mixers, some

form of neutralization of the grid-plate capacitance may be necessary to prevent excessive grid circuit loading.

3. Input and Output Equivalent Circuits. The input circuit of a triode or pentode

(*a*) EQUIVALENT INPUT CIRCUIT
OF A VACUUM TUBE MIXER

(*b*) EQUIVALENT OUTPUT CIRCUIT
OF VACUUM TUBE MIXER

I_p = I-F COMPONENT OF PLATE CURRENT
e_g = R-F GRID VOLTAGE
g_c = MIXER CONVERSION TRANSCONDUCTANCE
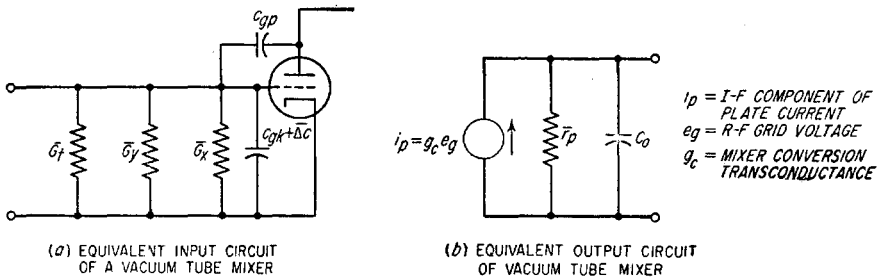
FIG. 7.49. Mixer equivalent circuits.

mixer is shown in Fig. 7.49. The tube input admittance Y_i to the r-f input signal is given by

$$Y_i = \bar{G}_t + \bar{G}_y + \bar{G}_x + j\omega_{rf}[c_{gk} + \overline{\Delta c} + c_{gp}(1 + \bar{A}\cos\theta)] \qquad (7.161)$$

where $\bar{G}_t$ = grid-input conductance due to electron transit time averaged over local-oscillator cycle

$\bar{G}_y$ = grid-input conductance due to cathode circuit inductance averaged over local-oscillator cycle

$\bar{G}_x$ = average grid conductance due to capacitive reactance in plate circuit and feedback through grid-to-plate capacitance c_{gp}

$\quad = -\omega_{rf}c_{gp}\bar{A}\sin\theta$

$\overline{\Delta c}$ = increase in grid-cathode capacitance due to electron transit time averaged over local-oscillator cycle

$\bar{A}$ = absolute voltage gain of mixer at r-f signal frequency averaged over local-oscillator cycle

$\theta = \theta_1 - \theta_2$ where θ_1 is the phase angle of the load impedance at the radio frequency and θ_2 is the phase angle of the load impedance plus the tube plate resistance at the radio frequency

ω_{rf} = r-f signal frequency, radians/sec

The output circuit of a triode or pentode mixer is shown in Fig. 7.49*b*. The mixer output admittance Y_o at the i-f frequency is

$$Y_o = \frac{1}{\bar{r}_p} + j\omega_{if}C_o \qquad (7.162)$$

where $\bar{r}_p$ = plate resistance of tube averaged over local-oscillator cycle

ω_{if} = intermediate frequency, radians/sec

C_o = total output capacity

4. Conversion Gain. The available conversion power gain G_c of a mixer is defined as the ratio of the available i-f signal power S_{if} at the mixer output to the available r-f signal power S_{rf} at the output of the r-f signal source. The available conversion power gain of a triode or pentode mixer is given by

$$G_c = \frac{S_{if}}{S_{rf}} = \frac{g_c^2 G_s \bar{r}_p}{(G_s + G_1)^2} \qquad (7.163)$$

where G_s = output conductance of r-f signal source

G_1 = sum of mixer tube input conductances given in Eq. (7.161)

$\bar{r}_p$ = mixer plate resistance averaged over local-oscillator cycle

g_c = conversion transconductance of mixer tube

S_{rf} = r-f signal power which r-f source could deliver to a conjugate load

S_{if} = i-f signal power which mixer could deliver to a conjugate load under matching conditions existing between r-f source and mixer input

5. *Noise Figure.* The equations for the noise figures of triode and pentode grounded cathode amplifiers [see Eqs. (7.125) through (7.127)] can be used to determine the noise figures of grounded cathode mixers provided that the impedance of the grid resonant circuit is negligible at the image frequency and provided the value of input resistance R_t due to electron transit time is averaged over the local-oscillator cycle. The value of the equivalent grid-noise resistance R_{eq} appearing in these equations must be determined by Eqs. (7.30) and (7.31). When the bandwidth of the mixer input circuit is wide enough to pass the image as well as the desired radio frequency, the noise figure of the grounded cathode mixer is given by

$$F = 2 + \frac{2R_s}{R_1} + \frac{2R_sT_c'}{T\bar{R}_t} + \frac{R_{eq}}{R_s}\left(1 + \frac{R_s}{\bar{R}_t} + \frac{R_s}{R_1}\right)^2 \qquad (7.164)$$

where R_s = source resistance, ohms

$\bar{R}_t$ = grid-cathode transit-time loading resistance averaged over local-oscillator cycle, ohms

R_{eq} = equivalent plate shot-noise grid resistance, ohms [see Eq. (7.30) or (7.31)]

R_1 = any external resistance shunting input, e.g., shunt resistance of grid resonant circuit, ohms

The increase in noise figure is due to the r-f noise within the tube and in R_s at the image frequency which is converted to the intermediate frequency by the heterodyne action.

The optimum source resistances for the grounded cathode mixer in the absence and presence of the image frequency noise at the mixer input are given by Eqs. (7.165) and (7.166), respectively.

$$R_s(\text{optimum}) = \frac{\bar{R}_a}{\sqrt{1 + \dfrac{\bar{R}_a{}^2}{R_1R_{eq}}\left(1 + \dfrac{R_1T_c'}{\bar{R}_tT}\right)}} \qquad \text{(image frequency noise absent)}$$

$$(7.165)$$

$$R_s(\text{optimum}) = \frac{\bar{R}_a}{\sqrt{1 + \dfrac{2\bar{R}_a{}^2}{R_1R_{eq}}\left(1 + \dfrac{R_1T_c'}{\bar{R}_tT}\right)}} \qquad \text{(image frequency noise present)}$$

$$(7.166)$$

where $\bar{R}_a = \dfrac{R_1\bar{R}_t}{R_1 + \bar{R}_t}$

The noise figure of the grounded grid mixer when image frequency noise is not present in the mixer output is given by

$$F = 1 + \frac{T_c'R_s}{T\bar{R}_t} + \frac{R_{eq}}{R_s}\left(\bar{g}_mR_s + 1 + \frac{R_s}{\bar{R}_t} + \frac{R_s}{\bar{r}_p}\right)^2 + g_c{}^2R_{eq}R_s \qquad (7.167)$$

where R_{eq} = equivalent plate shot-noise grid resistance as defined by Eq. (7.30)

$\bar{r}_p$ = plate resistance averaged over local-oscillator cycle

The optimum source resistance for this condition is given by

$$R_s(\text{optimum}) = \frac{1}{\sqrt{\dfrac{T'_c}{T\bar{R}_t R_{eq}} + g_c{}^2 + \left(\bar{g}_m + \dfrac{1}{\bar{R}_t} + \dfrac{1}{\bar{r}_p}\right)^2}}$$

(image frequency noise absent)   (7.168)

When the input circuit bandwidth is wide enough to pass the image as well as the desired radio frequency, the noise figure of the ground grid mixer is given by

$$F = 2 + 2\frac{T'_c R_s}{T\bar{R}_t} + \frac{R_{eq}}{R_s}\left(\bar{g}_m R_s + 1 + \frac{R_s}{\bar{R}_t} + \frac{R_s}{\bar{r}_p}\right)^2 + 2g_c{}^2 R_{eq} R_s$$

(image frequency noise present)   (7.169)

The optimum source resistance for this condition is given by

$$R_s(\text{optimum}) = \frac{1}{\sqrt{\dfrac{2T'_c}{T\bar{R}_t R_{eq}} + 2g_c{}^2 + \left(\bar{g}_m + \dfrac{1}{\bar{R}_t} + \dfrac{1}{\bar{r}_p}\right)^2}}$$

(image frequency noise present)   (7.170)

The noise figure of a triode or pentode mixer is always greater than the noise figure of the same tube operated as an amplifier.

6. Mixer Bias.   The bias for a triode or pentode mixer can be obtained from either an external bias voltage, cathode bias, or grid-leak bias.   If fixed grid bias is used, the conversion transconductance of the mixer tube is very sensitive to changes in the amplitude of the local-oscillator voltage.   The oscillator voltage may change with variations in the oscillator supply voltage, with tuning of the local oscillator, and if the local oscillator is injected into the grid circuit, with tuning of the r-f signal source. The conversion transconductance is less sensitive to the local-oscillator voltage amplitude if cathode bias is used and least sensitive if grid-leak bias is used.   This is due to the fact that the bias for maximum conversion transconductance, with a small applied local-oscillator voltage, is that bias where the g_m versus E_g characteristic has the steepest slope.   With cathode bias, the bias decreases as the local-oscillator signal is reduced.   If the local-oscillator signal is reduced to zero, the bias will be equal to the value determined by the quiescent operating conditions for the tube.   With grid-leak bias, the bias decreases with decreasing local-oscillator voltage until the bias is zero at zero applied local-oscillator signal.   If only grid-leak bias is used, there is the possibility that the plate current may be excessive if the local oscillator fails.   Therefore, when grid-leak bias is used, a small bypassed cathode resistor should be included to protect the tube in case of local oscillator failure.   If a pentode mixer is used, a series screen resistor may be used instead of the cathode resistor.   The series screen resistor should have a value such that with zero bias the plate and screen currents are not excessive.   If a series screen resistor is used in conjunction with grid-leak bias, the conversion transconductance exhibits less variations with changes in the magnitude of the applied local-oscillator voltage than in any of the other bias arrangements.[1]

[1] E. W. Herold, The Operation of Frequency Converters and Mixers for Superheterodyne Receivers, *Proc. IRE*, February, 1942.

7.5b. *Multigrid Mixer and Converter Tubes.* To minimize the difficulties encountered in triode and pentode mixers due to interaction between the local oscillator and the r-f signal source, several different types of special multigrid tubes have been developed. Some of these tubes are designed to operate as oscillators as well as mixers and are known as converters. The principal types of multigrid tubes used as mixers are (1) the pentagrid mixer, (2) the pentagrid converter, (3) the triode-hexode converter.

1. Pentagrid Mixer. The pentagrid mixer contains five grids as shown in Fig. 7.50. The first and third grids are control grids, the second and fourth are screen grids, and

(*a*) PENTAGRID MIXER WITH OUTER GRID INJECTION

(*b*) PENTAGRID MIXER WITH INNER GRID INJECTION
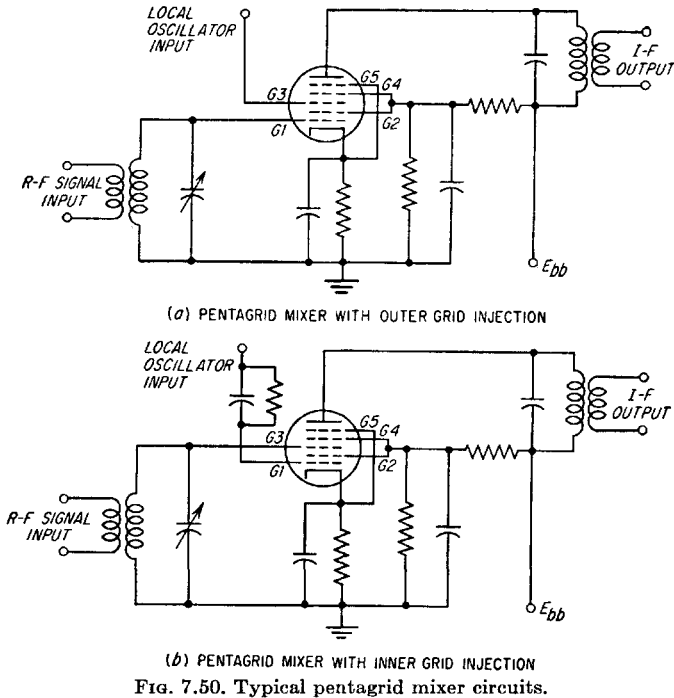
FIG. 7.50. Typical pentagrid mixer circuits.

the fifth is a suppressor grid. The local-oscillator signal can be applied to either $G1$ or $G3$. If it is applied to $G1$, the grid nearest the cathode, the tube is said to have inner grid injection. If it is applied to $G3$, the tube is said to have outer grid injection. The characteristics of the tube differ considerably for the two types of operation.

INNER GRID INJECTION. If the local-oscillator voltage is applied to $G1$, the cathode current flowing to the other tube elements is modulated at the local-oscillator frequency. Because the outer grid $G3$ to plate transconductance is a function of the tube current passing through this grid structure, the signal grid $G3$ to plate transconductance is varied at the local-oscillator frequency and the mixing action is analogous to that described for the triode mixer. The screen grid $G2$ minimizes the coupling between $G1$ and $G3$ and provides isolation between the signal and oscillator circuits. The grids $G4$ and $G5$ serve the same purposes as the screen and the suppressor grids in a pentode. Typical grid-plate transconductance versus bias characteristics for the pentagrid mixer are shown in Fig. 7.51 and illustrate how the bias voltage on either control grid effects the grid-to-plate transconductance of the other grid. Since the signal and oscillator grids are isolated, the oscillator grid can be driven into the

conducting region without loading the signal source. In practice, grid-leak bias for the oscillator grid $G1$ is used so as to ensure the maximum signal grid-to-plate transconductance.

Although grids $G2$, $G4$, and $G5$ provide sufficient shielding to eliminate any serious effects due to capacitive coupling between $G1$ and $G3$ and the plate, an internal coupling known as *space-charge coupling* exists between $G1$ and $G3$. The signal grid $G3$ is biased negatively, and the electrons passed by the oscillator injection grid $G1$ during the conduction portion of the oscillator cycle form a virtual cathode in the region between $G2$ and $G3$. The negative space charge forming this virtual cathode is proportional to the current reaching $G3$ and, hence, is proportional to the local-oscillator voltage during the interval in which the tube is conducting. When the tube current is cut off, the space charge is reduced to zero. When the negative space charge is increasing,

(a) MAXIMUM OUTER GRID-TO-PLATE
TRANSCONDUCTANCE FOR $E_{C3}=0$

(b) MAXIMUM INNER GRID-TO-PLATE
TRANSCONDUCTANCE FOR $E_{C1}=0$
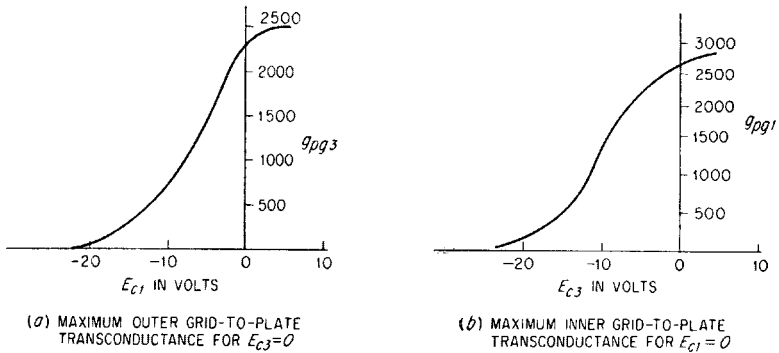
FIG. 7.51. Transconductance characteristics of typical pentagrid mixer tubes.

the space charge is repelled from the negative signal grid and induces an electron flow from the signal grid to the cathode through the external grid impedance. When the negative space charge decreases, as it does as soon as the applied local-oscillator voltage has passed its maximum positive value, the space charge moves closer to the signal grid and the induced electron flow in the signal-grid circuit is in the opposite direction. A current which varies at the local-oscillator frequency and which lags the local-oscillator voltage by approximately 90° is thus induced in the signal-grid circuit. Because of the finite transit time of the electrons in traveling from the cathode to the region of the negative space charge between $G2$ and $G3$, the induced current in the signal-grid circuit impedance lags the applied local-oscillator voltage by slightly more than 90°. This effect is equivalent to having a negative capacitance and a negative resistance in parallel between $G1$ and $G3$. The magnitude of the reactance decreases directly with frequency, and the negative resistance decreases as the square of frequency.

Coupling between the local oscillator and signal grids is usually not important unless an appreciable voltage at the local-oscillator frequency is developed across the signal-grid circuit impedance. This normally will only occur when the intermediate frequency is low relative to the signal frequency since under this condition the signal circuit impedance at the local-oscillator frequency is appreciable. The effect of the induced voltage depends upon its phase relative to the applied local-oscillator voltage. It may either increase or decrease the effective conversion transconductance by its modulation of the plate current. If the amplitude of the induced voltage exceeds the bias on $G3$, there will be grid conduction loading of the signal source.

The effects of space-charge coupling can be eliminated by the connection of a small capacitor and series resistor between $G1$ and $G3$ as shown in Fig. 7.52. The capacitive

reactance and positive resistance of these elements can be made to compensate for the space-charge coupling over a small band of frequencies. The normal procedure is to tune the signal grid circuit to resonance at the local-oscillator frequency and adjust the resistance and capacitance until no local-oscillator voltage appears across the signal circuit.

OUTER GRID INJECTION. When the r-f signal voltage is applied to G1 and the local-oscillator signal to G3, the voltage applied to the signal grid modulates the cathode current and the local-oscillator voltage varies the distribution of the cathode current between the screen grid and plate. Isola-tion between the r-f signal source and local oscillator is provided by screen grid G2. Typical transconductance between G1 and the plate as a function of G3 bias is shown in Fig. 7.51b. Coupling between the signal grid and the local-oscillator grid through the interelectrode capacitance existing be-tween these two grids may be important in some applications. When the r-f sig-nal is applied to G1 and the local oscillator to G3, the effect of space-charge coupling between grids is an apparent capacitive coupling. The magnitude of the space-charge coupling between the grids with outer grid local-oscillator injection is much smaller than that with inner grid injection

FIG. 7.52. Compensation of space-charge coupling effects between signal grid G3 and oscillator injection grid G1 in a penta-grid mixer.

because of the modifying effect of the screen grid between the r-f signal grid and the virtual cathode.

The input admittance of the signal grid is larger for outer grid injection of the local-oscillator signal than for inner grid injection. Some of the electrons which pass through the screen grid are stopped by the negative potential of the local-oscillator grid. These electrons are then reaccelerated toward the r-f signal grid by the screen potential, thereby inducing a current flow in the signal-grid external circuit. The admittance produced by this effect has a capacitive component which increases linearly with frequency and a conductive component which rises as the square of frequency. The input admittance thus produced is analogous to the conventional input admittance produced in triodes due to electron transit time except that the transit time of electrons in traveling from the screen to G3 and back to the screen is the predominate factor. As the bias applied to G3 approaches the value necessary for plate current cutoff, the input admittance of G1 approaches a maximum. The conductive component of input admittance due to this effect can exceed the con-ventional input conductance due to electron transit time by several times when the bias on G3 is near the cutoff value. This effect can be greatly reduced by designing the tube structure so that the electrons which are repelled by G3 travel a different return path and are collected by auxiliary electrodes instead of reentering the signal-grid space.

2. Pentagrid Converter. The pentagrid converter is very similar to the pentagrid mixer except that the tube also serves as the local oscillator as illustrated in Fig. 7.53. The cathode together with G1 and G2 form a Hartley oscillator which is electron-coupled to the other tube electrodes and the r-f signal is applied to G3. The grids G4 and G5 serve as the screen and suppressor grids, respectively. In all other respects the pentagrid converter is identical to the pentagrid mixer with inner grid injection.

3. Triode-hexode Converter. The triode-hexode converter consists of a triode oscillator and multigrid mixer as shown in Fig. 7.54. The oscillator grid G1 controls
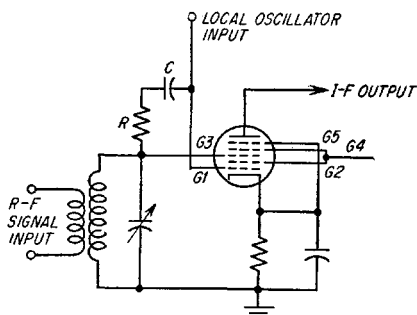
the flow of cathode current to the screen grid $G2$ of the mixer section of the tube.   The r-f signal is coupled to $G3$, and $G4$ acts as a conventional screen grid.   The principal advantage of the triode-hexode converter over the pentagrid converter is that the oscillator section is capable of operating at higher frequencies.   The same inter-electrode couplings exist in the triode-hexode converter that exist in the pentagrid mixer with inner grid injection.

4. Noise in Multigrid-mixer Tubes.   The noise present in all types of multigrid-mixer and converter tubes is much greater than that present in triode and pentode
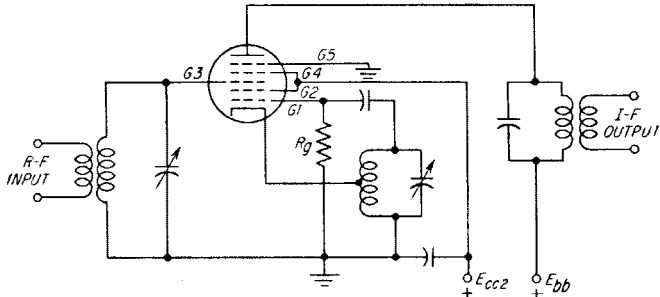
FIG. 7.53. Pentagrid converter.
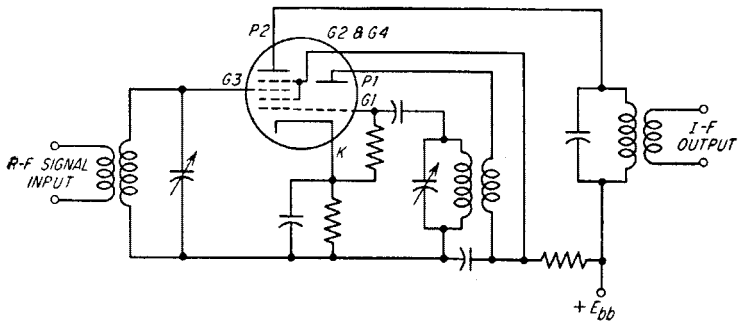
FIG. 7.54. Triode-hexode converter.

mixers.   The increase is due to an increase in partition noise which is caused by the relatively small portion of the total cathode current that flows to the plate (see Sec. 7.2*h*).   The equivalent noise resistance of a mixer of this type is given by Eq. (7.32).   The effects of grid-induced shot-effect noise must also be included in determining the noise figure of a multigrid mixer or converter.   The procedure is identical to that for the grounded-cathode triode or pentode mixer except for the increased value of R_{eq}.

7.5c. Vacuum Diode Mixers.   The diode mixer is the simplest type of frequency conversion device.   Although a diode mixer always introduces a conversion loss rather than a gain as do grid-controlled mixer tubes, the diode is the principal mixer type used at 500 Mc and above because of its superior noise figure at these frequencies. The circuit of a diode mixer is shown in Fig. 7.55.

1. Diode Conversion Conductance, Input and Output Impedances, and Conversion Gain.   If the r-f signal voltage is assumed to be much smaller than the local-oscillator voltage, the conductance of the diode can be assumed to vary as a function of the local-oscillator voltage only.   This action for a linear diode is illustrated in Fig. 7.56.

Since the diode is a two-terminal device, an i-f voltage across the diode output will combine with the local-oscillator signal to produce a current at the r-f signal frequency, and the i-f voltage will also combine with the r-f signal voltage to produce a current at the local-oscillator frequency. Because the i-f power developed in the diode output load is supplied by the r-f signal source, the r-f input impedance of the diode is a function of the i-f load impedance and the output impedance of the diode is affected by the r-f signal-source impedance. The diode con-

FIG. 7.55. Circuit of diode mixer.

ductance $g(t)$ can be expressed as a Fourier series whose fundamental frequency is the local-oscillator frequency ω_{lo}.

$$g(t) = a_o + \sum_{n=1}^{\infty} a_n \cos (n\omega_{lo}t) \tag{7.171}$$

The coefficients a_o and a_n of the Fourier series are given by

$$a_o = \frac{1}{2\pi} \int_0^{2\pi} g(t)\, d\,(\omega_{lo}t) \tag{7.172}$$

$$a_n = \frac{1}{\pi} \int_0^{2\pi} g(t) \cos (n\omega_{lo}t)\, d\,(\omega_{lo}t) \tag{7.173}$$

Since both the r-f and i-f voltages are present when the r-f signal is applied, the current
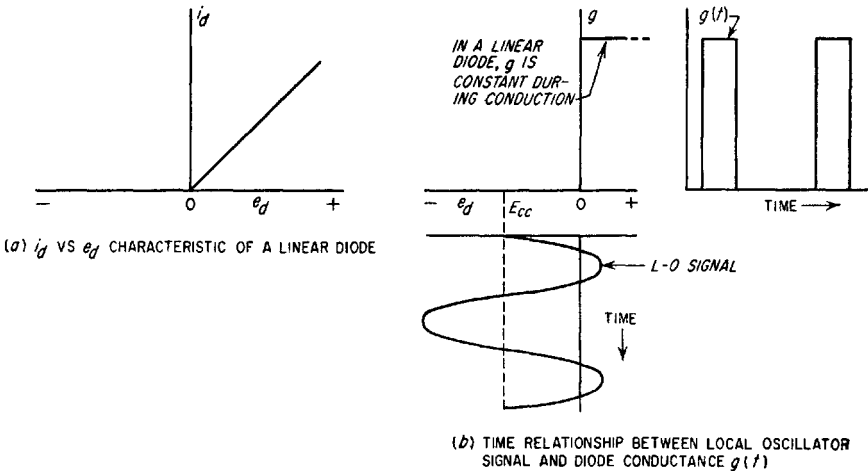
FIG. 7.56. Linear diode characteristics.

present in the diode due to an applied r-f signal is given by Eq. (7.174) provided that the diode load is a pure resistance at the intermediate frequency.

$$i = g(t)[E_{rf} \sin (\omega_{rf}t) - E_{if} \sin (\omega_{if}t)] \tag{7.174}$$

Substitution of the diode conductance as given by Eq. (7.171) yields

$$i = a_o[E_{rf} \sin (\omega_{rf}t) - E_{if} \sin (\omega_{if}t)]$$

$$+ E_{rf} \sin (\omega_{rf}t) \sum_{n=1}^{\infty} a_n \cos (n\omega_{lo}t)$$

$$- E_{if} \sin (\omega_{if}t) \sum_{n=1}^{\infty} a_n \cos (n\omega_{lo}t) \qquad (7.175)$$

Expansion of Eq. (7.175) gives

$$i = a_o E_{rf} \sin (\omega_{rf}t) - a_o E_{if} \sin (\omega_{if}t)$$

$$+ \frac{E_{rf}}{2} \sum_{n=1}^{\infty} a_n \sin (\omega_{rf} + n\omega_{lo})t$$

$$+ \frac{E_{rf}}{2} \sum_{n=1}^{\infty} a_n \sin (\omega_{rf} - n\omega_{lo})t$$

$$- \frac{E_{if}}{2} \sum_{n=1}^{\infty} a_n \sin (\omega_{if} + n\omega_{lo})t$$

$$- \frac{E_{if}}{2} \sum_{n=1}^{\infty} a_n \sin (\omega_{if} - n\omega_{lo})t \qquad (7.176)$$

Assuming that the intermediate frequency is given by

$$\omega_{if} = \omega_{rf} - n\omega_{lo}$$

the current components of importance in Eq. (7.176) are those at the frequencies $\omega_{rf} - n\omega_{lo}$, ω_{rf} and ω_{lo}. All other terms can be neglected. The currents at ω_{if} and ω_{rf} are given by

$$i_{rf} = a_o E_{rf} \sin (\omega_{rf}t) - \frac{a_n}{2} E_{if} \sin (\omega_{rf}t) \qquad (7.177)$$

$$i_{if} = -a_o E_{if} \sin (\omega_{if}t) + \frac{a_n}{2} E_{rf} \sin (\omega_{if}t) \qquad (7.178)$$

The conversion conductance g_c is given by

$$g_c = \frac{\delta i_{if}}{\delta E_{rf}}$$

$$= \frac{a_n}{2} \qquad (7.179)$$

and Eqs. (7.177) and (7.178) can be written as

$$i_{rf} = g_o E_{rf} \sin (\omega_{rf}t) - g_c E_{if} \sin (\omega_{rf}t) \qquad (7.180)$$
$$i_{if} = g_c E_{rf} \sin (\omega_{if}t) - g_o E_{if} \sin (\omega_{if}t) \qquad (7.181)$$

where g_o = average diode conductance and is equal to a_o as determined from Eq. (7.172)

The equivalent circuit[1] of a diode mixer satisfying Eqs. (7.180) and (7.181) is given in

[1] For greater detail concerning the development of this diode equivalent circuit see E. W. Herold, Frequency Mixing in Diodes, *Proc. IRE*, October, 1943.

Fig. 7.57. With the exception of the fact that the input and output frequencies are
different, the equivalent circuit of a diode
mixer is simply a symmetrical π attenua-
tor in which maximum power transfer
is achieved when the mixer is matched
on an image impedance basis. Under
this condition, both the i-f load resist-
ance R_{if} and the r-f source resistance R_s
are equal to the characteristic impedance
Z_m of the mixer network. The value of

FIG. 7.57. Equivalent circuit of diode mixer.

R_s and R_{if} which provides the maximum power transfer through the mixer is

$$R_s = R_{if} = Z_m = \frac{1}{\sqrt{g_o^2 - g_c^2}} \qquad (7.182)$$

The ratio of the mixer characteristic impedance Z_m to the average diode resistance
R_o is plotted as a function of g_c/g_o in Fig. 7.58a. As the ratio g_c/g_o approaches unity,
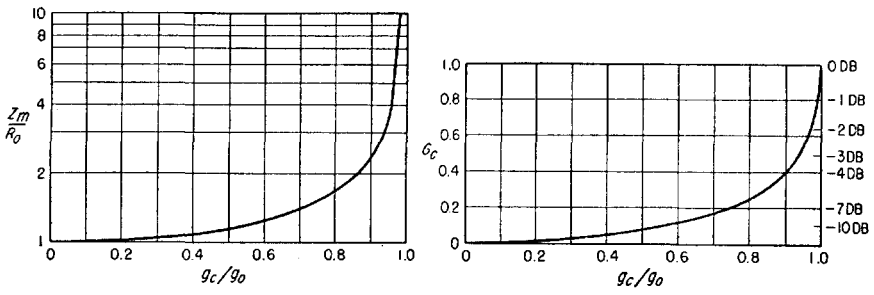
(a) RATIO OF THE MIXER CHARACTERISTIC IMPEDANCE
Z_m TO THE AVERAGE DIODE RESISTANCE $(1/g_o)$
AS A FUNCTION OF g_c/g_o

(b) CONVERSION POWER GAIN OF A DIODE MIXER AS
A FUNCTION OF THE RATIO OF CONVERSION CON-
DUCTANCE TO AVERAGE DIODE CONDUCTANCE FOR
THE CASE WHERE THE MIXER INPUT AND OUTPUT
IMPEDANCES ARE MATCHED

FIG. 7.58. Characteristic impedance and conversion power gain of a diode mixer.

the ratio Z_m/R_o approaches infinity. The value of the mixer input conductance g_i
for any value of i-f load resistance is

$$g_i = \frac{g_o^2 - g_c^2 + g_o g_{if}}{g_o + g_{if}} \qquad (7.183)$$

where $g_{if} = 1/R_{if}$
If the values of the r-f source resistance and the i-f load resistance are given by Eq.
(7.182), the conversion power gain G_c of the mixer is

$$G_c = \left[\frac{(g_c/g_o)}{1 + \sqrt{1 - (g_c/g_o)^2}} \right]^2 \qquad (7.184)$$

The conversion power gain is plotted as a function of g_c/g_o for this condition in Fig.
7.58b. Note that the gain is always less than unity.

The expressions for g_o and g_c are given by Eqs. (7.172) and (7.179), respectively.
The ratio of these two conductances can approach unity only as $g(t)$ becomes a narrow
pulse so that $\cos(\omega_{lo} t)$ is essentially unity for the duration of $g(t)$. This condition
can be met by applying a very large local-oscillator signal and a very large negative
bias so that the diode conducts only at the peaks of the local-oscillator cycle. The
local oscillator should be coupled to the r-f input circuit so as to present either a

high shunt impedance or a low series impedance in the input circuit at the r-f signal frequency.

2. Graphical Determination of g_o and g_c. The average diode conductance g_o and the conversion conductance g_c can be determined graphically by constructing a g versus e plot from the E-I characteristics of the diode and using the following approximate equations.[1]

$$g_o = \tfrac{1}{12}[g_1 + g_7 + 2(g_2 + g_3 + g_4 + g_5 + g_6)] \quad (7.185)$$

$$g_c = \tfrac{1}{12}[(g_7 - g_1) + (g_5 - g_3) + 1.73(g_6 - g_2)] \quad (7.186)$$

where g_c = conversion conductance for intermediate frequency of $\omega_{rf} - \omega_{lo}$

$g_1, g_2 \ldots g_7$ = values of diode conductance at points of local-oscillator cycle shown in Fig. 7.59

In an ideal diode having a linear E-I characteristic as shown in Fig. 7.56, the value of g_o, assuming zero bias, is $g/2$. When using Eq. (7.185) for this case, the value of g_4 must be taken as the average of the conductances at cutoff and when conducting, that is, $g_4 = g/2$. The conversion conductance as determined from Eq. (7.186) is approximately g/π. The ratio g_c/g_o is equal to $2/\pi$ and the conversion loss, as given by Eq. (7.184) and Fig. 7.58b, is 8.9 db.

FIG. 7.59. Graphical determination of average diode conductance g_o and conversion conductance g_c.

Example 7.11

One section of a 6AL5 diode is to be used as a mixer in the 100-Mc region. The i-f amplifier center frequency is 10 Mc, and the i-f bandwidth is 4 Mc. Determine the optimum local-oscillator injection voltage and diode bias and, under these conditions, find the mixer input impedance and conversion loss.

1. Determine the desired value of g_c/g_o.

From Fig. 7.58b and Eq. (7.184), it is evident that the minimum mixer conversion loss occurs when the ratio g_c/g_o approaches unity, assuming matched input and output impedances. Also, the largest values of g_c/g_o are associated with large values of Z_m as shown in Fig. 7.58a. The i-f load impedance R_{if} should match Z_m; therefore the optimum value of R_{if} increases as g_c/g_o increases. Thus, the largest i-f impedance which will provide the necessary bandwidth is desired.

2. Determine the maximum allowable value of diode i-f load impedance R_{if}.

Assume that the first i-f amplifier stage is to be a 6AH6 tube. The nominal value of c_i for the 6AH6 is 10.0 $\mu\mu$f. The diode output c_o is approximately 3.5 $\mu\mu$f. The socket and interstage wiring capacitance C_1 can be assumed to be about 5 $\mu\mu$f. The total shunt capacitance C_t at the diode output is therefore

$$C_t = c_o + C_1 + c_i$$
$$= 3.5 + 5.0 + 10.0$$
$$= 18.5 \ \mu\mu f$$

[1] E. W. Herold, The Operation of **Frequency Converters** and Mixers for Superheterodyne Receivers, *Proc. IRE*, February, 1942.

If it is assumed that the mixer load impedance determines the i-f amplifier bandwidth, the load resistance R_{if} is found from the relation

$$\Delta f = \frac{1}{2\pi R_{if} C_t}$$

and

$$R_{if} = \frac{1}{6.28 \times 4 \times 10^6 \times 18.5 \times 10^{-12}}$$
$$= 2{,}150 \text{ ohms}$$

3. Determine g_o, g_c, and R_{if} for the mixer.

From the E-I characteristic given in Fig. 7.60a, the diode conductance versus applied voltage characteristic is constructed (see Fig. 7.60b).

(a) I_d VS E_d CHARACTERISTIC

(b) g VS E_d CHARACTERISTIC

(c) DETERMINATION OF DIODE CONDUCTANCE THROUGHOUT LOCAL OSCILLATOR CYCLE
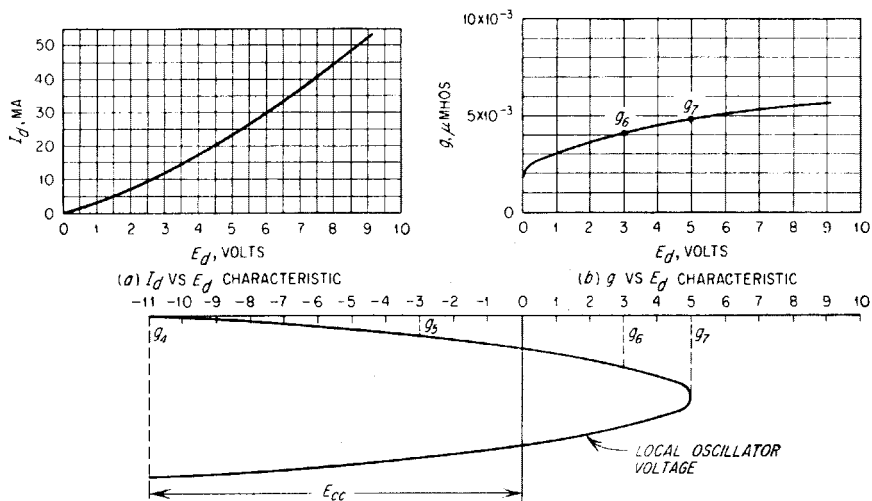
Fig. 7.60. Determination of diode conductance for Example 7.11.

From Eq. (7.186) it is seen that for maximum g_c, the conductance at points 5, 6, and 7 should be a maximum and the conductance at points 1, 2, and 3 should be a minimum. Also, to make the ratio of g_c/g_o approach unity, the conductance at point 4 should be a minimum [see Eq. (7.185)]. The relationship between the optimum value of R_{if} and the conductances g_o and g_c is given by Eq. (7.182). To make the optimum value of R_{if} be equal to 2,150 ohms as determined in step 2, the desired amplitudes of local-oscillator voltage and diode bias must be determined by successive approximations. For various assumed values of local-oscillator voltage and bias, the values of g_o and g_c must be determined graphically and the associated optimum values of R_{if} determined from Eq. (7.182). In this example, the optimum value of R_{if} is equal to 2,330 ohms if the peak value of local-oscillator voltage is equal to 16 volts and the diode bias is equal to 11 volts (see Fig. 7.60b).

$$g_1 = 0, \; g_2 = 0, \; g_3 = 0, \; g_4 = 0, \; g_5 = 0, \; g_6 = 4.0 \times 10^{-3}, \; g_7 = 4.9 \times 10^{-3}$$

From Eq. (7.185)

$$g_o = \frac{4.9 \times 10^{-3} + 2 \times 4.0 \times 10^{-3}}{12} = 1.075 \times 10^{-3}$$

From Eq. (7.186)

$$g_c = \frac{4.9 \times 10^{-3} + 1.73 \times 4.0 \times 10^{-3}}{12} = 0.985 \times 10^{-3}$$

From Eq. (7.182)

$$R_{if} = \frac{1}{\sqrt{(1.075 \times 10^{-3})^2 - (0.985 \times 10^{-3})^2}} = 2,330 \text{ ohms}$$

This is within a few per cent of the desired value of 2,150 ohms.
4. Determine the conversion power loss of the mixer.

$$\frac{g_c}{g_o} = \frac{0.985 \times 10^{-3}}{1.075 \times 10^{-3}} = 0.916$$

From Fig. 7.58b, the conversion power gain is 0.43, or a loss of 3.7 db.

3. Noise Figure of a Vacuum Diode Mixer. The noise current in a diode whose current is space-charge-limited is given by Eq. (7.14). The diode output conductance changes during the local-oscillator cycle in a diode mixer and is dependent upon the r-f source impedance. The average value of the diode output conductance at the intermediate frequency for a diode mixer having matched input and output impedances is $\sqrt{g_o^2 - g_c^2}$. It is this value which should be substituted for g in Eq. (7.14) for a diode mixer. In practice, it is observed that the noise current output of a diode mixer is greater than predicted by this modification of Eq. (7.14). The increase in noise is due to the fact that in a diode mixer the anode-cathode voltage is quite low and many electrons are elastically reflected from the anode. These reflected electrons disrupt the space charge and lower its effectiveness in reducing the plate shot noise from the cathode.[1] The noise current I_n of a diode mixer is given approximately by

$$I_n = \sqrt{0.644 \frac{T_e}{T} (4KT \, \Delta f) \sqrt{g_o^2 - g_c^2}} \tag{7.187}$$

where T_e = effective cathode temperature, i.e., that temperature at which tube would produce *observed* noise output
The term

$$t = 0.644 \frac{T_e}{T}$$

can be considered as the factor by which the noise output power of the diode equivalent resistance $\dfrac{1}{\sqrt{g_o^2 - g_c^2}}$ must be multiplied to get the actual diode noise output power. The value of t is usually between 2 and 10 for diodes having oxide-coated cathodes and depends on the particular tube type and operating conditions.

The noise figure of a diode mixer in which the r-f source and i-f load impedances are properly matched is

$$F = \frac{t}{G_c} \tag{7.188a}$$

$$F = t \left[\frac{1 + \sqrt{1 - (g_c/g_o)^2}}{(g_c/g_o)} \right]^2 \tag{7.188b}$$

where G_c = conversion gain of mixer
If a diode mixer is followed by an i-f amplifier, the over-all noise figure [see Eq. (7.55)] is given by

$$F_{db} = L_d + 10 \log_{10} (F_{if} + t - 1) \tag{7.189}$$

[1] For a more complete explanation see D. O. North, Fluctuation in Space-charge-limited Currents at Moderately High Frequencies, Part II, pp. 117–124, *RCA Rev.*, vol. 5, July, 1940.

where F_{db} = over-all noise figure, db

  L_d = conversion loss of mixer, db, that is, $10 \log_{10} (1/G_c)$

  F_{if} = noise figure of i-f amplifier expressed as power ratio

7.5d. Crystal Mixers. Another type of mixer which is widely used at frequencies above approximately 200 to 500 Mc is the crystal diode mixer. The crystal diode consists of a metal "catwhisker" in contact with a semiconductor material as shown in Fig. 7.61. The contact between the whisker and the semiconductor is maintained by the spring bend in the whisker and is not adjustable. The semiconductor material may be either silicon or germanium with small quantities of added impurities. Although silicon diodes have been used exclusively as mixers at 3,000 Mc and above,
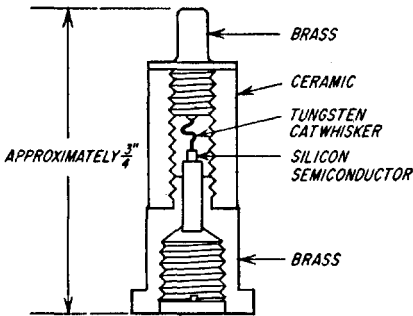
FIG. 7.61. Typical silicon mixer crystal.
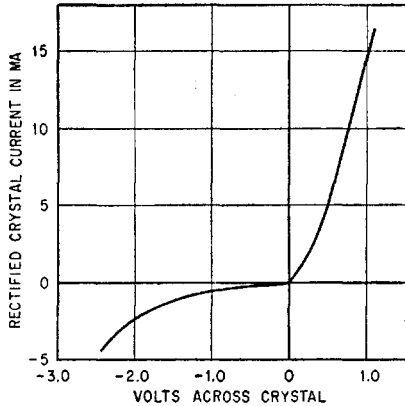
FIG. 7.62. *E-I* characteristic of typical silicon mixer crystal.

recent developments indicate that improved germanium crystals may provide equal noise figures throughout the entire microwave spectrum. Welded-contact germanium diodes are used in the 500-Mc region where their noise figures are less than those of vacuum-tube and silicon-diode mixers.

The action of a crystal diode as a mixer is exactly analogous to the operation of a vacuum-tube diode mixer. The crystal diodes conduct slightly in the reverse direction, however. A typical diode *E-I* characteristic is shown in Fig. 7.62. The optimum noise figures are achieved when the crystal diodes are operated with zero bias or with a very small negative bias. For this reason, crystal diodes are normally operated without bias. The equation for the noise figure of a crystal-diode mixer is of the same form as that of a diode mixer.

$$F = \frac{t_c}{G_c} \qquad (7.190)$$

where t_c = ratio of crystal noise power to noise power of a resistor at same temperature

  = unity at zero crystal current

The spectrum of the noise power generated in a crystal diode is not uniform with frequency as is resistor noise. Crystal noise is highest in the audio-frequency range and varies in intensity inversely with frequency from less than 50 cycles to around 1 Mc. Above 1 Mc the noise power per unit bandwidth tapers off and approaches thermal noise. The factor t_c varies from approximately 1.5 to 5 for crystal-diode mixers operating at intermediate frequencies greater than 1 Mc. Both the temperature factor t_c and the conversion gain of a crystal-diode mixer are affected by the level of the applied local-oscillator signal. Typical curves for t_c, $1/G_c$, and noise figure

are shown in Fig. 7.63 as a function of rectified crystal current, which is a measure of the applied local-oscillator power.   The optimum noise figure is usually achieved at about 0.5 ma of d-c crystal current.   This is equivalent to about 0.5 mw of local-oscillator power.   The maximum values of crystal conversion loss and temperature
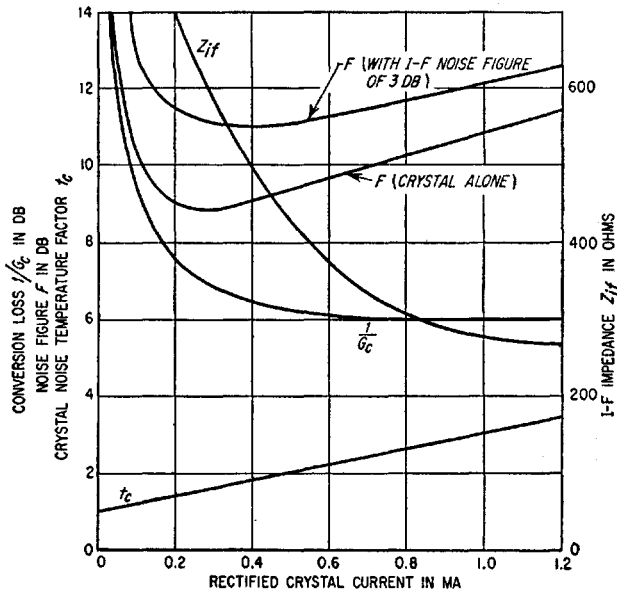
FIG. 7.63. Typical crystal mixer characteristics.

factor as specified by MIL (Military Standards) specifications are given for several common crystal-diode mixer types in Table 7.8.   The measurements are made using standard JAN (Joint Army-Navy Standards) crystal holders and 30-Mc i-f amplifiers.

TABLE 7.8. MIL RATINGS ON MIXER CRYSTAL DIODES

Crystal type	Freq., Mc	$1/G_c$ (max), db	t_c (max)	i-f impedance, ohms
1N21B	3,000	6.5	2.0	200–800
1N21C	3,000	5.5	1.5	200–800
1N23A	10,000	8.0	2.7	150–600
1N23B	10,000	6.5	2.7	150–600
1N23C	10,000	6.0	2.0	325–475
1N23D	10,000	4.5	1.7	300–400
1N25	1,000	8.0	2.5	100–400
1N26	25,000	8.5	2.5	300–600
1N28	3,000	7.0	2.0	250
1N263	9,000	6.0	1.4	150–250

NOTE: All diodes are silicon with the exception of the 1N263, which is a germanium diode.

The r-f and i-f impedances of crystal-diode mixers are dependent upon the level of the local-oscillator signal applied to the crystal.   The i-f impedance of a typical silicon crystal is shown in Fig. 7.63 as a function of the rectified crystal current.

The average value of i-f impedance at 0.4- to 0.5-ma rectified crystal current is approximately 350 to 500 ohms.

A typical microwave crystal mixer for the 3,000-Mc region is shown in Fig. 7.64a. The local-oscillator input is terminated in a 50-ohm resistive disk which is equal to the characteristic impedance of the coaxial lines in the mixer. The disk is used to provide a nearly constant load to the local oscillator as a function of frequency. The amount of local-oscillator power coupled to the crystal is varied by means of the adjustable probe. The distance from the junction of the local-oscillator input line and the adjustable probe line to the short circuit at the coupling screw is designed to be

(a) COAXIAL MIXER FOR 3,000 MC REGION

(b) WAVEGUIDE MIXER FOR 10,000 MC REGION
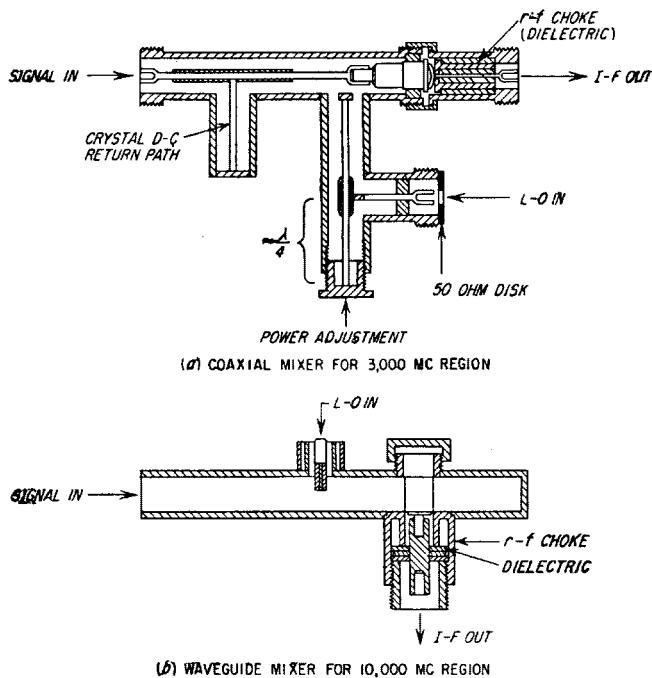
FIG. 7.64. Typical microwave crystal mixers.

approximately $\lambda/4$ at the r-f band center. A $\lambda/4$ open line or choke is used at the i-f output connector to effectively present a short circuit at the end of the crystal to the r-f and local-oscillator signals. The r-f input end of the crystal is short-cir-cuited at the intermediate frequency by the support stub which also provides a d-c return path for the crystal. In this type of mixer the local oscillator must be capable of supplying considerably more than the 0.5 mw of power required by the crystal in order that the local oscillator can be loosely coupled to the r-f signal line. If the local oscillator is tightly coupled to the r-f signal line, appreciable r-f signal power may be absorbed by the local-oscillator circuit, and the stability of the local oscillator may be adversely affected by changes in the load presented by the mixer.

A waveguide-type mixer for the 10,000-Mc region is illustrated in Fig. 7.64b. In this mixer the local oscillator is coupled into the guide with a probe. The amount of local-oscillator power injected is adjusted by varying the depth of the probe in the waveguide. The crystal is placed across the center of the waveguide with its axis parallel to the electric field. The r-f signal is kept out of the i-f output by means of an r-f choke in this line. The position of the crystal relative to the shorted end of the

waveguide is chosen so that the r-f impedance of the crystal is matched to the r-f transmission line. The probe coupling the local-oscillator signal into the waveguide provides a convenient means of coupling the output from a reflex klystron such as the 2K25, which has an output probe mounted in its base, directly into the mixer waveguide. The probe depth is kept quite small to maintain loose coupling to the waveguide to prevent appreciable loss of signal power into the local-oscillator circuit. This type of coupling presents a load admittance to the klystron which varies widely with frequency and is undesirable except in narrow frequency-range applications. In wide frequency-range receivers, it is desirable to either couple the local oscillator to the mixer by a directional coupler or to use a balanced mixer.

7.5e. Local-oscillator Noise and Balanced Mixers. The plate current shot-effect noise within the local-oscillator tube produces a noise voltage across the local-oscillator load impedance. The magnitude of this noise and its spectral distribution is

(a) BALANCED MICROWAVE MIXER

(b) EQUIVALENT CIRCUIT OF BALANCED
CIRCUIT MIXER AND I-F TRANSFORMER
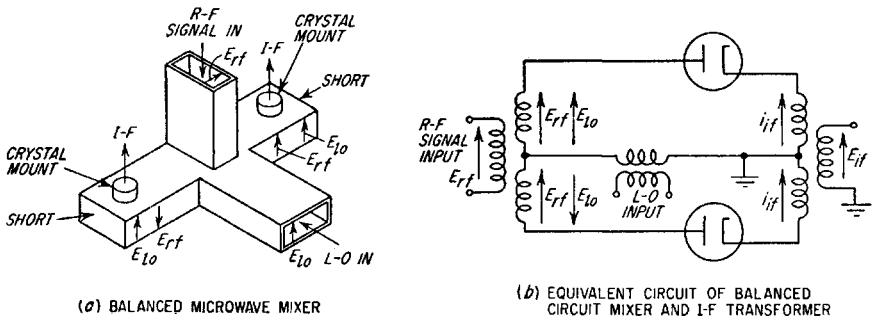
FIG. 7.65. Balanced microwave crystal mixer.

determined by the Q of the local-oscillator circuit. In addition, the plate current noise produces amplitude and frequency modulation of the local-oscillator signal itself. The magnitude of the modulation sidebands thus produced is a function of the sensitivity of the oscillator to plate-current modulation and the selectivity of the local-oscillator resonant circuit. If the noise components are present at frequencies separated from the local oscillator frequency by the intermediate frequency, they will heterodyne with the local-oscillator signal in the mixer to produce noise output at the intermediate frequency. At high microwave frequencies, the ratio of the intermediate frequency to the local-oscillator frequency may become small enough compared to the bandwidth of the local-oscillator resonant circuit that the local oscillator contributes appreciable noise to the mixer output. This additional noise is taken into account when determining the combined mixer and i-f amplifier noise figure by increasing the value of the temperature factor t_c in Eq. (7.189) by the amount necessary to account for the increased noise output. At input frequencies of approximately 10,000 Mc and intermediate frequencies of approximately 30 Mc, approximately 1.5 should be added to t_c as determined from Table 7.8 to account for the local-oscillator noise contribution.[1] This increase reduces to approximately 1.0 and 0.3 for 60- and 90-Mc intermediate frequencies, respectively.

The problems arising in microwave mixers due to local-oscillator noise and the difficulties in isolating the r-f signal source from the local oscillator may be reduced to a very great extent by the use of a balanced mixer. One type of balanced mixer and its equivalent circuit are shown in Fig. 7.65. The mixer utilizes the "magic T" type of hybrid junction, the action of which may be understood by reference to the

[1] For greater detail concerning the effects of local-oscillator noise on receiver noise figure, see R. V. Pound, "Microwave Mixers," vol. 16, chap. 5, Radiation Laboratory Series, McGraw-Hill Book Company, Inc., New York, 1948.

mixer equivalent circuit. The local-oscillator signal is applied in phase to each diode, and the r-f signal is applied in phase opposition to each diode. Since the phase of the i-f signal at the output of each diode is a function of the relative phases of the two heterodyned signals, the i-f current of one diode will be increasing while that of the other is decreasing, and the i-f voltage induced in the secondary of the i-f transformer due to the i-f currents from each diode will add in phase. However, noise-modulation components from the local oscillator will be in phase at each diode and will produce i-f output currents which will cancel each other. In a well-matched magic T, there is 20- to 40-db isolation between the local-oscillator input and the r-f signal input so that the local oscillator can be closely coupled to the mixer without introducing local-oscillator stability problems or coupling r-f signal power into the local-oscillator transmission line. The use of a balanced mixer instead of a single-ended mixer in a receiver having a 30-Mc i-f and a reflex klystron local oscillator and operating at 10,000 Mc may improve the noise figure by as much as 2 db.

7.6. Local Oscillators. The local oscillator in a superheterodyne receiver may consist of either a separate oscillator tube or tubes, or it may be a part of the mixer tube. At frequencies where it is possible to employ r-f amplifiers ahead of the mixer to minimize the receiver noise figure, or in receivers in which sensitivity is not of prime importance, converter tubes serving both as mixer and local oscillator are frequently used.

7.6a. General Considerations in Design of a Local Oscillator. There are several factors which must be given consideration in the design of a local oscillator for a particular application.

1. Oscillator Output. The local-oscillator signal which is applied to the mixer must have the correct voltage amplitude if maximum conversion transconductance is to be achieved. In triode, pentode, and multigrid mixer tubes, the local-oscillator peak voltage amplitude at the mixer should be approximately 3 to 10 volts. In vacuum diode mixers operated with a high negative bias to achieve minimum conversion loss, the applied local-oscillator signal may be in the order of 20 volts peak. In crystal diode mixers, the optimum local-oscillator signal is that which produces approximately 0.5 ma of rectified crystal current. This corresponds closely to 0.5 mw of local-oscillator power.

To minimize interaction between the local oscillator and the r-f signal source, it is desirable to loosely couple the local oscillator to the mixer. For this reason, the local-oscillator signal voltage should be five to ten times the required voltage at the mixer.

2. Frequency Stability. In most receivers, short-term oscillator stability must be such that the variations in local-oscillator frequency will not be greater than about 20 per cent of the i-f amplifier bandwidth. If the allowable warmup time of the receiver is insufficient to allow the components in the local-oscillator circuit to reach thermal equilibrium, the frequency drift due to changes in component values with temperature may be excessive unless temperature compensated parts are used. Variations in the voltages applied to the tube may also cause frequency changes in the oscillator.

3. Amplitude Stability. To prevent variations in the mixer conversion transconductance, the local-oscillator voltage amplitude must be held constant. If the tuning range of the receiver is a significant fraction of the local-oscillator frequency, the oscillator output may vary considerably with tuning because of variations in the Q of the oscillator resonant circuit. The effect of oscillator amplitude variations upon mixer conversion gain is minimized by using cathode bias or grid-leak bias in the mixer tube (see Sec. 7.5a).

4. Frequency of Operation. The frequency of the local oscillator will determine to a certain extent the tube type and circuit which is most suitable. Conventional triodes are used extensively at frequencies up to approximately 2,500 Mc. At

approximately 4,000 Mc and above, the effects of transit time in triodes are so severe that they cannot be used, and the reflex klystron is used almost exclusively to all other tube types.

7.6b. Triode and Pentode Oscillators. At broadcast and short-wave frequencies, any of the common types of oscillator circuits are satisfactory for local-oscillator use. The design data for these various types of oscillators are given in Sec. 6.2. The Hartley and the tuned-grid oscillators are the most frequently used types for fundamental operation at low frequencies. The electron-coupled oscillator is normally not used for fundamental operation because of the high harmonic content of the plate circuit signal. These harmonics may beat with undesired r-f signals to produce an i-f output from the mixer unless adequate preselection is present in the r-f amplifier stages preceding the mixer.

At frequencies above 100 Mc, the use of circuits utilizing mutual inductive coupling becomes difficult, and the Colpitts and tuned-plate tuned-grid oscillators are used extensively. Conventional triodes such as the 6AF4 may be used successfully up to 1,000 Mc in the Colpitts circuit using either lumped circuit elements or resonant lines. The "lighthouse" triodes, which have very closely spaced planar electrodes, and the "pencil" triodes, which have very closely spaced cylindrical electrodes, may be used in the grounded-grid version of the tuned-plate tuned-grid oscillator circuit up to 3,000 Mc. Both of these tube types have grid, cathode, and plate connections which form a part of the externally connected coaxial transmission-line resonant circuits (see Sec. 6.5).

7.6c. Klystron Oscillators. At frequencies above 4,000 Mc, the reflex klystron is used almost exclusively as the local oscillator in superheterodyne receivers. The principle of operation of klystron oscillators is discussed in Sec. 6.6a. An advantage of the reflex klystron is the ease with which it can be electronically tuned by variation of the repeller voltage.

7.6d. Oscillator Tracking. In superheterodynes, the local oscillator is tuned to a frequency differing from the r-f signal frequency by the intermediate frequency. Since the tuning of the local-oscillator circuit is ordinarily ganged to the tuning of the r-f circuits, the local oscillator must tune over either a greater or smaller percentage of its center frequency than the r-f circuits, depending upon whether the local oscillator is tuned below or above the signal frequency. If the fixed inductances in the r-f circuit and the local-oscillator circuit are identical, a specially shaped variable capacitor can be used in either circuit to provide exact tracking at all frequencies. If the local-oscillator signal is tuned above the r-f signal, the necessary relation between the two capacitors is

$$\frac{C_{lo}}{C_{rf}} = \frac{f_{rf}^2}{f_{lo}^2} \qquad (7.191)$$
$$= \frac{1}{\left(1 + \dfrac{f_{if}}{f_{rf}}\right)^2}$$

This relation must hold for all values of f_{rf} in the frequency band to be covered by the receiver.

In the usual case, the variable capacitors for the r-f circuit and the local-oscillator tuning are identical and are ganged together. To maintain a constant difference between the resonant frequencies of the two circuits, auxiliary capacitors must be added to one of the circuits. The local-oscillator circuit is ordinarily the one which is modified since there often are two or more r-f circuits which are tuned together. The general resonant circuits for the r-f stage and the local-oscillator stage are shown in Fig. 7.66a. By the proper combination of the oscillator circuit capacitances, it is possible to achieve exact tracking at three frequencies. This is illustrated in Fig. 7.66b. If the three frequencies at which the local-oscillator tuning is correct

are selected to be near the ends of the receiver tuning range and near the arithmetic center of the tuning range, the maximum values of deviation above and below the exact frequency will be nearly equal and of minimum value. In the 550- to 1,500-kc broadcast band, the maximum error in local-oscillator frequency need not exceed about 3 kc.

(a) GANGED OSCILLATOR AND R-F CIRCUITS HAVING IDENTICAL VARIABLE CONDENSERS

(b) TYPICAL CURVE OF LOCAL OSCILLATOR TRACKING ERROR
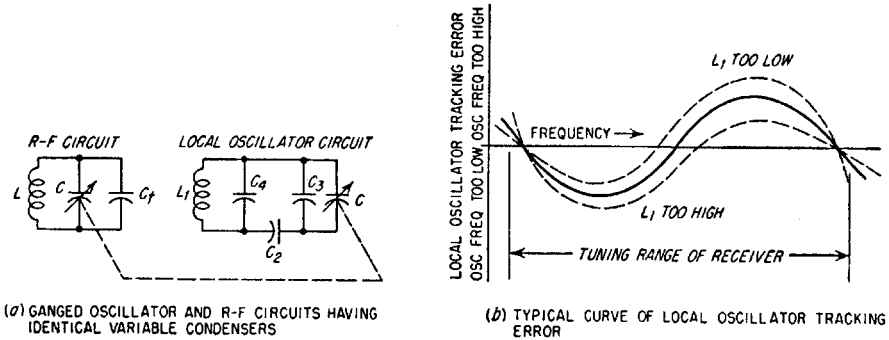
FIG. 7.66. Tracking of local-oscillator and r-f amplifier circuits.

The equations relating the circuit elements in the r-f stage and the local oscillator are as follows:[1]

C = variable tuning capacitance, $\mu\mu f$

C_3 = sum of trimmer capacitance, minimum capacitance of C, and tube input capacitance, $\mu\mu f$

C_t = sum of distributed capacitance of L, minimum capacitance of C, and tube input capacitance, $\mu\mu f$

C_2 = padder capacitance, $\mu\mu f$

C_4 = distributed capacitance of L_1, $\mu\mu f$

CASE 1. When $C_4 = 0$ or $C_4 \ll C_2$ (the usual case)

$$C_2 = C_o f_o{}^2 \left(\frac{1}{n^2} - \frac{1}{l^2} \right) \qquad (7.192)$$

$$C_3 = \frac{C_o f_o{}^2}{l^2} \qquad (7.193)$$

$$L_1 = \frac{l^2 L (C_2 + C_3)}{C_2 m^2} \qquad (7.194)$$

CASE 2. When $C_3 = 0$

$$C_2 = \frac{C_o f_o{}^2}{n^2} \qquad (7.195)$$

$$C_4 = \frac{C_o f_o{}^2}{l^2 - n^2} \qquad (7.196)$$

$$L_1 = \frac{l^2 L C_2}{m^2 (C_2 + C_4)} \qquad (7.197)$$

CASE 3. When C_4 is known

$$C_2 = A \left(\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{C_4}{A}} \right) \qquad (7.198)$$

$$C_3 = \frac{C_o f_o{}^2}{l^2} - \frac{C_2 C_4}{C_2 + C_4} \qquad (7.199)$$

$$L_1 = \frac{l^2 L (C_2 + C_3)}{m^2 (C_2 + C_4)} \qquad (7.200)$$

[1] This information follows that contained in "The Radiatron Designers' Handbook," Amalgamated Wireless Valve Company, Sidney, Australia, 1941.

CASE 4.   When C_3 is known

$$C_2 = \frac{C_o f_o^2}{n^2} - C_3 \qquad (7.201)$$

$$C_4 = \frac{C_2 B}{C_2 - B} \qquad (7.202)$$

$$L_1 = \frac{l^2 L (C_2 + C_3)}{m^2 (C_2 + C_4)} \qquad (7.203)$$

Check Formula

Equation for oscillator frequency resulting from circuit constants

$$f_{lo} = m \sqrt{\frac{f_{rf}^2 + n^2}{f_{rf}^2 + l^2}} \qquad (7.204)$$

Equations for l^2, m^2, and n^2 in terms of oscillator constants:

$$l^2 = \frac{C_o f_o^2 (C_2 + C_4)}{C_2 C_4 + C_3 (C_2 + C_4)} \qquad (7.205)$$

$$m^2 = \frac{L C_o f_o^2}{L_1} \left(C_4 + \frac{C_2 C_3}{C_2 + C_3} \right) \qquad (7.206)$$

$$n^2 = \frac{C_o f_o^2}{C_2 + C_3} \qquad (7.207)$$

where f_{if} = intermediate frequency, Mc

f_1, f_2, f_3 = frequencies in megacycles at which exact tracking is to be obtained

$a = f_1 + f_2 + f_3$

$b^2 = f_1 f_2 + f_1 f_3 + f_2 f_3$

$c^3 = f_1 f_2 f_3$

$d = a + 2f_{if}$

$l^2 = \dfrac{b^2 d - c^3}{2 f_{if}}$

$m^2 = l^2 + f_{if}^2 + ad - b^2$

$n^2 = \dfrac{c^3 d + f_{if}^2 l^2}{m^2}$

C_o = maximum tuning capacitance in r-f stage, $\mu\mu f$

f_o = lowest frequency to which r-f stage is to be tuned, Mc

$L = \dfrac{25,330}{C_o f_o^2} \qquad \mu h$

$$A = C_o f_o^2 \left(\frac{1}{n^2} - \frac{1}{l^2} \right)$$

$$B = \frac{C_o f_o^2}{l^2} - C_3$$

Case 1 is of most practical importance since it is usually desirable to place the trimmer capacitor in shunt with the tuning capacitor rather than in shunt with the coil to minimize the effects of variations in tube input capacitance and minimum tuning capacitance. If this is done, C_4 is just the distributed capacitance of the coil and is usually so small compared to C_2 that it can be considered as a part of C_3 with negligible error.

After the proper circuit-element values have been determined, the tracking should be adjusted by experimentally adjusting the trimmer and padder capacitors. The shunt trimmer capacitor has a predominant effect upon the high-frequency crossover frequency, and the series padder has a predominant effect upon the low-frequency crossover frequency. The middle crossover frequency is controlled by adjustment of the inductance in the local-oscillator circuit.

Example 7.12

Determine the proper circuit elements for a local oscillator in a receiver which is to tune from 500 to 1,500 kc and has an intermediate frequency of 465 kc. The variable capacitors in the local-oscillator circuit and the r-f circuit are identical sections of a two-gang unit. The maximum capacitance of each section including tube capacitance is 365 $\mu\mu$f, and the minimum capacitance of each section is 20 $\mu\mu$f including tube capacitance.

Solution

1. Determine the required maximum and minimum values of capacitance in the r-f circuit.

Since $\omega^2 = \dfrac{1}{LC}$

$$L = \frac{1}{\omega_a{}^2 C_b} = \frac{1}{\omega_b{}^2 C_a}$$

where ω_a = minimum operating frequency, radians/sec
ω_b = maximum operating frequency, radians/sec
C_a = minimum r-f tuning capacitance, $\mu\mu$f
C_b = maximum r-f tuning capacitance, $\mu\mu$f
Therefore

$$\frac{1}{(2\pi \times 500 \times 10^3)^2 (365 + \Delta C) \times 10^{-12}} = \frac{1}{(2\pi \times 1,500 \times 10^3)^2 \times (20 + \Delta C) \times 10^{-12}}$$

The value of ΔC required to satisfy the above relationship is 23.1 $\mu\mu$f. Therefore, 23.1 $\mu\mu$f must be added in parallel with the tuning capacitor in the r-f amplifier to cover the band of 500 to 1,500 kc with the tuning range of the variable capacitor.

2. Determine the proper inductance for the r-f section.

$$L = \frac{1}{\omega_a{}^2 C_b}$$
$$= \frac{1}{(2\pi \times 500 \times 10^3)^2 \times 388.1 \times 10^{-12}}$$
$$= 261 \ \mu\text{h}$$

3. Determine the three frequencies at which exact tracking is desired.
As indicated in Fig. 7.66b the two end frequencies of exact tracking should be slightly inside the maximum and minimum desired tuning frequencies.

$$\text{Let } f_1 = 550 \text{ kc} = 0.55 \text{ Mc}$$
$$f_2 = 950 \text{ kc} = 0.95 \text{ Mc}$$
$$f_3 = 1,450 \text{ kc} = 1.45 \text{ Mc}$$

4. Determine the other constants necessary for the solution of Case 1.

(1) $a = f_1 + f_2 + f_3 = 0.550 + 0.950 + 1.450 = 2.950$
(2) $b^2 = f_1 f_2 + f_1 f_3 + f_2 f_3 = 2.697$
(3) $c^3 = f_1 f_2 f_3 = 0.7576$
(4) $d = a + 2f_{if} = 3.88$
(5) $l^2 = \dfrac{b^2 d - c^3}{2f_{if}} = 10.44$
(6) $m^2 = l^2 + f_{if}{}^2 + ad - b^2 = 19.40$
(7) $n^2 = \dfrac{c^3 d + f_{if}{}^2 l^2}{m^2} = 0.2678$

5. Determine C_2, C_3, and L_1.

$$C_2 = C_o f_o{}^2 \left(\frac{1}{n^2} - \frac{1}{l^2} \right)$$
$$= 353.0 \ \mu\mu\text{f}$$
$$C_3 = \frac{C_o f_o{}^2}{l^2}$$

$$C_3 = 9.29 \ \mu\mu f$$

$$L_1 = \frac{l^2 L(C_2 + C_3)}{m^2 C_2}$$

$$= 144 \ \mu h$$

6. Apply the check formula to check the computation

$$f_{lo} = m \sqrt{\frac{f_{rf}{}^2 + n^2}{f_{rf}{}^2 + l^2}}$$

Since exact tracking is to occur at $f_{rf} = 0.950$ Mc, the value of f_{lo} as determined from the check formula should be $f_{rf} + f_{if}$, or 1.415 Mc. Substitution of the values for f_{rf}, n^2, l^2, and m gives a value of f_{lo} equal to 1.415 Mc, which provides the desired check.
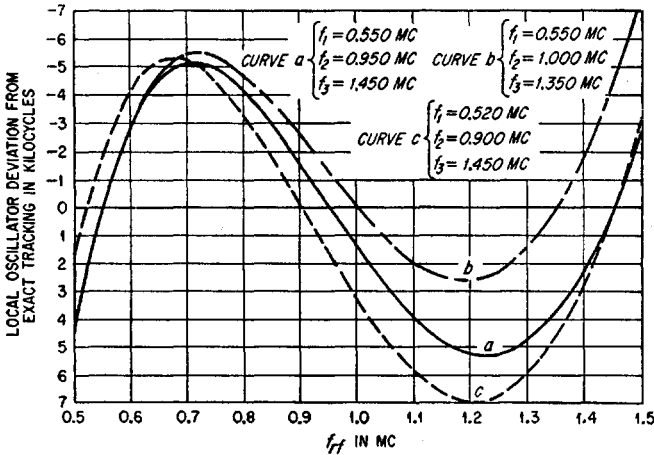
FIG. 7.67. Local-oscillator tracking error of Example 7.12.

7. The tracking error Δf existing at any frequency within the band is given by

$$\Delta f = m \sqrt{\frac{f_{rf}{}^2 + n^2}{f_{rf}{}^2 + l^2}} - (f_{rf} + f_{if})$$

The tracking error as a function of f_{rf} is plotted as curve a in Fig. 7.67. Curves b and c in Fig. 7.67 illustrate the effects of various values of f_1, f_2, and f_3.

7.6e. Automatic Frequency Control. It is often desirable to control automatically the frequency to which the receiver is tuned by controlling the local-oscillator frequency. Such control systems are known as automatic frequency control (AFC) systems. They are two general classes: difference-frequency AFC and absolute-frequency AFC. In the difference-frequency AFC system, the frequency of the local oscillator is automatically maintained at a fixed increment from some reference signal. Normally, the reference signal is the signal it is desired to receive, and the fixed increment is the intermediate frequency of the receiver. Difference-frequency AFC systems are used in receivers where the frequency stability of either the desired signal or the local-oscillator signal is inadequate to maintain the desired difference in frequency between the local oscillator and the signal to be received.

In absolute-frequency AFC systems, the frequency of the local oscillator is maintained constant at a predetermined frequency. Such systems find application in receivers fixed-tuned to a specified frequency and in receivers requiring a very stable local-oscillator frequency as is the case in moving target indicator radar systems.

1. *Difference-frequency AFC Systems.* A block diagram for a difference-frequency
AFC system is shown in Fig. 7.68. The local-oscillator signal is heterodyned with
the reference signal, and the difference signal is fed to a discriminator. The output
of the discriminator is a d-c voltage which is proportional to the difference between
the desired difference frequency and the actual difference frequency. The d-c voltage
from the discriminator is used as a con-
trol voltage to vary the frequency of the
local oscillator, thereby compensating un-
desired frequency drifts. If the local
oscillator is a reflex klystron, the d-c
control voltage is applied to the repeller
electrode. If the local oscillator is a con-
ventional tube, the oscillator frequency
can be varied by applying the d-c control
voltage to the grid of a reactance tube
connected across the oscillator resonant

FIG. 7.68. Difference-frequency AFC system.

circuit. This type of system constitutes a closed-loop feedback system in which
frequency deviations produce d-c control signals having the correct polarity to cause a
correction of the oscillator frequency. The degree of frequency stabilization achieved
is a function of the loop gain and is given by Eq. (7.208).

$$\frac{\Delta f'}{\Delta f} = \frac{1}{1 + k_1 k_2 A} \qquad (7.208)$$

where $\Delta f'$ = frequency change with AFC
Δf = frequency change without AFC
k_1 = transfer constant of discriminator, volts/Mc
k_2 = transfer constant of klystron repeller or reactance tube plus oscillator,
 Mc/volts
A = gain in feedback loop, i.e., circuit gain between discriminator output
 and local-oscillator input

Discriminators. The general requirement of the discriminator shown in Fig. 7.68
is to provide a d-c voltage output which is zero when its input signal is equal to the
desired frequency and which has an amplitude proportional to the difference between
the correct frequency and the actual frequency when the input signal frequency is
incorrect. The polarity of the voltage output must reverse in sign as the frequency
of the input signal is increased from below to above that of the correct frequency.
This general characteristic is shown in Fig. 7.69a.

Numerous circuit configurations will provide the characteristic shown in Fig. 7.69a.
One of the simplest is shown in Fig. 7.69b. The discriminator consists of two con-
ventional diode detectors connected so that the output signal is the difference between
the outputs of the two detectors. One diode is connected in reverse with respect to
the other so that the polarity of one detector is negative and the other is positive.
The resonant circuits are tuned to different frequencies, one above and one below
the desired crossover frequency. The Q of the resonant circuits and the separation
between the two resonant frequencies determine the width of the discriminator
response curve. Although this type of discriminator is quite simple, it requires that
the two input circuits be isolated from each other.

One of the most commonly used discriminators is the Foster-Seeley circuit shown
in Fig. 7.70. Primary and secondary circuits are resonant at the same frequency and
are inductively coupled with less than critical coupling. Primary voltage and the
secondary induced voltage are in phase, and at resonance, the voltage across C_2 lags
by 90° the voltage which is induced in the secondary L_2. Thus, the voltage E_2 lags

(a) TYPICAL DISCRIMINATOR CHARACTERISTIC

WITH REFERENCE TO FIG. 7.68, $f_o = f_{if}$

(b) SIMPLE DISCRIMINATOR CIRCUIT

(c) RESPONSE OF SIMPLE DISCRIMINATOR CIRCUIT
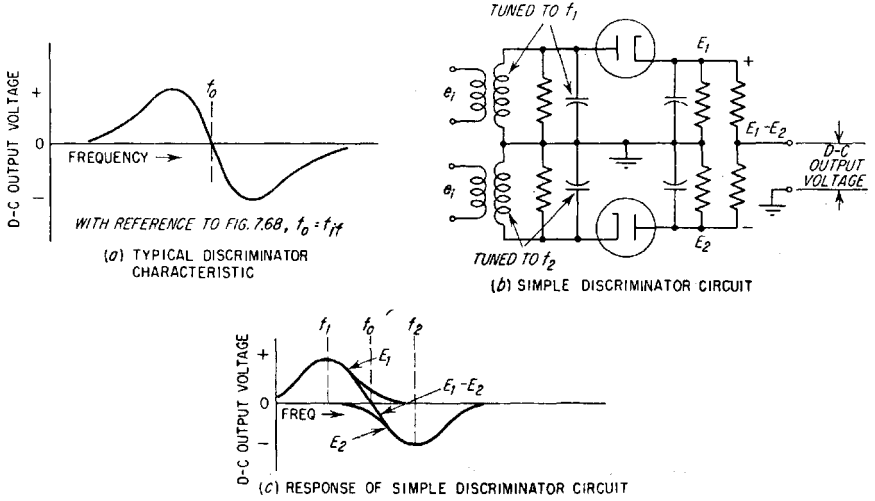
FIG. 7.69. Discriminator characteristics.
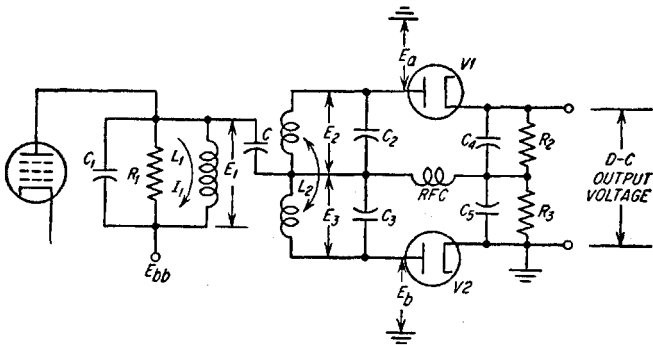
(a) SCHEMATIC OF FOSTER–SEELEY DISCRIMINATOR

(b) AT RESONANCE   (c) ABOVE RESONANCE   (d) BELOW RESONANCE

VOLTAGE RELATIONSHIPS

(e)
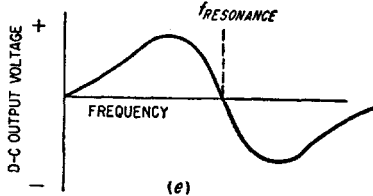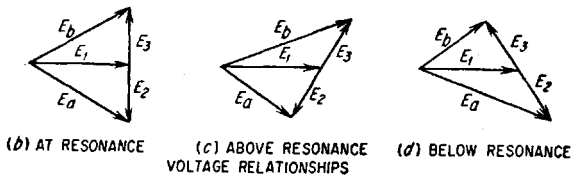
FIG. 7.70. Foster-Seeley discriminator circuit.

E_1 by 90°, and the voltage E_3 leads E_1 by 90° as shown in Fig. 7.70b. The relation-ships between primary and secondary voltages above and below resonance are shown in Fig. 7.70c and d. Since the cathodes of the diode detectors are bypassed to ground at the signal frequency by C_4 and C_5, the voltage impressed across the diode $V1$ is E_a and the voltage impressed across the diode $V2$ is E_b. The rectified current from $V1$ flows through R_2 and the r-f choke which provides a d-c return path for the rectified current. Similarly, the rectified current from $V2$ flows through R_3 and the r-f choke. Since the rectified currents flow in opposite directions, the net voltage across R_2 and R_3 is zero when E_a and E_b are equal and R_2 and R_3 are equal. When E_a exceeds E_b,
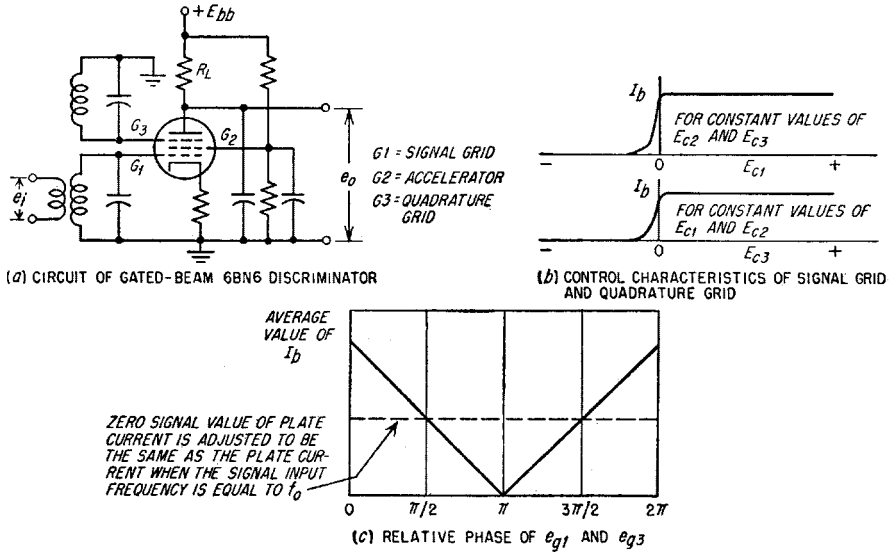
FIG. 7.71. Gated-beam discriminator.

the output voltage is positive. The discriminator characteristic shown in Fig. 7.70e is thus achieved. The polarity of the discriminator curve can be reversed by reversing the secondary winding or by reversing both diodes. For values of mutual coupling between primary and secondary less than critical, the frequency separation of the peaks of the discriminator characteristic is approximately

$$\Delta f = \frac{f_o}{Q_s} \qquad (7.209)$$

where Δf = peak separation, cps
  f_o = resonant frequency, cps
  Q_s = Q of secondary circuit including loading of diodes

A tube designed primarily for use as a combined limiter and discriminator is a 6BN6. It is known as a "gated-beam discriminator," and a circuit for its use as a limiter and discriminator is shown in Fig. 7.71a. The accelerator electrode is shaped to form the electrons into a sharp beam and provide the signal grid and the quadrature grid with very sharp control characteristics as shown in Fig. 7.71b. This provides the desired limiting action and allows the plate current to flow in pulses. The widths of the pulses are dependent upon the phase difference between the quadrature-grid voltage and the signal-grid voltage.

If the signal-grid and quadrature-grid voltages are large enough to obtain good limiting action (about 10 to 20 volts peak), the average value of plate current will

be a linear function of the phase difference between e_{g1} and e_{g3}. The external circuit connected to the quadrature grid consists of a parallel resonant circuit tuned to the desired center frequency f_o. At f_o the impedance of the resonant circuit is a pure resistance, and because of space-charge coupling (see Sec. 7.5b) a voltage is developed across this resistance which lags the signal voltage by 90°. If the frequency of the incoming signal is lower than f_o, the resonant circuit appears inductive and the phase of the voltage developed across the resonant circuit lags the signal voltage by less than 90°. If the frequency of the incoming signal is higher than f_o, the resonant circuit appears capacitive and the phase of the voltage developed across the resonant circuit lags the signal voltage by more than 90°. Thus, the average value of plate current is varied as the frequency of the applied signal changes. The output voltage
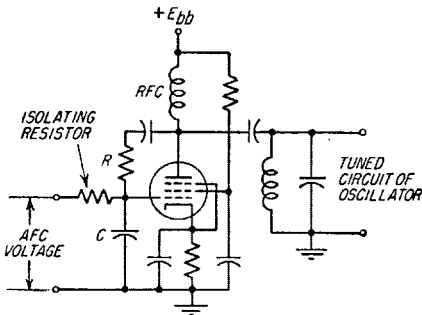
Fig. 7.72. Reactance tube circuit.

from such a discriminator is not zero when the incoming signal frequency is f_o. The cathode bias is normally adjusted to provide a quiescent output voltage which is equal to the output voltage when an input signal at frequency f_o is present. Changes in the frequency of the applied signal either increase or decrease the output voltage.

Reactance Tube Circuits. If AFC is to be applied to a local oscillator of the conventional triode type, the tuning of the oscillator is accomplished by varying the reactance of a reactance tube. The circuit of a reactance tube using a pentode is shown in Fig. 7.72. The reactance tube is connected in parallel with the resonant circuit of the local oscillator. The grid-cathode voltage of the reactance tube is a portion of the voltage existing across the oscillator resonant circuit shifted in phase approximately 90°. The reactance tube plate current is therefore in quadrature with the voltage across the oscillator resonant circuit, and the tube appears as a reactance shunted across the oscillator. The reactance can be made either capacitive or inductive by having the grid-cathode voltage lead or lag the oscillator voltage by 90°. In the circuit of Fig. 7.72, the grid-cathode voltage lags the oscillator voltage by 90°, thereby producing an effective inductive reactance across the oscillator. Neglecting tube capacities, the admittance presented to the oscillator by the reactance tube and the RC phase shifting network shown in Fig. 7.72 is given by

$$Y = \frac{1}{r_p} + \frac{j\omega C}{1 + j\omega RC} + \frac{g_m}{1 + j\omega RC} \qquad (7.210)$$

where ω = frequency of operation, radians/sec

g_m = grid-plate transconductance of reactance tube

If 90° phase shift is obtained across the phase-shift network, the reactive component of the impedance presented by the tube is $j\omega RC/g_m$. At high frequencies the interelectrode capacitances of the tube must also be considered. The general effect of these capacities is to increase the resistive component of impedance presented by the reactance tube to the oscillator.

In an AFC application, the AFC voltage is applied as a bias change to the reactance tube, thereby changing the g_m of the tube. The result is a change in both the tube reactance and the resonant frequency of the oscillator-tuned circuit. The oscillator frequency change that can be achieved with the reactance tube depends upon the relative values of the tube reactance and the tuned circuit passive reactances and also upon the variation in g_m that can be achieved in the tube.

Gas-tube AFC Circuits. Gas-tube AFC circuits have received widespread use in pulse applications. In a typical circuit, one thyratron produces a voltage sawtooth which causes the receiver local oscillator to sweep over a wide frequency range. Any r-f pulse which is received is converted to an i-f pulse and is fed into a discriminator. The polarity of the video output pulse from the discriminator establishes whether the local-oscillator frequency was too low or too high at the instant the pulse was received. The output pulses from the discriminator are applied to another thyratron which takes control from the first thyratron whenever the pulse polarity is correct to fire the second thyratron.[1] In addition to disabling the first thyratron, the reduction in plate voltage at the second thyratron is applied to the local oscillator, thereby causing the oscillator frequency to shift in the direction which reduces the oscillator-frequency error. After the second thyratron deionizes, its plate voltage increases, thereby shifting the oscillator frequency in the other direction. When the drift becomes large enough that a discriminator pulse will again fire the second thyratron, the cycle will repeat itself.

This type of AFC does not have proportional control as does the ordinary discriminator type. Instead, the local-oscillator frequency is continually varied slightly above and below the desired frequency at a slow rate.

Diode-phantastron AFC Circuits. This circuit also has found wide use in pulse applications. In this circuit a discriminator, diode detector, and phantastron sawtooth generator are used to provide a "search" sawtooth of voltage to initially sweep the local oscillator in frequency. When pulses of the proper polarity are received from the discriminator, the phantastron ceases to produce a sawtooth and acts as a conventional d-c amplifier.[2]

2. Absolute-frequency AFC Systems. In an absolute-frequency AFC system, the frequency of the local oscillator is stabilized on an absolute-frequency basis rather than relative to another signal. The same circuit considerations apply for absolute-frequency AFC systems as for relative-frequency systems. However, the discriminator used in the AFC loop must be tuned to the desired local-oscillator frequency rather than to an intermediate frequency as in the relative-frequency AFC system.

7.7. AM Detectors. In general, the signal present at the output of the i-f amplifier of a superheterodyne or at the output of the r-f amplifier in a trf receiver (see Sec. 7.8) consists of a carrier having modulation sidebands symmetrically located above and below the carrier frequency ω_c by an amount equal to the original modulating frequencies. For the special case of a single modulating frequency ω_m, the waveform at the detector input appears as shown in Fig. 7.73a and can be expressed as

$$e_s = E_s(1 + m \sin \omega_m t) \sin \omega_c t \qquad (7.211)$$

$$= E_s \sin \omega_c t - \frac{m}{2} E_s \cos (\omega_c + \omega_m)t + \frac{m}{2} E_s \cos (\omega_c - \omega_m)t$$

which $\quad e_s$ = instantaneous voltage applied to detector

$\qquad E_s$ = peak carrier voltage at frequency ω_c

$\qquad \omega_m$ = modulating frequency

$(\omega_c + \omega_m)$ = upper modulation sideband frequency

$(\omega_c - \omega_m)$ = lower modulation sideband frequency

$\qquad m$ = modulation factor (see Sec. 5.2)

The purpose of the detector is to demodulate the output signal from the i-f or r-f amplifier and present at the detector output the original modulating frequencies. The signal at the output of the detector should ideally be an exact reproduction of the

---

[1] For a detailed explanation of this type of AFC, refer to "Principles of Radar," pp. 458–461, McGraw-Hill Book Company, Inc., New York, 1952.

[2] For a detailed explanation see vol. 23, MIT Series, "Microwave Receivers," Sec. 3.13, pp. 64–69, McGraw-Hill Book Company, Inc., New York, 1948.

original modulating signal. In practical detectors there may be phase, amplitude, and frequency distortion of sufficient magnitude to cause the detected output signal to differ significantly from the original modulating signal.

7.7a. Linear Diode Detection. The most frequently used type of detector is the diode. The analysis of diode detectors is most conveniently divided into (1) large-signal rectification and (2) small-signal rectification (see Sec. 7.7*b*).

The term *linear diode detection* is usually applied to the rectification of large signals, i.e., to the rectification of input signals having peak values greater than approximately one or more volts. Large-signal rectification is based on the principle that the diode
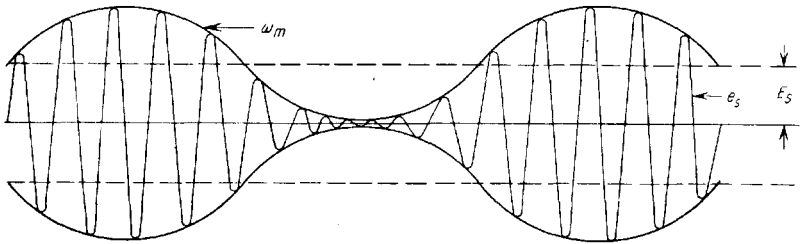
(*a*) WAVEFORM OF AMPLITUDE MODULATED CARRIER
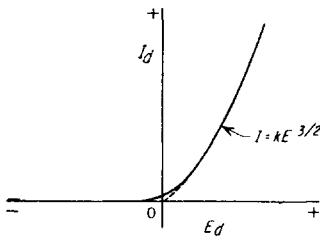
(*b*) TYPICAL VACUUM DIODE
E-I CHARACTERISTIC
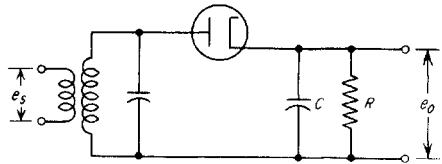
(*c*) BASIC CIRCUIT OF DIODE DETECTOR

Fig. 7.73. Diode detector.

is an "on" or "off" device, i.e., it is assumed to have infinite resistance in the back direction and a constant resistance in the forward direction. In practice this does not preclude the use of crystal diodes provided their back-to-front resistance ratio is very high. In linear diode detection, the diode conducts only on the peak values of the input signal, as in a half-wave rectifier with a capacitor-input filter, and the rectified output voltage is usually 50 to 90 per cent of the peak value of the input voltage. As the name implies, a detector of this type is very nearly linear, and the rectified output voltage is very nearly proportional to the envelope of the input signal. Nonlinearities in the diode characteristics have but a slight effect on the over-all detection linearity.

The *E-I* characteristic of a typical diode is shown in Fig. 7.73*b*, and the basic circuit in Fig. 7.73*c*. In the following analysis of linear diode detection it is assumed that the input signal has a peak value of approximately one or more volts, that the frequency of the input signal is much higher than any modulation frequencies, and that the plate-cathode capacitance of the diode is negligible. Values of R and C in the diode load circuit are selected on the following bases:

1. R is made much larger than the forward resistance of the diode to obtain high rectification efficiency.

2. C is made small enough that the amplitude and phase of the highest modulation frequency are not seriously affected by the time constant RC.

3. C is made as large as (2) will allow to reduce the carrier frequency ripple voltage appearing across R and to maximize the rectification efficiency.

If an unmodulated signal is applied to the detector input, a d-c voltage will be developed across R and C because the diode conducts only during the positive half cycle
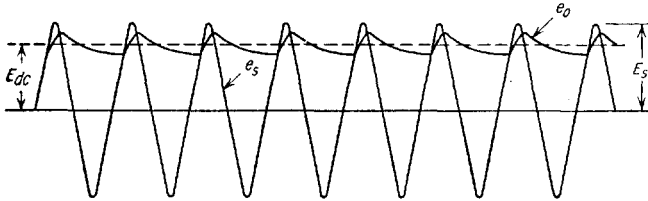
FIG. 7.74. Voltage relationships in diode detector.
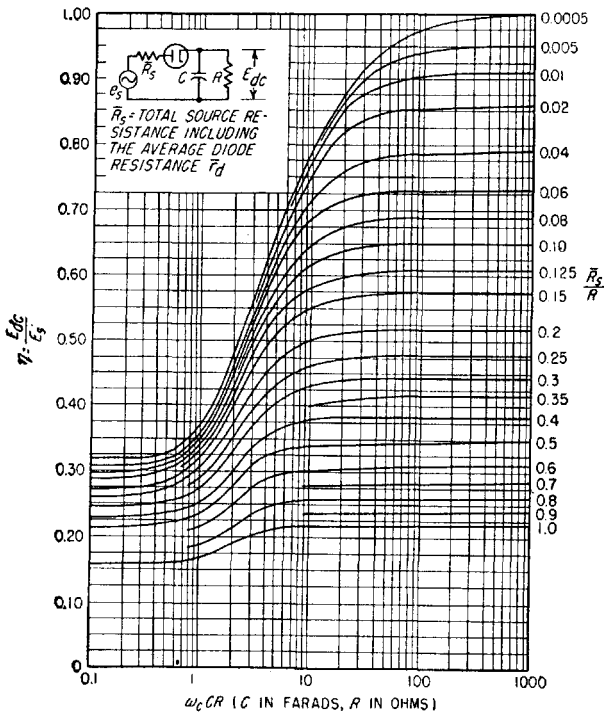
FIG. 7.75. Rectification efficiency of diode detector.

of the applied signal. The voltage developed across C during the positive half cycle cannot completely discharge during the negative half cycle, and after a number of cycles, the voltage across C builds up to a value slightly less than the peak value of the input signal as shown in Fig. 7.74. As the voltage across R and C approaches the input signal peak value E_s, the diode conducts for only the very small portion of the positive half cycle in which e_s exceeds the instantaneous output voltage e_0. The charging current that flows through the diode during the positive peaks when averaged over an entire cycle is equal to the average discharge current that flows through R. The instantaneous output voltage e_0 consists of a d-c component E_{dc} plus an a-c

component called the ripple voltage which has a fundamental frequency component equal to that of the input signal e_s.   The efficiency η of rectification of a diode detector including the effect of the source resistance is defined as

$$\eta = \frac{E_{dc}}{E_s} \qquad\qquad (7.212)$$

where E_{dc} = d-c voltage across R and C

$\qquad E_s$ = peak value of input-signal voltage

The efficiency of rectification achievable in a particular diode detector can be determined from Fig. 7.75 where η is plotted as a function of $\omega_c CR$ for various values of the
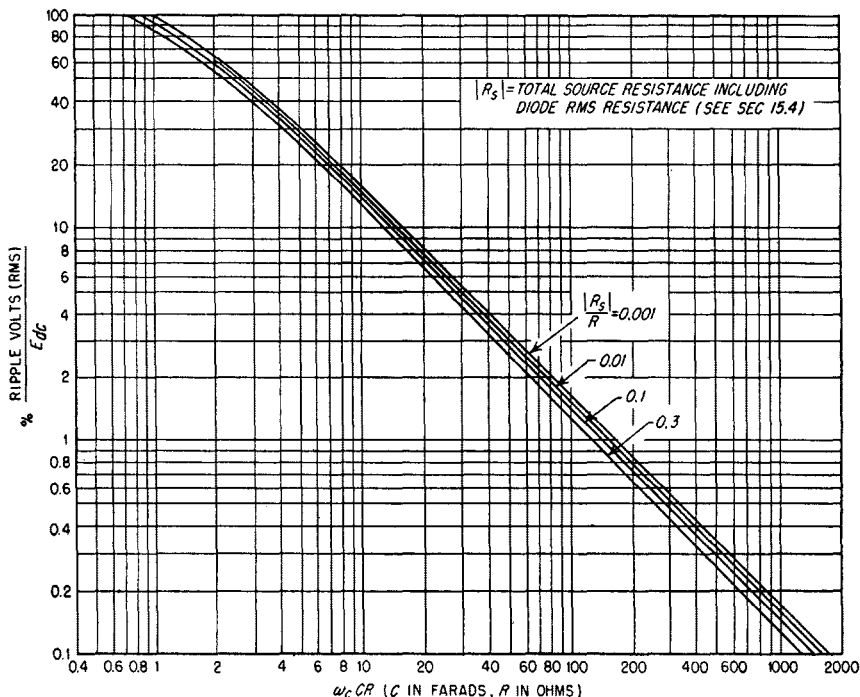
FIG. 7.76. Ratio of ripple voltage to d-c voltage at output of diode detector.

ratio $\bar{R}_s/R$.   The ratio of the rms amplitude of the ripple voltage to the d-c voltage across C and R is given in Fig. 7.76 as a function of $\omega_c CR$ for various values of $|R_s|/R$.

Figures 7.75 and 7.76 are based on the average and rms values of the diode resistance, respectively (see Sec. 15.4); however, these values are nearly equal to one another and to the diode peak resistance as shown by Eqs. (15.4) and (15.5).   Relatively good accuracies can usually be obtained if the diode resistance in Figs. 7.75 and 7.76 is taken to be equal to the diode resistance when the current through the diode is equal to four times the average value of the rectified current through the load resistance.

When the input signal is amplitude-modulated, the average value of the current through the diode and the voltage across R and C vary with the envelope of the input voltage, thus reproducing the modulation across R and C.

The relationships between the voltage applied to the diode, the average rectified current I_{dc}, and the d-c output voltage E_{dc} can be determined from a diode rectification diagram.   Diode rectification diagrams are the equivalent of E-I character-

istics for triode tubes and are usually available from tube manufacturers.   A sample
diode rectification diagram for the 6H6 diode is shown in Fig. 7.77.   The average
rectified current is given as a function of d-c output voltage for various values of rms
input voltage.   Since the d-c output voltage has been shown to be negative, the rectifi-
cation characteristics apply for a diode polarity as shown in Fig. 7.78.   A negative d-c
output voltage is usually desirable since it provides a source of AVC voltage.
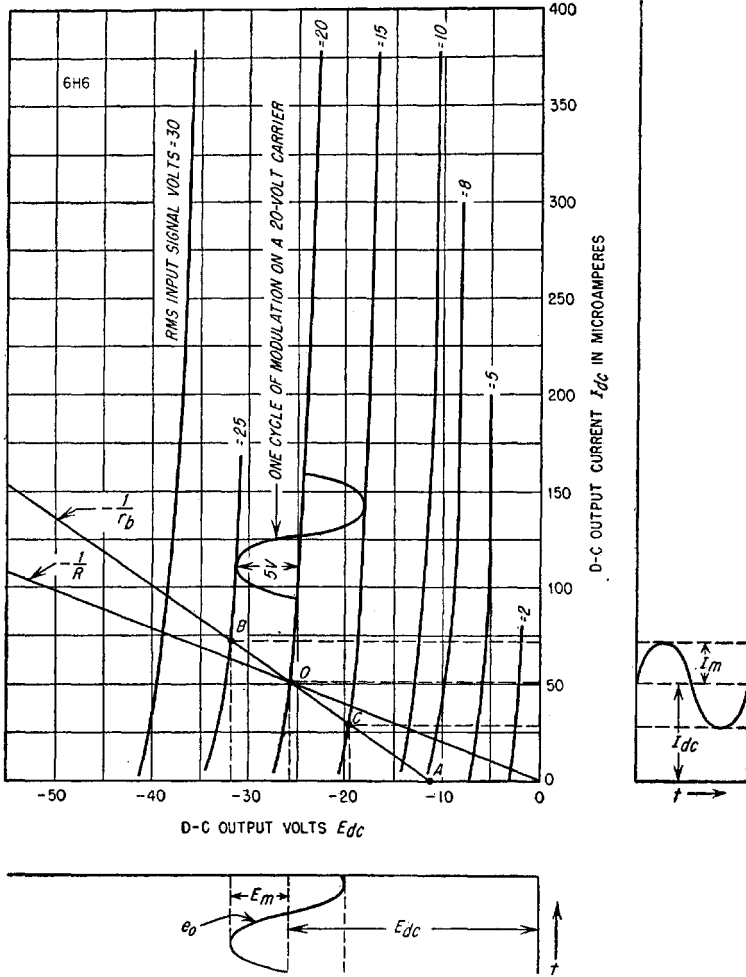
FIG. 7.77. Diode rectification diagram.

A diode rectification diagram can be constructed from the data acquired by making
the following measurements.   Connect the diode as shown in Fig. 7.73c with a micro-
ammeter in series with R.   The measurements can be made at any convenient input
frequency provided that the product $\omega_c CR$ is equal to approximately 100 or more.
While maintaining a constant peak value of input voltage, vary R and measure I_{dc}.
Measure or calculate E_{dc} from the relation

$$E_{dc} = I_{dc}R$$

and plot the resulting curve.   Repeat this procedure for as many different values of input voltage as necessary.

A d-c load line can be constructed on the diode-rectification diagram by drawing a line from the origin through all points satisfying the relation

$$\frac{E_{dc}}{I_{dc}} = R$$

where R = d-c load resistance in detector circuit.   The quiescent point is the inter-section of the d-c load line with the curve having a value equal to the unmodulated value of the input voltage.   As in tube amplifier circuits, the diode a-c load line may
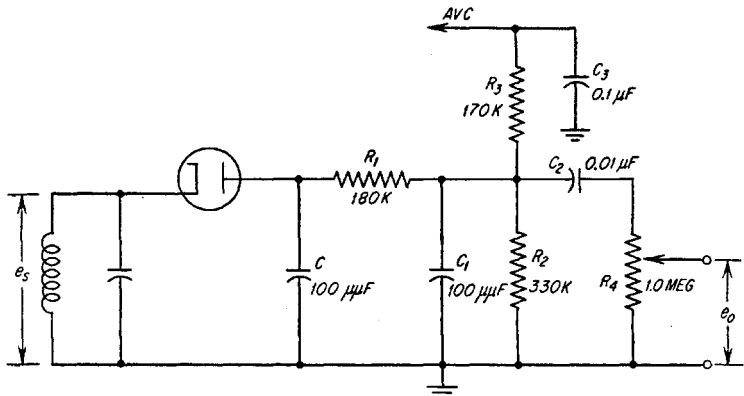
FIG. 7.78. Typical diode detector circuit for a broadcast receiver.

differ from the d-c load line.   As an example, consider the diode detector circuit shown in Fig. 7.78.   The purpose of R_1, C, and C_1 is to filter the ripple at the plate of the diode resulting from the rectification of the carrier. R_3 and C_3 provide a filter to eliminate the audio-frequency signals and allow a portion of the d-c voltage across C to be coupled to the preceding i-f stages for automatic volume control. C_2 prevents the d-c at the junction of R_1 and R_2 from biasing the audio-amplifier stage connected to R_4.   For the circuit of Fig. 7.78, the d-c load resistance is

$$R = R_1 + R_2 = 510 \text{ kilohms}$$

and the a-c load resistance r_b at the modulation frequencies where the reactances of C_2 and C_3 are very small is

$$r_b = R_1 + \frac{1}{1/R_2 + 1/R_3 + 1/R_4} = 281 \text{ kilohms}$$

The d-c load line and a-c load line of Fig. 7.78 are shown in Fig. 7.77 for a quiescent point determined by assuming an rms unmodulated input signal of 20 volts.   The a-c and d-c load lines are shown to cross at the d-c quiescent point O.   This will be exactly true only if the diode characteristics are linear within the swing of the modulation cycle.   Because the a-c load line is steeper than the d-c load line, the percentage of modulation can be increased only until the negative peak of the modulation cycle reaches point A, otherwise the output current is reduced to zero even though the input signal is not equal to zero.   If the per cent modulation is increased above this value, the negative peaks will be clipped as shown in Fig. 7.79a.   If negative-peak clipping

(a) NEGATIVE PEAK CLIPPING DUE TO $r_b < R$ AND HIGH PERCENTAGE MODULATION

(b) DIAGONAL CLIPPING DUE TO REACTIVE LOAD AND HIGH PERCENTAGE MODULATION

a = PERIOD OF EXPONENTIAL DISCHARGE
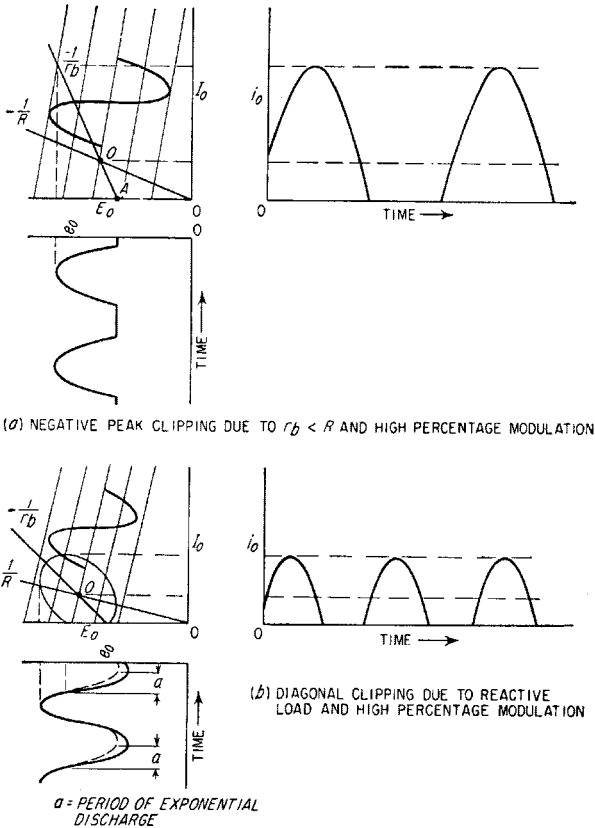
FIG. 7.79. Clipping due to diode a-c load impedance different from diode d-c load impedance.

is to be prevented, the ratio m of the peak a-c load voltage to the d-c load voltage must be

$$m \leq \frac{r_b}{R} \qquad (7.213)$$

The maximum allowable modulation factor of the input signal is

$$m_{max} = \frac{e_1 - e_2}{e_1} \qquad (7.214)$$

where e_1 = value of input signal at point O
e_2 = value of input signal at point A

The distortion of the detected modulation due to nonlinearity in the diode character-istics can be determined in a manner exactly analogous to the method discussed in Sec. 3.3b for triode amplifiers except that the input signal is taken as the largest value of modulation signal impressed upon the carrier.

At the lower modulation frequencies, where the reactances of C_2 and C_3 cannot be considered to be equal to zero, and at the higher modulation frequencies, where the reactances of C and C_1 cannot be considered to be infinite (see Fig. 7.78), the diode load will be a complex impedance having both resistive and reactive components and the path of operation on the diode rectification diagram will be an ellipse. As the percent-

age of modulation is increased from that shown in Fig. 7.77, the ellipse will increase in size until the bottom of the ellipse becomes tangent to the zero current line at point A. If the percentage of modulation is further increased, diagonal clipping of the detected modulation will occur as shown in Fig. 7.79b.

   1. *Equivalent Circuit of a Linear Diode Detector.*   A comparison of a diode rectification diagram (see Fig. 7.77) with the E_b-I_b characteristics of a triode reveals a striking similarity. If the diode characteristics are defined by Eqs. (7.215) and (7.216), an equivalent circuit similar to that of a triode amplifier can be constructed for a linear diode detector.
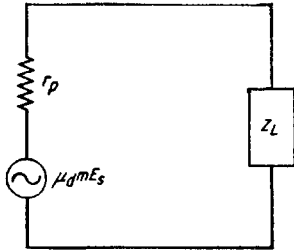
Fig. 7.80. Equivalent circuit of a linear diode detector.

$$\mu_d = \frac{\Delta E_{dc}}{\Delta E_s}\bigg]_{I_{dc}=\text{const}} \qquad (7.215)$$

$$r_p = \frac{\Delta E_{dc}}{\Delta I_{dc}}\bigg]_{E_s=\text{const}} \qquad (7.216)$$

where μ_d = detection μ of diode
   r_p = detection plate resistance of diode

The equivalent circuit is shown in Fig. 7.80.   The equivalent voltage generator for the detected modulation frequencies is

$$E_m = \mu_d m E_s \qquad (7.217)$$

where m = modulation factor at diode input
   E_m = peak value of detected modulation
   E_s = peak value of unmodulated carrier at diode input

The equivalent amplification A_d of the diode detector, expressed as the ratio of the detected modulation voltage E_m across the diode load impedance Z_L to the amplitude of the modulation at the diode input, is given by

$$A_d = \frac{E_m}{mE_s} = \mu_d \frac{Z_L}{r_p + Z_L} \qquad (7.218)$$

where Z_L = diode load impedance at frequency of the detected modulation

   2. Input Impedance of a Linear Diode Detector.   The input impedance of a linear diode detector can be determined from the diode forward resistance r_d and the efficiency of rectification η.   Assuming RC to be large enough that the voltage across C does not change appreciably between periods of diode conduction, i.e., assuming zero ripple across the diode load, the average power P_d delivered to the detector by the carrier is

$$P_d = \frac{E_s^2}{2\pi r_d}\left(\cos^{-1}\eta - \eta\sqrt{1 - \eta^2}\right) \qquad (7.219)$$

where E_s = peak value of carrier at diode input
   η = efficiency of detector assuming source resistance to be equal to zero (see Fig. 7.75)

The general form of the equation expressing the power dissipated in a resistor having a voltage with a peak value equal to E_s applied across it is

$$P = \frac{E_s^2}{2R} \qquad (7.220)$$

Therefore the effective diode input resistance R_i from Eq. (7.219) is

$$R_i = \frac{\pi r_d}{\cos^{-1}\eta - \eta\sqrt{1 - \eta^2}} \qquad (7.221)$$

The diode appears as a constant load resistance to the signal source throughout the entire i-f cycle, even though the diode conducts for only part of the cycle. This occurs because of the filtering action of the resonant circuit in parallel with the detector input. Equation (7.221) is plotted in Fig. 7.81. If the input signal is amplitude-modulated, the input resistance of the detector to the sideband frequencies will be a function of the a-c load impedance at the modulation frequency corresponding to the difference between that sideband frequency and the carrier frequency. If the diode loss is assumed negligible, the effective input resistance R'_i at a particular sideband frequency is given by

$$R'_i = \frac{|r_p + Z_L|^2}{2\mu_d{}^2 r_b} \qquad (7.222)$$

where Z_L = diode load impedance at detected modulation frequency
r_b = resistive component of Z_L

If the diode load at the modulation frequency is purely resistive and much larger than r_p, Eq. (7.222) reduces to

$$R'_i = \frac{r_b}{2\mu_d{}^2} \qquad (7.223)$$

FIG. 7.81. Effective input resistance of a diode detector as a function of the detector efficiency where $R_s = 0$ (see Fig. 7.75).

3. *Effect of a Linear Diode Input Impedance upon the Modulation of an Input Signal.* The peak signal voltage appearing across the input terminals of a diode detector is a function of the detector effective input impedance R_i, the signal source resistance R_s, and the peak generator open-circuit voltage E_s.

$$E_i = E_s \frac{R_i}{R_s + R_i} \qquad (7.224)$$

Since the diode input impedance R'_i at the sideband frequencies may be different from the input impedance R_i at the carrier frequency, the effective percentage of modulation of the signal at the input to the diode may be changed if the source resistance is an appreciable fraction of the diode detector input resistance. The relation between the percentage modulation at the diode input and the percentage modulation at the generator is given by

$$\frac{\text{Modulation percentage at diode input}}{\text{Modulation percentage at generator}} = \frac{R'_i}{R_i}\left(\frac{R_i + R_s}{R'_i + R'_s}\right) \qquad (7.225)$$

where R_s = generator impedance at carrier frequency
R'_s = generator impedance at sideband frequencies

The ratio of the effective modulation percentage at the diode detector input to the percentage modulation at the original signal source is also affected by the frequency-response characteristics of any intervening stages, for example, r-f amplifier, mixer, i-f amplifier.

Example 7.13

Design a linear diode detector using a 6AL5 diode. Assume that the rms unmodulated carrier at the diode input is 20 volts at 1 Mc and that 80 per cent amplitude modulation

is to be rectified with negligible distortion.   The upper -3-db frequency of the detector circuit is to be 10 kc, and the carrier ripple voltage at the output of the detector is not to exceed 1 volt rms.   The diode rectification diagram for a 6AL5 is shown in Fig. 7.82.

Solution

1. Determine $\omega_c RC$.
   The rectification diagrams supplied by tube manufacturers are prepared on the basis that the diode load capacitance C is large enough to make $\omega_c RC$ equal to approximately 100 or more.   For smaller values of $\omega_c RC$, the rectification diagram will be in error as evidenced by Fig. 7.75.   The maximum value of RC is determined by the highest modulation frequency to be passed with -3-db attenuation.   Thus,

$$RC = \frac{1}{2\pi f_{max}}$$
$$= \frac{1}{2\pi \times 10^4}$$
$$= 15.9 \times 10^{-6}$$

Therefore $\omega_c RC = 6.28 \times 10^6 \times 15.9 \times 10^{-6}$
$$= 100$$

FIG. 7.82. Diode rectification curves for Example 7.13.

and the rectification diagram can be used with sufficient accuracy.

2. Determine the diode load resistance R.
   R should be large compared to the diode forward resistance for high efficiency and low ripple.   Therefore let the diode load resistance be equal to 100 kilohms.   The 100-kilohm load line in Fig. 7.81 shows the efficiency at the operating point to be

$$\eta = \frac{E_{dc}}{E_s} = \frac{25.8}{20\sqrt{2}} = 0.912$$

3. Determine C.

$$C = \frac{100}{\omega_c R}$$
$$= \frac{100}{2\pi \times 10^6 \times 10^5}$$
$$= 159 \ \mu\mu f$$

4. Determine the ripple voltage at the detector output.
   From Fig. 7.75 it is found that for $\eta = 0.912$ and $\omega_c RC = 100$, the ratio $\bar{R}_s/R$ is equal to 0.009.   From Fig. 7.76 the ratio of rms ripple to E_{dc} is found to be approximately 1.7 per cent.   Therefore

$$E_{ripple} = 0.017 \times 25.8$$
$$= 0.439 \text{ volt rms}$$

If the ripple had exceeded the allowable 1.0 volt an additional LC or RC filter would have to be added to the detector load circuit.

5. Determine the minimum value of grid resistor R_c for the next stage.
   The minimum value of the input voltage with 80 per cent modulation is

$$E_{s(min)} = 0.20 \times 20 = 4.0 \text{ volts rms}$$

Assuming linear diode rectification, the minimum a-c load resistance can be determined from Eq. (7.213).

$$r_b = mR$$
$$= 0.8 \times 100,000$$
$$= 80,000 \text{ ohms}$$

Since $$r_b = \frac{R_c R}{R_c + R}$$

$$R_c = 400,000 \text{ ohms minimum}$$

6. Determine the input resistance at ω_c.

From step 4 it was established that R_s/R was equal to 0.009. Since the diode rectification curves are based on zero source resistance, r_d/R is equal to 0.009. Therefore

$$r_d = 0.009R = 900 \text{ ohms}$$

From Fig. 7.81

$$\frac{R_i}{r_d} = 62$$

and

$$R_i = 55,800 \text{ ohms}$$

If the detector input does not include a tuned circuit, the effective diode input resistance will not remain constant throughout the i-f cycle and the conventional

(a) DIODE DETECTOR WITHOUT LOAD

(b) DIODE RECTIFICATION OF SMALL SIGNALS
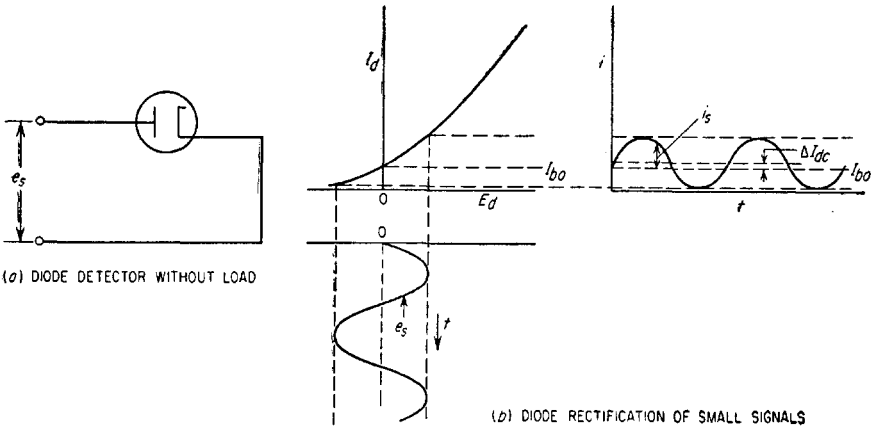
FIG. 7.83. Small-signal diode rectification.

rectification diagram cannot be used. In this case either a special rectification diagram including source resistance can be prepared as described in Sec. 7.7a, or the linear diode detector can be analyzed utilizing Figs. 7.75 and 7.76 only. In the latter case, it is satisfactory to assume that the diode average and rms resistance values are equal to the resistance of the diode when conducting a current four times the rectified current. This requires an initial estimate of the rectified diode current which can later be compared with the value determined from Fig. 7.75.

7.7b. *Small-signal Diode Detectors.* The E-I characteristic of a typical diode is similar to the solid curve in Fig. 7.73b and has considerable curvature at $E_d = 0$. The current flow at $E_d = 0$ occurs because a certain number of the electrons emitted by the cathode have initial velocities sufficient to enable the electrons to reach the plate even when the plate-cathode voltage is zero or even slightly negative.

The detection of small signals in a diode detector is accomplished by utilizing the nonlinearity of the diode characteristic near the origin. Shown in Fig. 7.83b are the current and voltage relationships for a diode connected as shown in Fig. 7.83a and having a small-signal input. Note that the diode conducts for 360° of the small input signal instead of for only a few degrees as in the case of a linear diode detector. From the figure it can be seen that throughout the positive half cycle of the applied voltage the diode current will be increased more than it is decreased during the negative half cycle of the applied voltage. The result is that the average diode current is increased from I_{bo}, the average diode current with no applied signal, to I'_{bo}. This increase in the average diode current due to the presence of a signal constitutes the method of small-signal detection in diodes.

The instantaneous current variations due to an input signal in the circuit shown in Fig. 7.83 can be approximated by a Taylor's series evaluated at the operating point O.

$$i_s = \frac{\partial i_d}{\partial e_d} e_s + \frac{1}{2!} \frac{\partial^2 i_d}{\partial e_d{}^2} e_s{}^2 + \frac{1}{3!} \frac{\partial^3 i_d}{\partial e_d{}^3} e_s{}^3 + \cdots \qquad (7.226)$$

where $e_s = E_s \sin \omega_c t =$ instantaneous value of applied voltage

$i_s =$ instantaneous diode current due to e_s

The derivatives are to be taken at the operating point. This method is analogous to the method of determining the harmonic distortion in amplifiers. In diode detection, however, only those terms contributing to an increase in the average diode current are
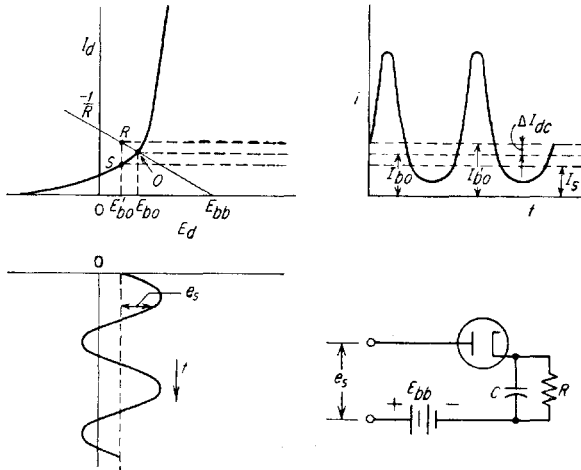
FIG. 7.84. Current and voltage relationships in a small-signal diode detector.

significant and need be considered. The odd derivative terms do not contribute to the average current and may be neglected. The second derivative term contributes by far the greatest amount to the change in average current and ordinarily is the only one considered. If only the second derivative term is considered, the average diode current increases as the square of the amplitude of the applied voltage and the detector is commonly referred to as a *square-law detector*. The increase in average current ΔI_{dc} for a square-law detector is given by

$$\Delta I_{dc} = \frac{1}{4} \frac{\partial^2 i_d}{\partial e_d{}^2} E_s{}^2 \qquad (7.227)$$

where the second derivative is taken at the operating point.

Since the change in average diode current is a function of the rate of change of slope of the diode characteristic at the operating point, that is, $\partial^2 i_d / \partial e_d{}^2$, the diode should be biased so that the operating point O is positioned where the diode characteristic curvature is greatest. Note that the detector is a square-law detector provided $\partial^2 i_d / \partial e_d{}^2$ is not equal to zero and that the term square-law detector does not stem from any particular E-I characteristic.

If a load consisting of R and C parallel, which is assumed to have negligible reactance at the input frequency, is inserted in series with the diode, the current and voltage relations are as shown in Fig. 7.84. The bias voltage E_{bb} is necessary to locate the dynamic operating point S in the region of greatest curvature on the diode character-

istic. The quiescent operating point is point O. When the r-f signal is applied to the detector input, the average diode current increases from I_{bo} to I'_{bo}. This increase in average current increases the voltage drop across R and decreases the average diode plate-cathode voltage from E_{bo} to E'_{bo}. The instantaneous value of diode current when the instantaneous applied voltage e_s is zero is I_s and is determined by the intersection of the diode characteristic curve with E'_{bo}. The a-c component of diode current varies
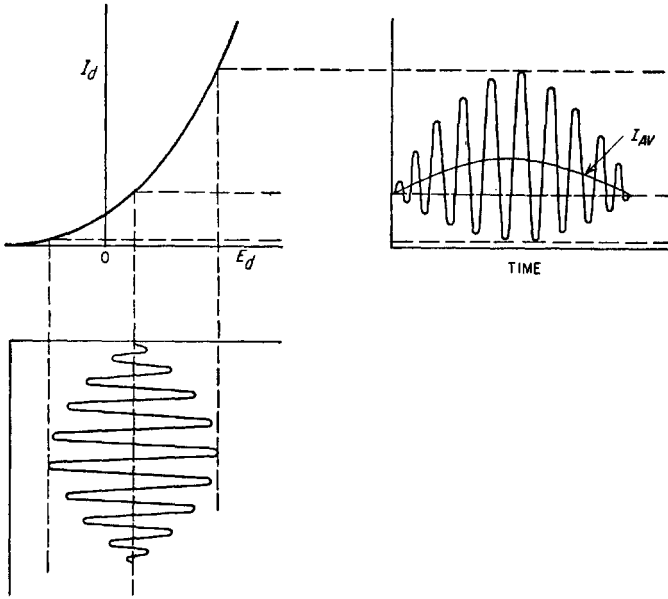
FIG. 7.85. Demodulation of small signals in a diode detector.

about the value I_s. The increase in average diode current from the zero signal condition is

$$\Delta I_{dc} = I'_{bo} - I_{bo} \tag{7.228}$$

If the points S and O are assumed close enough together so that the slope of the characteristic curve can be considered constant over that interval, the increase in average diode current can be expressed as

$$\Delta I_{dc} = \frac{E_s^2}{4} \frac{\partial^2 i_d}{\partial e_d^2} \frac{r_d}{r_d + R} \tag{7.229}$$

where r_d = dynamic resistance of diode at operating point
If the diode E-I characteristic is defined by the relationship $I_d = kE_d^2$, where k is a constant, the second derivative is equal to $2k$. For this case, Eq. (7.229) is given by

$$\Delta I_{dc} = \frac{kE_s^2 r_d}{2(r_d + R)} \tag{7.230}$$

If the input signal to a square-law detector consists of a modulated carrier, the modulation signal is detected in the manner illustrated in Fig. 7.85. The modulation results in a variation in the average value of diode current because of the nonlinear diode characteristic. However, the diode current contains components at the carrier frequency, the modulating frequency, the sum and difference between the carrier and

modulating frequencies, and the 2nd harmonic of each of these signals.   This can be seen by considering the input voltage e_s in Eq. (7.226) to consist of a carrier modulated by a signal frequency giving

$$e_s = E_s \sin \omega_c t + \frac{mE_s}{2} [\cos (\omega_c - \omega_m)t - \cos (\omega_c + \omega_m)t] \qquad (7.231)$$

Evaluation of Eq. (7.226) for this condition results in the frequencies mentioned.   In addition, if more than one modulating signal is present, the diode current contains components at the sum and difference frequencies of every pair of sideband frequencies present.   Thus, signal components are present at the output of a square-law detector
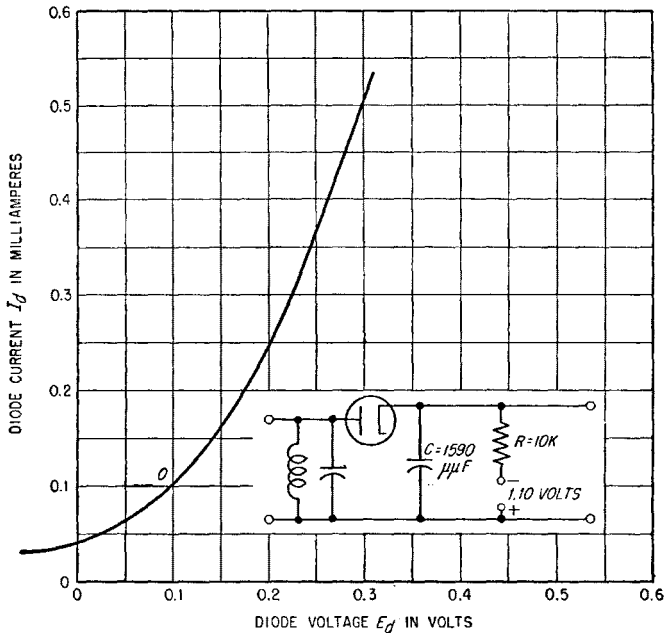
Fig. 7.86. Characteristic curve of the diode for the small-signal detector in Example 7.14.

which were not present in the original signal.   The extraneous signals which are produced cannot always be filtered out because they may fall within the necessary bandpass of the detector output circuit.

In determining the RC load of the small-signal detector, C should be chosen so that its reactance is much smaller than r_d at the signal frequency to develop the maximum signal voltage across r_d.   Load resistance R and supply voltage E_{bb} should be selected to make the operating point S occur at the point of greatest curvature on the diode E-I characteristic.   The maximum output voltage is obtained at a given operating point S by making E_{bb} large and selecting R to provide operation at the point S.   The maximum value of R may be limited by the required bandwidth of the diode load impedance composed of R and C in parallel.   If C is essentially a short circuit at the input frequencies, the input impedance of the small-signal nonlinear diode detector is equal to the diode r_d evaluated at the operating point.

Example 7.14

Given the characteristic in Fig. 7.86, design a small-signal detector for use in detecting a 1-Mc amplitude modulated carrier.   The bandwidth of the load impedance must be 10 kc-

1. Determine the desired operating point on the diode characteristic curve.
The optimum operating point is the point of maximum curvature, O in Fig. 7.86. For small input signals, the operating point will not shift appreciably from point O and the diode resistance r_d will remain essentially unchanged.

2. Determine the diode dynamic resistance r_d at the operating point

$$r_d = \frac{\Delta E_d}{\Delta I_d} = 1{,}000 \text{ ohms}$$

3. Determine C.
At $f = 1$ Mc, X_c should be $\ll r_d$. Therefore let

$$X_c = 0.1 \, r_d \qquad \text{at } f = 1 \text{ Mc}$$

$$C = \frac{1}{2 \times 3.14 \times 10^6 \times 0.1 \times 1{,}000}$$

$$= 1{,}590 \; \mu\mu\text{f}$$

4. Determine R.
Since the -3-db upper frequency of the diode load impedance is to be 10 kc

$$R = \frac{1}{\omega C} \qquad \text{at } f = 10 \text{ kc}$$

$$= \frac{1}{2 \times 3.14 \times 10^4 \times 1{,}590 \times 10^{-12}}$$

$$= 10{,}000 \text{ ohms}$$

5. Determine the necessary bias to locate the operating point at point O with a load resistance of 10 kilohms.

$$\Delta E = I_o R$$
$$= 0.1 \times 10^{-3} \times 10{,}000 = 1.0 \text{ volt}$$

Therefore, source voltage must equal $E_o + \Delta E = 1.10$ volts.

7.7c. Grid-leak Detectors. The detection of an amplitude-modulated signal can be accomplished in grid-controlled tubes as well as in diodes. If the detection is accomplished in the grid circuit, the detector is commonly referred to as a grid-leak detector. In Fig. 7.87 the grid and cathode act as the plate and cathode, respectively, of a diode detector. The d-c bias and the detected modulation of the input signal appear across a grid leak composed of R and C in parallel. As in the diode detector, grid-leak detection can be either linear or square-law, depending upon the amplitude of the incoming signal.

Fig. 7.87. Circuit of a linear grid-leak detector.

1. Linear Grid-leak Detection. If the signal applied to a grid-leak detector has an amplitude of several volts, the action in the grid circuit is very much like that of a linear diode detector. Typical waveforms for the grid circuit are shown in Fig. 7.88. The current that flows during that portion of a cycle of the applied signal that the grid is positive with respect to the cathode charges the grid-leak capacitor C to nearly the peak value of the applied signal. During the remainder of the cycle of applied signal, some of the charge on C is discharged through R. The voltage across R and C thus consists of a d-c component due to the carrier, an a-c component due to the modulation of the carrier, and a ripple component at the signal frequency. The instantaneous grid-cathode voltage is equal to the applied signal voltage minus the voltage developed across the grid leak

Because the tube-plate current contains a large amplitude component at the applied signal frequency, the plate circuit must contain an LC or RC filter to minimize this ripple.

R and C should be selected on the same basis as for the linear diode detector discussed in Sec. 7.7a.

The plate signal voltage e_o is given approximately by

$$e_o = \frac{-\mu Z_L E_m}{r_p + Z_l} \qquad (7.232)$$

where E_m = modulation component of voltage developed across grid leak

Z_L = plate-load impedance at modulation frequencies

(a) APPLIED SIGNAL VOLTAGE

(b) GRID-CATHODE VOLTAGE WAVEFORM

(c) PLATE CURRENT

(d) FILTERED PLATE VOLTAGE

FIG. 7.88. Waveforms in a linear grid-leak detector.

2. *Small-signal Grid-leak Detection.* When the applied signal amplitude is small, the grid-leak detector behaves very much like a square-law diode detector. In the region near zero bias, the E_c-I_c characteristic of the tube has considerable curvature and, because of the initial velocity of the electrons emitted by the cathode, the grid-cathode current is not zero at zero bias. Operation is identical to that shown in Figs. 7.84 and 7.85 for the small-signal diode detector. The increase in average current during the positive modulation crest increases the average voltage across the grid leak. This represents an increase in the negative bias between the grid and cathode, causing the average tube plate current to decrease. The instantaneous plate current varies in accordance with the amplitude of the applied signal, and these variations must be filtered out at the plate of the tube with an RC or LC filter.

In the design of a small-signal grid-leak detector, the E_c-I_c characteristic of the tube should be obtained as in the case of the diode and the values of R and C for the grid

(a) CIRCUIT OF PLATE DETECTOR

(b) PROJECTED CUTOFF FOR PLATE DETECTION
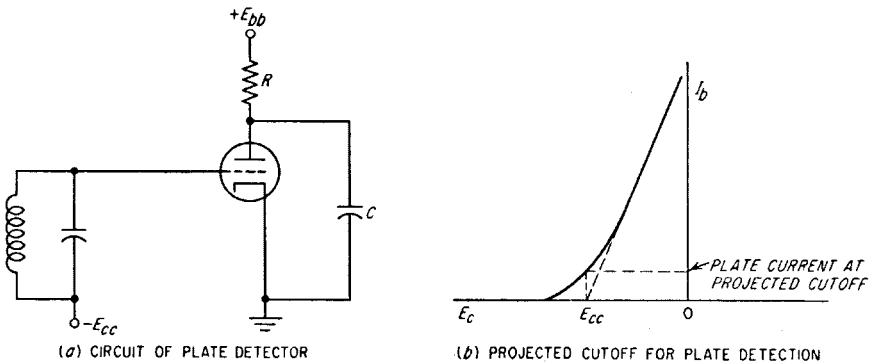
FIG. 7.89. Linear plate detector.

leak determined just as they were obtained for the diode detector. The tube should be biased to achieve the greatest curvature of the E_c-I_c curve at the operating point. Once the modulation voltage across the grid leak has been determined, the output signal voltage from the tube can be calculated by means of Eq. (7.232).

7.7d. Plate Detectors. Detection of a modulated carrier can also be accomplished in the plate circuit of a vacuum tube. The circuit of such a detector is shown in Fig. 7.89a. The plate detector may be either a large-signal (linear) detector or a small-signal (square-law) detector.

1. Linear Plate Detection. In the linear plate detector, the tube is biased to its "projected cutoff" as shown in Fig. 7.89b. In the absence of the shunting capacitor, a large input signal causes the plate current to increase almost linearly during the positive half cycle of the applied signal and causes the plate current to be cut off during the majority of the negative half cycle. The plate current varies as shown in Fig.

(a) GRID-CATHODE VOLTAGE

(b) PLATE VOLTAGE
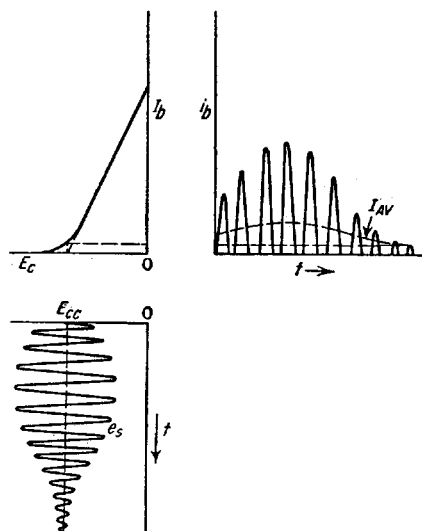
(c) PLATE CURRENT

Fig. 7.90. Waveforms in a linear plate detector with resistive plate load.
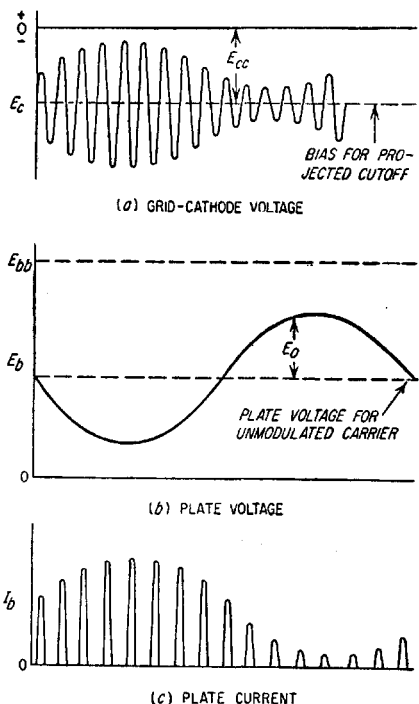
Fig. 7.91. Waveforms in a triode linear plate detector.

7.90, and the average plate current is approximately equal to I_{peak}/π. If a capacitor is placed in parallel with the load resistor as shown in the figure, the capacitor will take a charge which is a function of the tube resistance and the plate-load resistance. If the tube is a triode, a large plate-load resistor can be used, thereby making the time constant determined by the tube and the capacitor much shorter than the time constant determined by the plate-load resistor and the capacitor. If this is the case, the average voltage across the capacitor will be slightly greater than the minimum instantaneous plate voltage which exists across the tube in the absence of the capacitor. The average voltage across the capacitor will decrease as the average tube current is increased, as is the case when the amplitude of the input carrier is increased. It is in this manner that the modulation on the carrier is detected in the plate circuit. In the presence of an input carrier, the plate current flows for somewhat less than one-half the input cycle since the decreased plate voltage due to effect of C reduces the

value of grid bias required for cutoff.   However, the peak plate current is greater than when R alone is present because of the lower plate voltage source impedance provided by the capacitor.   The voltage and current relationships are as shown in Fig. 7.91.

The performance of a linear plate detector can be determined by a rectification
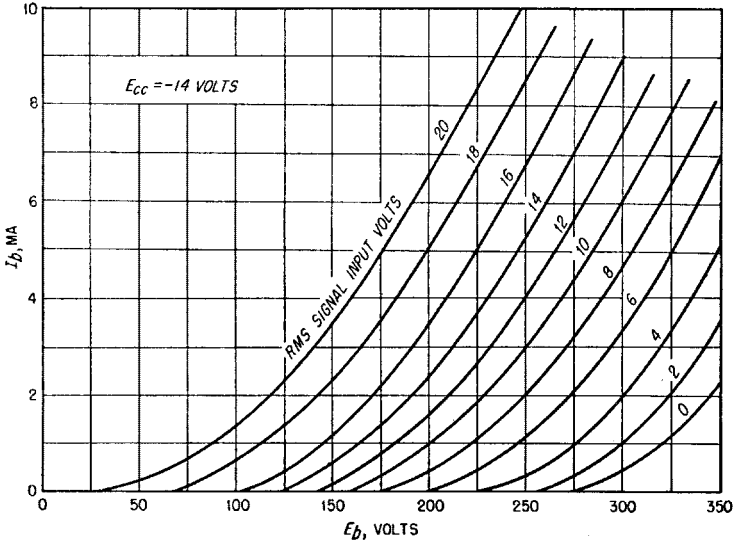
FIG. 7.92. Rectification diagram for typical triode linear plate detector.

diagram as shown in Fig. 7.92.   The rectification diagram shows the average tube current as a function of plate voltage for a particular value of grid bias and various values of input signal.   The plate-load line is constructed exactly as is done for the tube as an amplifier, the plate voltage equaling the supply voltage when the tube current is zero.   The proper value of grid bias is obtained for a particular supply

(a) TRIODE LINEAR PLATE DETECTOR
    EQUIVALENT CIRCUIT

(b) PENTODE LINEAR PLATE DETECTOR
    EQUIVALENT CIRCUIT
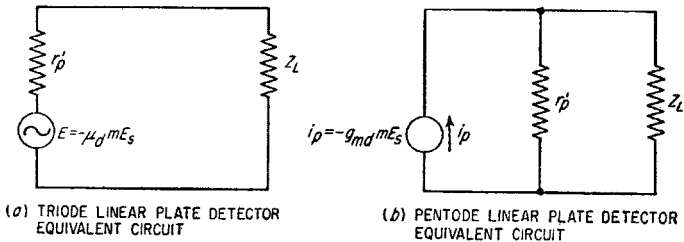
FIG. 7.93. Plate detector equivalent circuits.

voltage and load impedance by determining the projected cutoff of the tube from the I_b-E_c characteristic as shown in Fig. 7.90b.

An equivalent output circuit for the triode plate detector is shown in Fig. 7.93a. The effective voltage gain A of the detector is

$$A = \frac{-E_o}{mE_s} = -\mu_d \frac{Z_L}{r_p' + Z_L} \qquad (7.233)$$

where E_o = peak output voltage at modulation frequency

$\quad\quad m$ = modulation factor

$\quad\quad E_s$ = peak carrier voltage

$\quad\quad \mu_d$ = detection μ of tube (see Fig. 7.92)

$$\quad\quad = \frac{\partial E_{dc}}{\partial E_s}\bigg|_{I_b=\text{const}}$$

$\quad\quad r_p'$ = effective plate resistance of detector (see Fig. 7.92)

$$\quad\quad = \frac{\partial E_b}{\partial I_b}\bigg|_{E_s=\text{const}}$$

$\quad\quad Z_L$ = load impedance

In practice, μ_d will usually be between 0.7 and 0.9 times the μ as an amplifier, and r_p' will be about three times r_p for the tube as an amplifier. Exact values of μ_d and r_p' can be taken directly from the rectification diagram.

In pentode plate detectors the plate load resistance R is ordinarily much smaller than r_p'. The equivalent plate circuit for a pentode plate detector is shown in Fig. 7.93b. The effective voltage gain A is

$$A = \frac{-E_o}{mE_s} = -g_{md}Z_L \tag{7.234}$$

where g_{md} = detection transconductance for tube

$$g_{md} = \frac{\partial I_b}{\partial E_s}\bigg|_{E_b=\text{const}}$$

Rectification diagrams can be constructed for the pentode plate detector just as for the triode plate detector. However, because r_p' is much larger than R and the tube current is nearly independent of the plate-load impedance, the addition of a capacitor C across R does not change the average tube current significantly. Therefore, if C is large, the current through R will be nearly constant and will be approximately equal to I_{peak}/π where I_{peak} is the peak current through R in the absence of C. Thus, C does not improve the detection efficiency, but only serves to filter the ripple voltage across R.

In the pentode plate detector, g_{md} is approximately equal to $1/\pi$ times the g_m of the tube as an amplifier, and r_p' is about three times the r_p of the tube as an amplifier. Exact values of g_{md} and r_p' can be taken from the rectification diagram for the tube.

Example 7.15

A 6AH6 pentode is to be used as a linear plate detector. Determine the tube operating conditions and effective gain for a plate supply of 300 volts, screen supply of 150 volts, and an input signal which has a 1-Mc carrier, a 1.5-volt peak amplitude, and which is 100 per cent modulated. The detector upper −3-db frequency limit shall not be lower than 10 kc. For exact design, a rectification diagram should be constructed for the tube. However, approximate results can be obtained as follows:

1. Determine the projected cutoff bias for the 6AH6.

From Fig. 7.94, the projected bias is found to be approximately −3.6 volts.

2. Determine the plate-load resistor R.

R should be as large as possible while still allowing adequate frequency response and low distortion. For good linearity at the positive crest of the modulation, the load line should intersect the bias curve representing the smallest instantaneous value of grid-cathode voltage above the knee of the characteristic curve. Thus, the minimum instantaneous bias is equal to −0.6 volt since the peak input-signal voltage is 3 volts and the quiescent bias is −3.6 volts. Therefore, a load resistor of approximately 10 kilohms should be used (see Fig. 7.94).

3. Determine the value of C.

The capacitor C shunting R should be as large as permitted by the required modulation bandwidth.
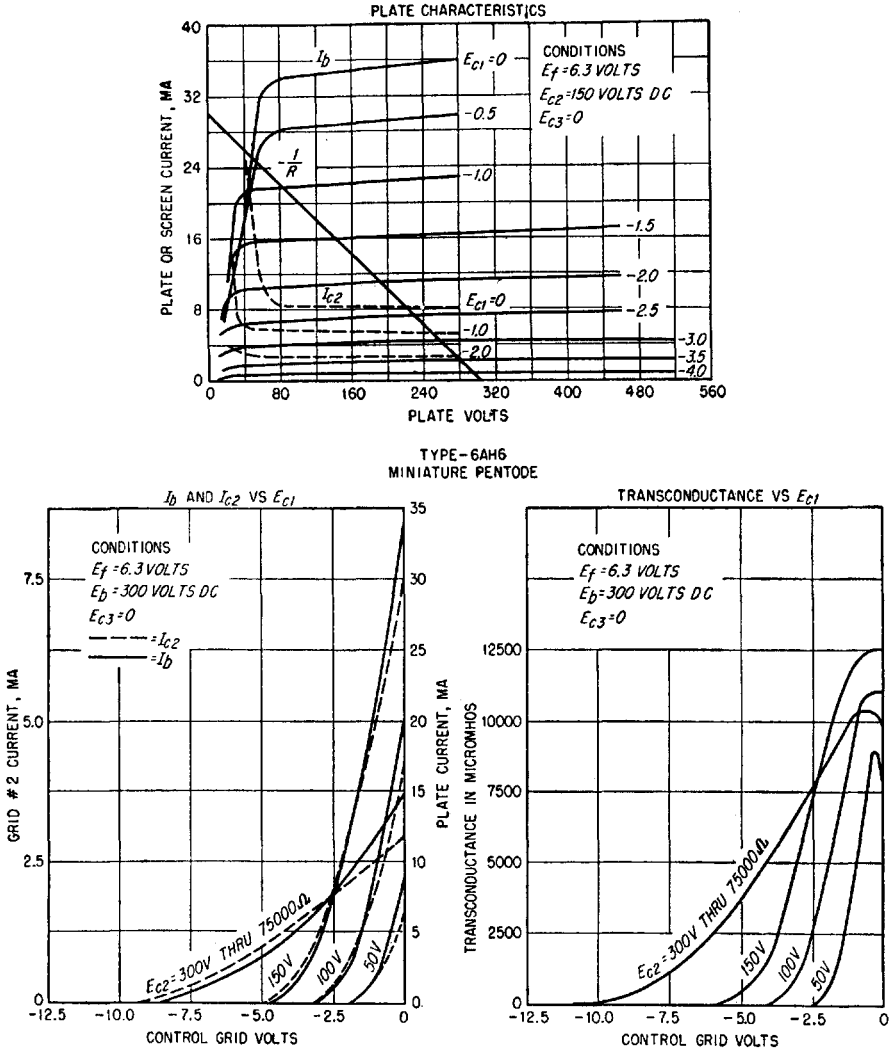
FIG. 7.94. Tube Characteristic Curves for Example 7.15.

$$C = \frac{1}{2\pi \times 10^4 \times 10^4}$$
$$= 1{,}590 \ \mu\mu f$$

4. Determine the approximate gain of the detector.

Neglecting the effect of the reactance of C at the modulation frequencies, the detector gain is given by

$$A = -g_{md}R$$

The detection transconductance is approximately equal to $1/\pi$ times the peak value of transconductance. For the unmodulated carrier, $g_{m(peak)}$ occurs at $E_c = -3.6 + 1.5$, or -2.1 volts.

$$g_{m(peak)} = 9{,}000 \ \mu mhos$$

Therefore

$$A = \frac{-9 \times 10^{-3} \times 10^4}{\pi}$$

$$= -28.6$$

The complete circuit of the detector is shown in Fig. 7.95.

2. Square-law Plate Detectors. The square-law or small-signal plate detector is very similar to the linear plate detector except that the tube bias is adjusted to the point of maximum curvature on the dynamic transfer characteristic in order to provide the maximum change in average plate current as the input signal amplitude varies with modulation. The out-
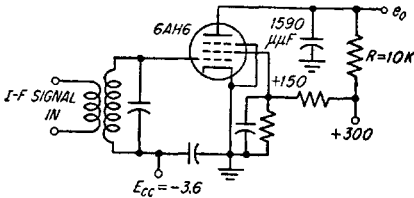
FIG. 7.95. Schematic of a linear plate detector.
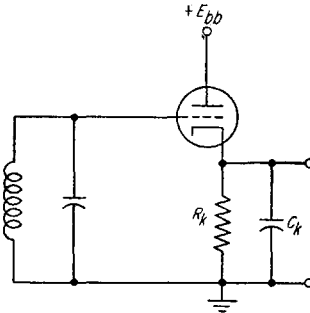
FIG. 7.96. Infinite-impedance detector.

put signal will contain components at the carrier frequency, the modulation frequency, the sum and difference frequencies of these two signals, and harmonics of these frequencies just as in the case of other non-linear detectors.

7.7e. Infinite-impedance Detector. The infinite-impedance detector consists of a cathode follower biased nearly to cutoff. The bias is usually obtained by the voltage developed across the cathode load resistor which is made quite large. The circuit is shown in Fig. 7.96. Since the tube is near cutoff for the condition of zero applied signal, plate current can increase during the positive half cycle of the applied signal, but is reduced to zero during the negative half cycle. This results in an increase in the average current through the tube which is a function of the amplitude of the applied signal. The time constant R_kC_k is made sufficiently large that the voltage across the capacitor changes only slightly from cycle to cycle of the applied signal and maximum rectification efficiency is obtained. The increased voltage across C_k due to the presence of an input signal carrier keeps the tube cut off for all but the peaks of the applied signal, and the resulting current and voltage waveforms are identical with that of the linear diode detector except that the current comes from the plate supply instead of the signal source. A
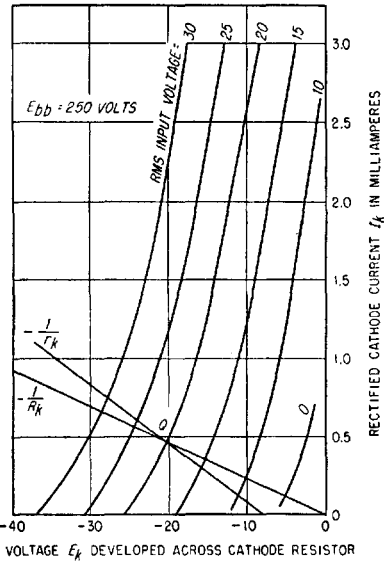
FIG. 7.97. Rectification diagram of typical infinite-impedance detector.

rectification diagram can be constructed for a tube operated as an infinite-impedance detector which is very similar to the rectification diagram of the diode detector.   A typical diagram is shown in Fig. 7.97.   The d-c and a-c load lines are constructed in a manner analogous to that for the diode detector.   However, since there is some current through R_k when no input signal is present, the $e_s = 0$ curve does not pass through the point $E_k = 0$, $I_k = 0$.   This allows the a-c load resistance to be appreciably less than the d-c load resistance without negative-peak clipping of 100 per cent amplitude-modulated signals.
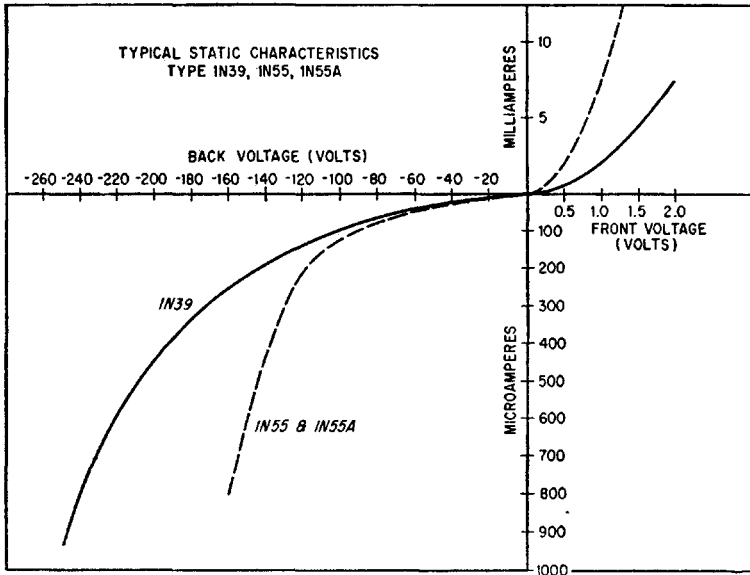
FIG. 7 98. Typical germanium diode characteristics.

The $g_m R_k$ product necessary to achieve a particular detector gain (always less than unity) is given by Eq. (7.235) provided $\omega_c R_k C_k \geq 100$.

$$g_m R_k = \frac{\pi}{\cos^{-1} A - \sqrt{1/A^2 - 1}} \tag{7.235}$$

where $A = \dfrac{\Delta E_k}{\Delta E_s} =$ detector gain $(A < 1)$

The detector gain can also be taken directly from the rectification diagram provided that the product $\omega_c R_k C_k$ is approximately 100 or more.

The infinite-impedance detector has the advantage of very high input impedance, better linearity than the plate detector, and a certain degree of freedom from clipping of the negative modulation peaks.   However, it has the disadvantages of less than unity gain and poorer linearity than the diode detector.

7.7f. *Crystal Diode Detectors.*   The design of linear crystal diode detectors is exactly the same as for linear diode detectors.   However, certain considerations must be taken into account in their use.   The crystal diode has the advantage of requiring no filament power, having no hum pickup between cathode and filament, having only about 0.5-$\mu\mu$f shunt capacitance, and having a lower forward resistance than vacuum diodes. The disadvantages of crystal diode detectors are smaller forward current and back voltage ratings than vacuum diodes, poorer linearity than vacuum diodes, and con-

siderable change in characteristics with variations in ambient temperature. Because
of the finite back resistance of a crystal diode, it should be used in applications where
the d-c load resistance is much smaller than the diode back resistance. If this con-
dition is not met, the detector efficiency will be reduced by current flow through the
diode during that portion of the cycle in which a thermionic diode would be cut off.
The characteristics of typical germanium detector crystals are shown in Fig. 7.98.

A special type of crystal diode detector is that which is used to detect very low level
signals in the "crystal video" type of receiver. This type of crystal is discussed in
Sec. 7.9.

7.7g. Considerations in Design of Detectors for Video and Pulse Applications. In
designing detectors for use in broadband circuits for pulse reception, the following
additional factors should be considered:
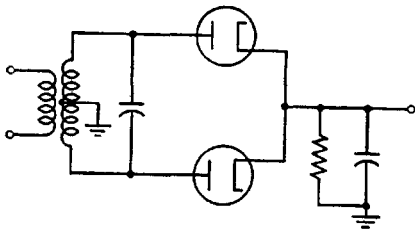
FIG. 7.99. Push-pull diode detector.
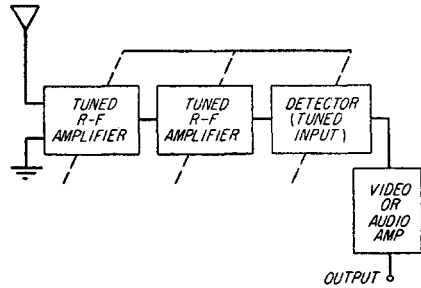
FIG. 7.100. Block diagram of a tuned r-f receiver.

1. Because the detector load resistance is usually low (a few thousand ohms),
the forward resistance of a diode detector should be as low as possible for maximum
detection efficiency.

2. The i-f signal applied to the input terminals of a diode detector is divided across
the capacitor divider formed by the diode shunt and the diode load capacitances.
To obtain the maximum i-f voltage across the diode, the diode shunt capacitances
should be much smaller than the diode load capacitance.

3. The time constant of the load impedance of a video detector is determined by the
allowable decay time of a detected pulse. The charging time constant of a diode
detector is less than the discharge time constant since the forward resistance of the
diode is less than the load resistance R.

4. A push-pull detector as shown in Fig. 7.99 provides a ripple voltage of twice the
i-f frequency across the diode load impedance. This permits easier filtering of the
ripple voltage in the detector load circuit without attenuation of the higher video
frequencies present in the received signals. The push-pull detector also provides a
more accurate reproduction of pulses because of the full-wave rectification. This
may be an important factor in certain applications, such as moving target indicator
radar systems.

5. Germanium crystal diodes have a considerably smaller forward resistance than
vacuum diodes, usually less than 100 ohms at 1 volt, and also have a low shunt
capacitance, about 0.5 $\mu\mu f$. This makes the crystal diode particularly useful in wide-
band applications where a small diode load resistance is required.

6. Video peaking circuits can be used in the detector load circuits just as in video
amplifier circuits.

7.8. Tuned Radio Frequency Receivers. A block diagram of a tuned radio fre-
quency (trf) receiver is shown in Fig. 7.100. It consists of several cascaded r-f
amplifier stages all tuned to the signal frequency followed by a detector and either an

audio or video amplifier.   The number of r-f stages depends upon the gain required to amplify the weakest signal to be received to a level suitable for detection.

The trf receiver is characterized by several inherent advantages and disadvantages as compared to the superheterodyne receiver.   These are:

Advantages

1. Simplicity of design
2. No-image frequency

Disadvantages

1. Poorer selectivity
2. Difficulty of tracking several r-f stages over any appreciable bandwidth

The r-f amplifiers used in the trf receiver are identical with the r-f or i-f amplifiers discussed in Sec. 7.4.   The untuned-primary tuned-secondary type of interstage coupling is most commonly used.   Double-tuned coupling circuits may be used, but the complication of achieving satisfactory tracking of the primary and secondary tuning of several cascaded stages is usually not justified based on the benefits of such coupling.
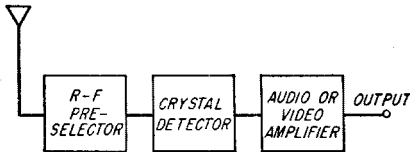
FIG. 7.101. Block diagram of a crystal video receiver.

The detector circuits for a trf receiver are identical with the detectors described in Sec. 7.7.   The detector is normally of the large signal or linear type, although a square-law detector can be used if little r-f amplification is desired.

7.9. Crystal Video Receivers.   The block diagram of a crystal video receiver is shown in Fig. 7.101.   It generally has no r-f amplification ahead of the crystal detector, although selectivity is sometimes achieved by employing bandpass filter networks ahead of the crystal.   It is the special case of the trf receiver where the r-f amplification has been omitted.   The detector is a small-signal or square-law detector, and all of the amplification is obtained by audio or video amplifiers following the detector.

The crystal video receiver is characterized by the following advantages and disadvantages:

1. Broad r-f bandwidth (unless filter networks are employed)
2. No local oscillator
3. No tuned amplifiers
4. Poor sensitivity compared to a superheterodyne or a trf receiver
5. High video or audio amplification required
6. Square-law detector

The crystal detectors used in a crystal video receiver are very similar to those discussed in Sec. 7.7f.   The semiconductor material is usually silicon because of its greater nonlinearity for small signals.   Because of the curvature of the E-I characteristics, the rectified crystal current is proportional to the square of the applied voltage.   The rectified current I_{dc} in microamperes can be expressed as

$$I_{dc} = BP \qquad (7.236)$$

where P = input power, μw

B = a constant of crystal diode and is a measure of current sensitivity of crystal, μa/μw

Equivalent output circuits of a crystal detector are shown in Fig. 7.102. The crystal can be considered as a voltage source E_{dc} having an internal resistance r_d such that

$$E_{dc} = I_{dc}r_d = r_d BP \qquad (7.237)$$

Crystal Detector Sensitivity and Figure of Merit. The noise voltage developed by the internal resistance of the diode is given by Eq. (7.238) provided that no direct current flows through the crystal.

$$E_n = \sqrt{4KT\,\Delta f r_d} \qquad (7.238)$$

where Δf = the video bandwidth

Any direct current through the crystal increases the noise temperature factor t_c above unity as discussed in Sec. 7.5d. This is illustrated in Fig. 7.63. Therefore the crystal should be capacitively coupled to the input of the video amplifier following the crystal. At low frequencies, the noise generated within the first amplifier stage following the detector can be considered as having originated in an equivalent resistance R_{eq} in series with the tube grid (see Sec. 7.2h). The total noise of the crystal and first amplifier tube referred to the crystal output circuit is

(a) EQUIVALENT CURRENT GENERATOR CIRCUIT
(b) EQUIVALENT VOLTAGE GENERATOR CIRCUIT

FIG. 7.102. Equivalent output circuits of a square-law crystal detector.

$$E_n = \sqrt{4KT\,\Delta f(r_d + R_{eq})} \qquad (7.239)$$

This results in a signal-to-noise ratio at this point of

$$\frac{S}{N} = \frac{P}{\sqrt{4KT\,\Delta f}}\left(\frac{r_d B}{\sqrt{r_d + R_{eq}}}\right) \qquad (7.240)$$

The term $r_d B/\sqrt{r_d + R_{eq}}$ is a function of the crystal diode and tube only and is known as the figure of merit (FM) of the crystal. The JAN committee on electron tubes has standardized on an R_{eq} of 1,200 ohms in determining the detector crystal figure of merit.

The maximum signal-to-noise ratio is obtained when the crystal impedance r_d is a maximum. Any loading of the crystal output is undesirable, although it may be necessary in order to achieve the required bandwidth. Typical values of r_d are between 4,000 and 24,000 ohms.

For unity signal-to-noise ratio at the crystal output, the required input power P in watts is

$$P = \frac{\sqrt{4KT\,\Delta f}}{FM} \qquad (7.241)$$

TABLE 7.9. CHARACTERISTICS OF TWO VIDEO DETECTOR CRYSTALS

Characteristics	Crystal type	
	1N31	1N32
Frequency	9,000 Mc	3,000 Mc
Video impedance, ohms	6,000–23,000	5,000–20,000
Figure of merit	55	100
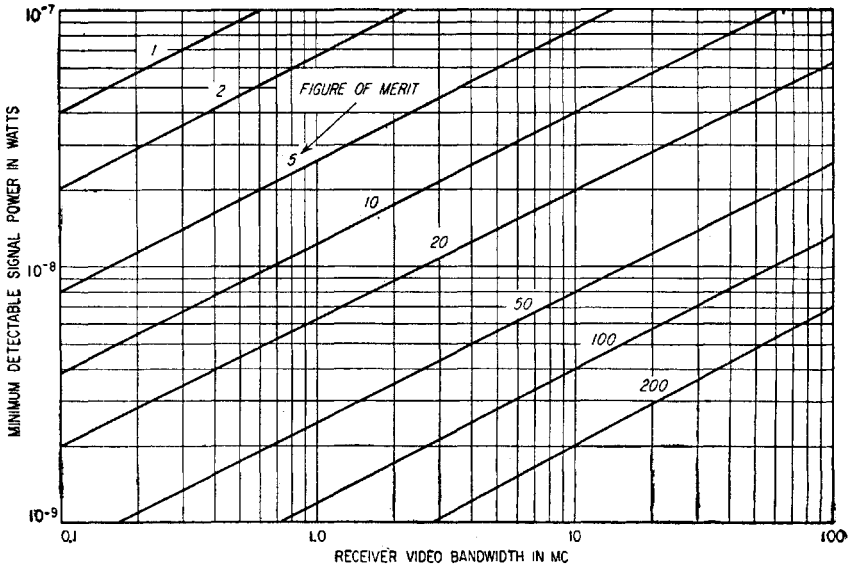
Fɪɢ. 7.103. Minimum detectable signal power P versus bandwidth for various values of crystal figure of merit.

Equation (7.241) is plotted in Fig. 7.103 where the minimum detectable signal power is plotted as a function of video bandwidth Δf for various values of FM. The figures of merit for two JAN-type crystal detectors are given in Table 7.9.

Direct current flowing through the crystal greatly increases the noise output of the crystal at low frequencies. Therefore the crystal should be capacitively coupled to the next stage as shown in Fig. 7.104 if the tube bias is such that grid current could flow.

R-F Impedance of Crystal Detector. The equivalent input circuit of a crystal diode for small signals is shown in Fig. 7.105. The resistance R_b of the crystal barrier, i.e., the resistance at the point of contact between the cat-whisker and the semi-conductor, is nonlinear and causes
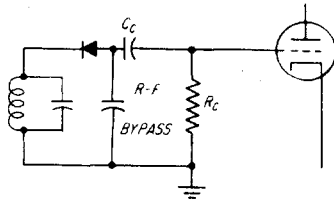
Fɪɢ. 7.104. Method of coupling video crystal to amplifier input.
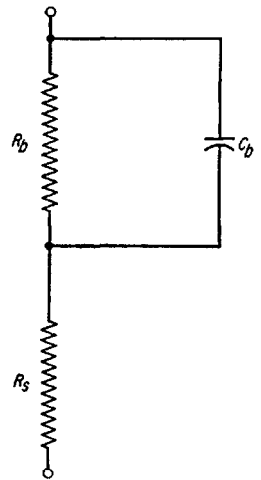
Fɪɢ. 7.105. Equivalent circuit of a crystal diode.

rectification of low-level signals as discussed in Sec. 7.7*b*. The barrier capacitance C_b shunts the barrier resistance and at very high frequencies reduces the rectification efficiency. The series resistance R_s is called the spreading resistance and results from the constriction of current in the semiconductor near the point contact. The spread-

ing resistance is constant and is typically 10 to 30 ohms for silicon diodes. It prevents C_b from being tuned out by an external reactance.

7.10. Superregenerative Receivers. A superregenerative receiver is an r-f amplifier or plate detector having sufficient positive feedback to cause oscillation. The receiver is caused to go in and out of oscillation by a control signal known as the quench signal. Typical quench-signal frequencies are between 10 kc and 1 Mc. Very high gains are possible. A one-tube circuit is capable of detecting the noise voltage existing at the tuned circuit input. The three possible modes of operation are (1) separate quenching, logarithmic mode, (2) separate quenching, linear mode, and (3) self-quenched mode.

7.10a. Logarithmic Mode. A schematic of a superregenerative detector is shown in Fig. 7.106. The grid circuit is resonant at the desired signal frequency and is coupled
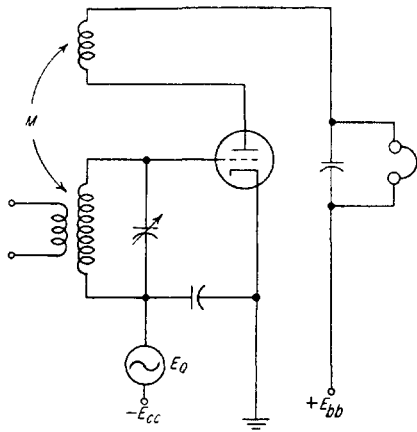
FIG. 7.106. Circuit of superregenerative detector.
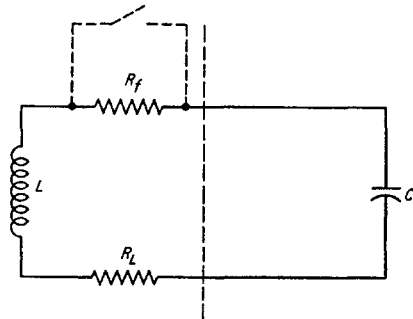
FIG. 7.107. Equivalent superregenerative detector resonant circuit.

to the antenna or signal source as shown. The mutual coupling between the plate coil and grid circuit is sufficient to allow oscillations to build up when the grid bias is raised above the cutoff value. The fixed grid bias E_{cc} is made greater than cutoff, and a control voltage is superimposed on the grid bias which periodically decreases the bias thereby permitting oscillations to start building up in the circuit. When the quench signal returns the grid bias sufficiently negative to stop the tube from conducting, the oscillations in the grid-tuned circuit decay exponentially. The equivalent grid circuit of the superregenerative detector is shown in Fig. 7.107. In the absence of the quench voltage, the circuit consists of C in parallel with L having some positive series resistance R_L. During the interval that the quench voltage causes the tube to conduct, the mutual coupling between the plate and grid circuits effectively introduces negative resistance R_f in series with L. The instantaneous voltage e_c across C is given by

$$e_c = E e^{-Rt/2L} \cos \omega_o t \qquad (7.242)$$

where E = voltage across C at time $t = 0$

R = net circuit resistance, $R = R_L - R_f$

ω_o = resonant frequency of circuit = $\sqrt{\dfrac{1}{LC} - \left(\dfrac{R}{2L}\right)^2}$

At all times there is a voltage present across C due to thermal noise, atmospheric noise, etc. Without the presence of the quench voltage, these voltages exponentially

decay at a rate determined by the Q of the circuit. When the quench voltage causes the tube to conduct, however, the feedback resistance R_f exceeds R_L in magnitude and the exponent in Eq. (7.242) becomes positive. As a result, the voltage across C exponentially increases from its value at the instant the quench voltage is applied until a state of equilibrium is reached. The frequency of the exponentially increasing oscillation is ω_o. The amplitude of the envelope of the oscillation in the interval before grid and/or plate current saturation is reached is given by

$$e_c = E_n e^{-Rt/2L} \qquad (7.243)$$

where E_n = noise voltage across C at $t = 0$

When the amplitude of the oscillations becomes large enough to drive the grid into the regions of grid conduction and cutoff during the positive and negative peaks of the oscillation, the amplification of the tube is reduced until the magnitude of R_f equals R_L and the amplitude of the oscillations remains constant. If a signal voltage E_s is present across C at the instant the quench voltage is applied, the envelope of the oscillation is given by
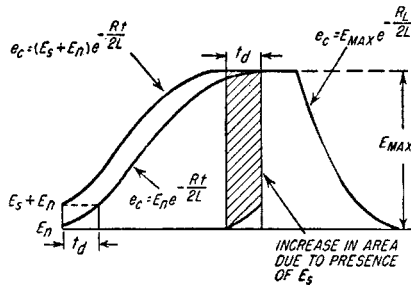
FIG. 7.108. Envelope of oscillations in logarithmic mode.

$$e_c = (E_s + E_n)e^{-Rt/2L} \qquad (7.244)$$

and the condition of saturation is reached sooner than when noise only is present. This is illustrated in Fig. 7.108. The envelope of oscillation starting from the level $E_s + E_n$ reaches saturation amplitude t_d seconds before the envelope of the oscillation starting from E_n. The difference in area ΔA under the two curves is given very closely by

$$\Delta A = t_d E_{\max} \qquad (7.245)$$

where $E_{\max}$ = amplitude of oscillation when equilibrium has been attained (see Fig. 7.108)

The time interval t_d is given by

$$t_d = -\frac{2L}{R} \log_e \frac{E_s + E_n}{E_n} \qquad (7.246)$$

The increase ΔE in the average voltage across C due to the presence of a signal is

$$\Delta E = -\frac{2L f_Q E_{\max}}{R} \log_e \frac{E_s + E_n}{E_n} \qquad (7.247)$$

where f_Q = frequency of applied quench voltage

The quantity $f_Q E_{\max} 2L/R$ can be made equal to several volts. Therefore, even though E_s may be only a few microvolts, if the noise voltage is the same order of magnitude, the gain of the stage will be in the order of a million.

  1. Considerations for Maximizing the Gain in Logarithmic Mode Operation. The gain of a superregenerative detector operating in the logarithmic mode can be maximized by making the terms $E_{\max}$, $2L/R$, and f_Q as large as possible. However, the terms are interdependent and must be considered simultaneously. General considerations are:

  a. As the magnitude of $2L/R$ is increased, the rate of increase of the oscillations is reduced and the period for which the quench voltage must be applied is increased.

b. As E_{max} is increased, the interval required for the oscillation buildup to the level of E_{max} and the interval required for the oscillation to decay to a level lower than the originating signal are increased.

c. The higher the Q of the resonant circuit itself, i.e., the higher $\omega_o L / R_L$, the longer the decay interval.

(a) VARIATION OF μ WITH E_c FOR TYPICAL TRIODE

(b) SINUSOIDAL QUENCH VOLTAGE   (c) RECTANGULAR QUENCH VOLTAGE

(d) VARIATION IN NET CIRCUIT RESISTANCE WITH QUENCH VOLTAGE

(e) OSCILLATION ENVELOPES FOR SINUSOIDAL AND RECTANGULAR QUENCH VOLTAGES
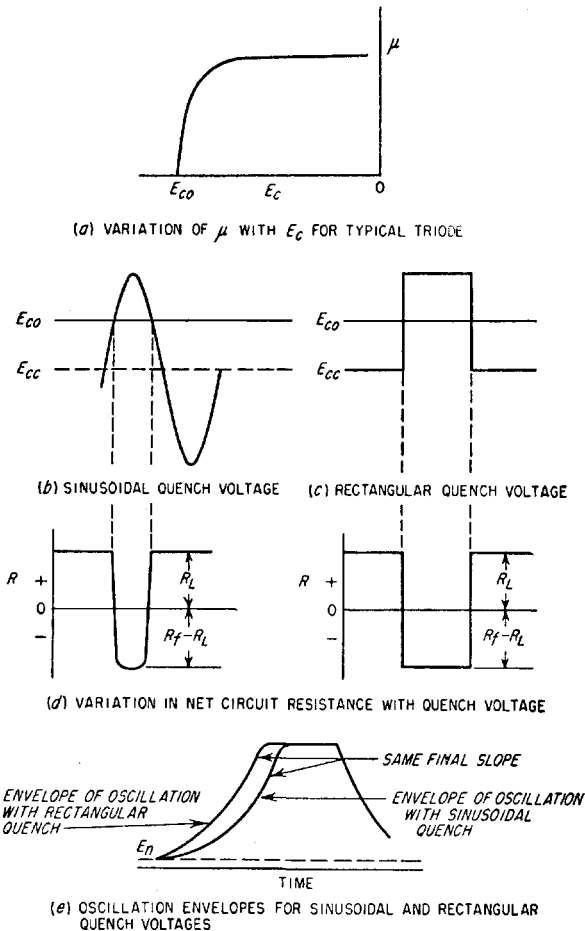
FIG. 7.109. Effect of quench-voltage waveshape upon envelope of oscillation buildup.

d. The longer the buildup and decay periods, the lower must be the maximum value of the quench frequency.

e. The quench frequency f_Q must be enough higher than the desired signal modulation frequencies to allow effective filtering of f_Q from the output, and in any case f_Q must be at least twice the highest modulation frequency to be detected.

f. The gain can be increased by reducing R [see Eq. (7.247)]. This can be accomplished by either decreasing the amount of feedback or by reducing the Q of the resonant circuit itself. The net effective circuit resistance R must be maintained negative, however, in order for the circuit to oscillate.

2. Quench-voltage Waveform. The waveform of the quench voltage is usually rectangular or sinusoidal. The gain with a sinusoidal quench voltage is slightly

greater than with a rectangular quench voltage because of the fact that the net circuit resistance R requires a longer period to change from its initial value of R_L to the final value $R_L - R_f$ if the quench voltage has a slow rate of rise.   A low negative value of R causes the oscillation to build up slowly so that the time interval t_d [see Eq. (7.246)] in Fig. 7.108 will be greater for sinusoidal quench than for rectangular quench.   This effect is illustrated in Fig. 7.109.

3. Modulation Reproduction.   The response of a superregenerative receiver operating in the logarithmic mode to a modulated carrier is illustrated in Fig. 7.110.   For
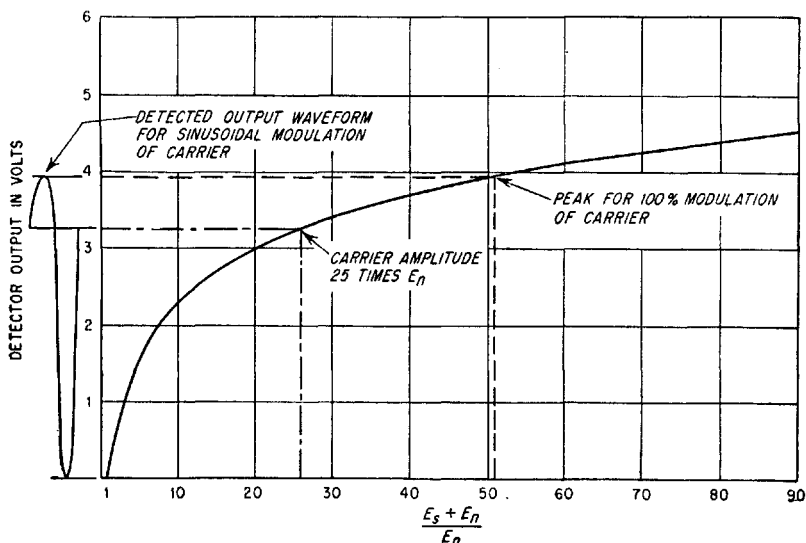
FIG. 7.110. Typical detector output characteristic for logarithmic mode.   The figure illustrates the distortion in the output for a 100 per cent modulated carrier for the case where the unmodulated carrier is 25 times greater than the noise voltage.

small percentages of modulation, the distortion is only slight.   However, as the percentage modulation approaches 100 per cent, an effective "volume expansion" is experienced and the distortion becomes quite severe.

The logarithmic response also provides a type of automatic volume control since the modulation of a high-amplitude carrier is compressed compared to that of a lower-amplitude carrier.   This causes the detected modulation from carriers which have a constant small-percentage modulation to be nearly equal regardless of the carrier amplitude.

7.10b. Linear Mode.   The linear mode of operation is shown in Fig. 7.111.   The length of the period during which the quench voltage is applied is reduced so that the tube does not reach equilibrium before the quench voltage is removed.   The average value of grid voltage during one quench cycle is greater when a signal is present than when noise alone is present since the envelope of oscillations increases exponentially from a higher initial voltage.   The envelope of the oscillations has, therefore, increased to a higher voltage when the quench voltage is removed and decays to zero from this higher voltage.   If it is assumed that the net circuit resistance R is a constant for the interval that the quench voltage causes the tube to conduct, the voltage gain A for the linear mode can be determined from

$$A = \frac{\Delta E}{E_s} = f_Q \left[\left(-\frac{2L}{R} + \frac{2L}{R_L} \right) e^{-Rt_1/2L} + \frac{2L}{R} - \frac{2L}{R_L} e^{[-Rt_1 + R_L(t_1 - t_2)]/2L} \right] \quad (7.248)$$

where t_1 = time interval during which quench voltage is applied

$t_2 = 1/f_Q$

ΔE = change in average voltage across C due to presence of E_s

Maximum gain is obtained in the linear mode by having a high value of $-R/2L$, that is, large positive feedback, and a high quench frequency. The interval during
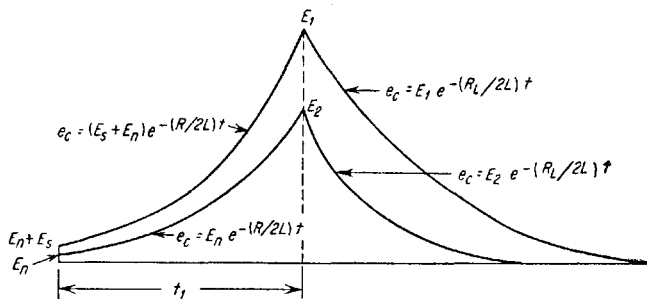
FIG. 7.111. Envelope of oscillations in the linear mode.

which the quench voltage is applied must be short enough so that the oscillation does not reach equilibrium when the largest amplitude signal to be amplified without distortion is received. The interval between applications of the positive quench voltage must be of sufficient duration to allow the envelope of oscillation to decay to a level below that of the original signal. The minimum interval between applications of the positive quench voltage is given by

$$\frac{1}{f_{Q(\text{max})}} = -\frac{R}{R_L} t_1 + \frac{2L}{R_L} \log_e \frac{E_s}{E_o} \qquad (7.249)$$

where t_1 = period of application of quench voltage

$\dfrac{E_s}{E_o}$ = ratio of original signal amplitude to amplitude of decayed oscillations

As a factor of safety, the ratio E_s/E_o is usually made equal to 5 or 10. Because of the above factors, the interval during which the quench voltage is applied is normally made considerably smaller than the interval between applications of quench voltage.

Since the gain in the linear mode is a constant and is independent of the input signal amplitude, the linear mode super-regenerative detector reproduces modulation with a minimum of distortion.

7.10c. Self-quenched Detector. If a superregenerative detector is connected as shown in Fig. 7.112, it will operate in a self-quenched fashion. Oscillations initially build up from the level of noise

FIG. 7.112. Circuit of a self-quenched super-regenerative detector.

within the tuned circuit until the grid is driven into the positive grid region during the peaks of the oscillations. Blocking capacitor C is then charged with the polarity shown in Fig. 7.112 by the flow of grid current during the positive peaks of the oscillations.

R is made large enough so that the bias voltage created by the current flowing through R in discharging C is great enough that the tube is cut off for such a large portion of each oscillation cycle that it cannot overcome the circuit losses during the

rest of the cycle.   After building up exponentially from noise to the level where grid current flows during the peaks of oscillation, the oscillation decays rather gradually until the instantaneous grid bias is always greater than cutoff for the tube. At this point, the oscillations decay exponentially.   The value of C must be large enough that the time constant RC is sufficient to prevent the bias voltage from decreasing appreciably before the oscillations decay to the point where the tube no longer conducts.   Provided that R and C are large enough to cause intermittent oscillation, the exact value of RC is determined by the desired quench frequency. The grid-voltage waveform for self-quenched operation is shown in Fig. 7.113.

When a signal is present it adds to the bias voltage across R and causes the oscillations to start sooner than would have been the case without the signal.   The envelope
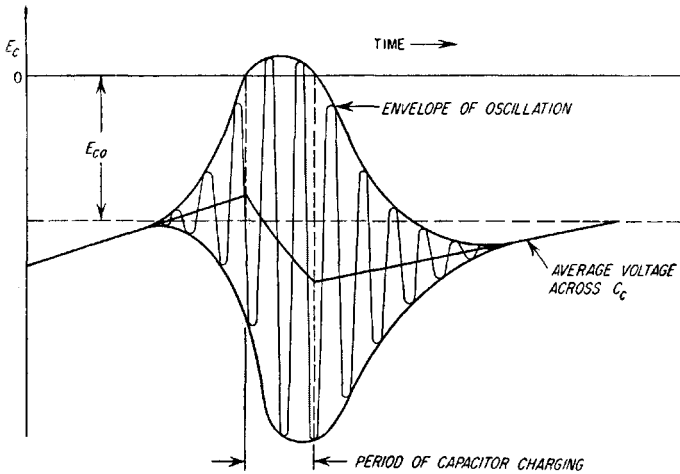
FIG. 7.113. Grid-voltage waveform of a self-quenched superregenerative detector.

of the oscillation is the same each time regardless of the presence of a signal, so that the result of a signal is to increase the average current through the tube by decreasing the period between oscillations, i.e., the quench frequency is increased.

In the self-quenched detector, the RC time constant should be adjusted to provide the highest possible quench frequency and still allow the oscillations to decay to a sufficiently low level before the initiation of another train of oscillations.   The minimum period between oscillations is given by Eq. (7.249).

7.10d. Selectivity.   The superregenerative detector is normally not preceded by any r-f amplifier stages; therefore, the selectivity of such a detector is dependent upon the response of a single resonant circuit.   The selectivity or adjacent band rejection of a single resonant circuit is considerably poorer than the corresponding selectivity of several cascaded tuned circuits having the same over-all −3-db bandwidth (see Fig. 7.13).

However, the effective Q of the detector resonant circuit is considerably higher than $\omega L/R_L$.   As the quench voltage is applied and the grid bias reaches cutoff, the tube begins to amplify, but the grid bias must increase to a value somewhat above cutoff before the tube μ is high enough to cause the net circuit resistance to become negative and for oscillations to start.   During the interval when the grid bias is between cutoff and the value where oscillations start, the net circuit resistance is positive but decreasing from R_L to O and the effective Q of the resonant circuit is increasing.   This regenerative period increases the Q of the circuit several times.

7.10e. Noise in Superregenerative Detectors. The superregenerative detector detects the presence of a carrier and the modulation of that carrier by "sampling" the signal at the quench frequency rate. The result is that the detected modulation of a carrier varies in accordance with the actual modulation of the carrier as shown in Fig. 7.114. The highest possible detection frequency is $f_Q/2$, and all higher modulation frequencies are converted to some frequency between zero and $f_Q/2$.

The noise spectrum of the superregenerative-detector input circuit contains all frequencies within the passband of the detector resonant circuit in proportions determined by the Q of the circuit. The noise voltage variations in the detector resonant circuit would produce noise output from a linear detector having components from zero to the noise bandwidth of the resonant circuit. However, because of the sampling process, these variations are all reduced to detected frequency variations between zero
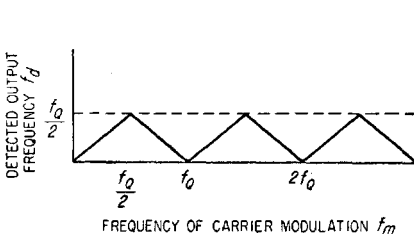
FIG. 7.114. Detected modulation frequency versus carrier modulation frequency for a superregenerative detector.
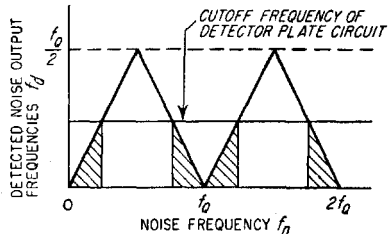
FIG. 7.115. Noise output from a superregenerative detector.

and $f_Q/2$ in this type of detector. This is illustrated in Fig. 7.115. Consequently, the noise output of the detector in the region of zero to $f_Q/2$ is considerably higher than for other detectors. The noise voltage per unit bandwidth at the detector output is inversely proportional to the Q of the resonant circuit and the quench frequency.

7.10f. Pulse Reception. The superregenerative receiver can be used for pulse reception by providing adequate bandwidth in the r-f circuit and by raising the quench frequency so that one or more oscillation buildups occur within the duration of the shortest pulse to be received.

The discharge time constant of the detector circuit must also be minimized to provide adequate resolution between successive pulses.

7.11. Frequency-modulation Receivers. In frequency modulation, information is transmitted by varying the frequency of a constant amplitude carrier in accordance with the intelligence to be transmitted (see Sec. 5.5). Detection of a frequency-modulated signal consists of the conversion of the frequency variations of the carrier to amplitude variations. Although any circuit which provides signal attenuation which varies as a function of frequency can be used as an f-m detector when followed by a conventional amplitude detector, the most frequently used types of f-m detectors are the Foster-Seeley discriminator, the gated-beam discriminator, and the ratio detector.[1] The Foster-Seeley discriminator requires limiting of the amplitude of the received carrier prior to detection in order to reject incidental amplitude modulation from the output. The gated-beam discriminator utilizing the 6BN6 tube provides limiting of the signal amplitude within the tube itself. The ratio detector circuit also provides good rejection of any amplitude modulation.

Frequency-modulation receivers are identical to other receiver types designed for amplitude modulation except for the detector circuit. Superheterodyne receivers are

[1] The first two of the above detectors are discussed in Sec. 7.6e. For details of the ratio detector see S. W. Seeley, The Ratio Detector, *RCA Rev.*, vol. 8, pp. 201–236, June, 1947.

used almost exclusively in reception of commercial broadcast f-m transmissions. However, other receiver types can be used in special applications.

7.12. Receiver Measurements. Several laboratory measurements are frequently made in the design and testing of receivers. The most important of these are:

1. Bandwidth, measurement and alignment
2. Skirt selectivity and image rejection
3. Gain
4. Noise figure

7.12a. Bandwidth, Measurement and Alignment. The bandwidth of a receiver is normally taken as the frequency band between the frequencies where the receiver gain has dropped 3 db from the midband value. Bandwidth measurements are most conveniently made with a sweep frequency generator, marker generator, detector, and
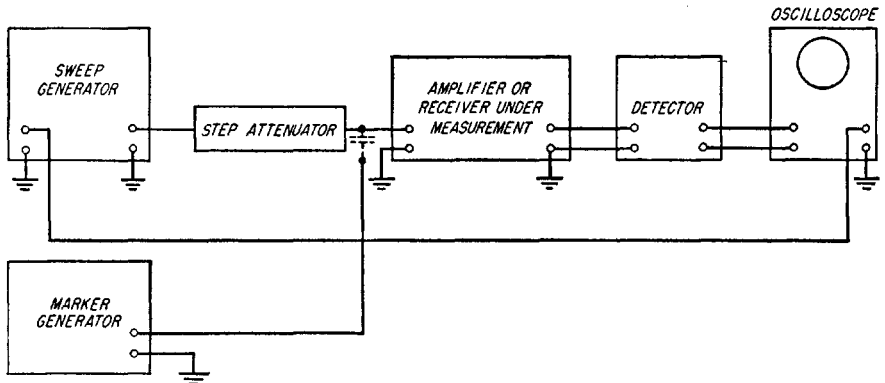
Fig. 7.116. Block diagram of test setup for bandwidth measurements using a sweep frequency generator.

oscilloscope. A sweep frequency generator is an r-f or i-f oscillator whose frequency is mechanically or electrically varied over a band of frequencies somewhat wider than the receiver bandwidth and which has an adjustable center frequency. The power output from the generator is maintained as nearly constant as possible over the swept frequency band. The test setup is illustrated in Fig. 7.116.

The output impedance of the step attenuator must be equal to the source impedance for which the receiver or amplifier is designed, or an impedance matching network must be used between the attenuator and the receiver or amplifier input. If a detector other than that of the receiver under test is used, the input impedance of the detector should be correct for the stage to which it is connected. The marker generator is simply a calibrated c-w signal generator covering the desired frequency range. It is either capacitively or inductively coupled to the amplifier under test as lightly as possible and still have its signal observable on the oscilloscope. Either the horizontal sweep is synchronized with the frequency sweep of the sweep generator, or, if available, a sweep voltage from the generator is amplified by the oscilloscope horizontal amplifier to provide the horizontal sweep. A typical oscilloscope waveform is shown in Fig. 7.117a. The marker generator will produce a small pip or disturbance on the passband waveform at the spot corresponding to the frequency to which the marker generator is tuned due to the beat between the sweep generator signal and the marker signal when the two signals are at nearly the same frequency. The -3-db bandwidth of the receiver is most accurately determined by the method illustrated in Fig. 7.117b. The marker pip is set to coincide to the same vertical

positions on the normal passband waveform that the top of the passband waveform come to when 3 db of attenuation is introduced by the step attenuator. The −3-db frequencies can then be read from the marker generator calibrated dial.

The sweep frequency generator is extremely useful in the alignment of a receiver or amplifier since it gives a picture of the entire passband. The individual tuned circuits can be tuned to the proper design frequency by using a grid-dip oscillator. A grid-dip oscillator is a tunable oscillator having a calibrated frequency dial and containing a d-c milliammeter in the grid circuit. If the coil of the grid-dip oscillator resonant circuit is closely coupled to the resonant circuit whose frequency is to be measured, the oscillator grid current will dip sharply when the oscillator is tuned to

(a) OSCILLOSCOPE WAVEFORM
(b) METHOD OF ACCURATELY DETERMINING RECEIVER BANDWIDTH
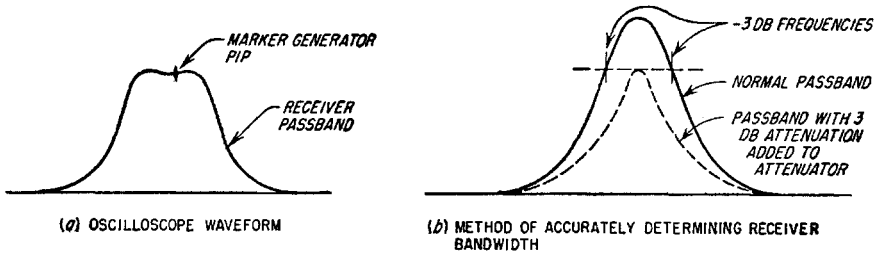
FIG. 7.117. Oscilloscope bandpass waveforms.
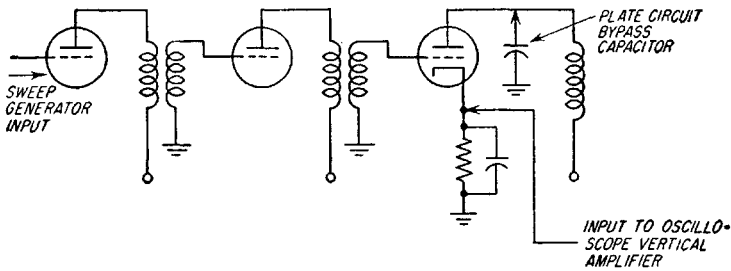
FIG. 7.118. Method of detecting passband when external detector cannot be used.

the resonant frequency of the other circuit. The over-all passband can then be observed and trimmed up, using the sweep frequency generator.

When the passband of only a portion of an amplifier or receiver is being observed, it is often inconvenient to couple in an external detector. This difficulty can be overcome by the method shown in Fig. 7.118. The oscilloscope input is connected directly to the bypassed cathode resistor of the stage following that portion of the amplifier or receiver under test. Because of nonlinearities in the tube characteristics, the tube will act as a detector and a low-level detected output will appear across the cathode resistor. The plate circuit should be bypassed to ground when this technique is utilized to prevent the plate circuit impedance from affecting the passband waveform appearing at the cathode.

7.12b. Skirt Selectivity and Image Rejection. The skirt selectivity of a receiver passband and the amount of image rejection provided by an r-f amplifier are most easily measured using a calibrated c-w signal generator and a step attenuator. The test setup is shown in Fig. 7.119. The signal generator is initially set at the center of the receiver passband, and, with all or nearly all of the attenuation of the step attenuator in, the receiver gain is adjusted to give a convenient d-c voltage level at the detector output. This d-c output should be considerably above the d-c output of the detector

due to noise, but below limit level in the receiver.   The frequency of the c-w generator is then changed in steps, and, maintaining a constant output from the generator, the attenuation of the attenuator is reduced an amount just sufficient to maintain the detector d-c output level constant.   The attenuation at each frequency relative to the attenuation at midband is recorded and plotted versus frequency to obtain a graph of skirt selectivity and/or image rejection.

7.12c. Gain.   The gain of a receiver is usually measured in terms of voltage gain, i.e., the ratio of the peak voltage across the receiver output terminals with its load connected to the peak voltage across the receiver input terminals.   Since the receiver normally includes a detector, this will be the ratio of the peak modulation voltage across the receiver load impedance to the peak value of the input carrier.   Thus
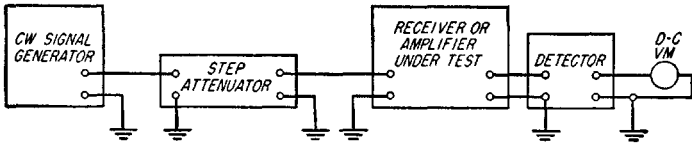
Fig. 7.119. Block diagram of test setup for measuring skirt selectivity and image rejection of a receiver.
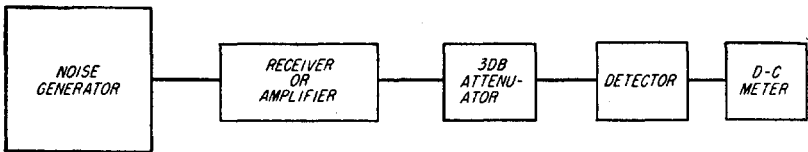
Fig. 7.120. Test procedure for measuring receiver noise figure utilizing a noise generator.

defined, it is a function of the percentage modulation of the carrier.   In receivers designed for pulse reception, the gain of the receiver is usually taken to be the ratio of the peak amplitude of the voltage pulse across the receiver load impedance to the peak value of the r-f pulse voltage across the receiver input terminals.

7.12d. Noise Figure.   The noise figure F of a receiver is given by Eq. (7.51) (see Sec. 7.2j).   The noise figure is a measure of the noise contributed to the receiver output by the receiver itself.   The noise figure of a receiver, or any portion of a receiver, can be most accurately measured by utilizing a noise generator which produces an accurately known amount of noise throughout the r-f passband of the receiver or amplifier under test.   The test setup is illustrated in Fig. 7.120.   Since the noise figure of the receiver is a function of the source impedance, the noise generator must present the same impedance to the receiver input terminals as exists in use.

At frequencies up to 1,000 Mc, temperature-limited vacuum diodes are used as noise generators.   The mean-square noise current component of diode plate current when temperature-limited is given by Eq. (7.13).   The circuit of a typical diode noise generator is shown in Fig. 7.121.   The available noise power from the generator is that noise power which the generator could deliver to a resistance equal to the output resistance R_a of the generator.   This is given by

$$P_n = \frac{eI_b \, \Delta f R_a}{2} \qquad (7.250)$$

The procedure for measuring noise figure with this type of noise generator is as follows.   When the noise generator is connected to the receiver as shown in Fig. 7.120, but with zero diode current, the d-c voltage at the detector output is observed.

An attenuation of 3 db is then added after the receiver but ahead of the detector as indicated in Fig. 7.120. The noise diode current is then increased until the d-c voltage at the detector output is equal to the value previously obtained with zero diode current. For this value of diode d-c current, the noise power at the output of the receiver has been doubled, indicating that the noise power delivered from the generator has produced an output noise power equal to that present at the receiver output in the absence of the noise generator. Therefore, considering the signal output power S_0 of Eq. (7.51) to be the noise output power contributed by the noise generator, the denominator of Eq. (7.51) is unity. The noise figure is then given by

$$F = \frac{S_i}{N_i} = \frac{eI_b \,\Delta f R_a}{2KT \,\Delta f} \qquad (7.251)$$
$$\simeq 20 \, I_b R_a$$

The noise figure is thus obtained by a method whereby the bandwidth of the receiver does not enter into the measurement. This is particularly advantageous because the −3-db bandwidth of a receiver and the effective noise bandwidth are not necessarily equal. The 3-db attenuator shown in Fig. 7.120 is desirable because it provides a method of measurement independent of the detector characteristic. If the attenuator is not used, the method of measurement is identical except that the actual noise output power is doubled by the noise generator. To determine this accurately, the detector must be precalibrated by a calibrated signal generator, or, alternately, the doubling of the output noise power can be measured by an rms-

FIG. 7.121. Temperature-limited diode noise generator.

reading r-f voltmeter connected at the detector input. In this case, the noise generator output is adjusted until the r-f voltmeter reading increases to 1.414 times the reading with zero diode current.

In the diode noise generator described, the input terminals of the receiver should be connected directly to the noise generator output terminals if possible and the stray shunt capacitances introduced tuned out by the inductance L of Fig. 7.121. If a coax section is employed, it should be appreciably less than $\lambda/10$ where λ is the wavelength of the signal to be received. The inductance L (see Fig. 7.121) should be adjusted to tune out the effect of the capacitance introduced by the coax section.

In the measurement of the noise figure of a superheterodyne with a noise generator, the image response of the receiver, if significant, must be taken into consideration. If the image response is not effectively eliminated by r-f preselection and if the noise generator has an output at the receiver image frequency, the total noise output from the receiver due to the noise generator will be due to the sum of the noise powers contributed by the generator at the desired and the image frequencies. Since an actual signal will be present only at the desired frequency, the result obtained will be incorrect. If the noise generator has equal noise-power output at the desired and image frequencies and if there is no r-f preselection, the true receiver noise figure will be 3 db greater than measured by the noise-generator technique.

When the effective noise bandwidth of the receiver under test is known, the noise figure of the receiver can be determined readily by a calibrated c-w signal generator having an output impedance equal to the source impedance required by the receiver. With the signal generator connected but turned off, the d-c voltage at the output of the detector in the receiver is measured. The signal generator is then turned on and

its output increased until the d-c output from the detector is 1.45 times the value with the signal generator turned off. When this condition exists, assuming a linear detector, the signal power into the detector is equal to the noise power. The receiver noise figure is then given by

$$F = \frac{S_i}{KT\,\Delta f} \qquad (7.252)$$

where S_i = signal power output from generator

7.13. Practical Considerations in the Design of a Receiver. In addition to the basic design of the various circuits in a receiver, there are a number of practical considerations which must be taken into account by the designer.

7.13a. Plate Circuit Decoupling. In the design of the r-f and i-f amplifier portions of receivers, sufficient decoupling must be employed between the plate returns of successive stages to prevent amplifier instability due to positive feedback through the common plate supply. The feedback path is shown in Fig. 7.122. At the resonant frequency of the tuned amplifiers, the plate supply will usually present a relatively high inductive reactance due to lead lengths. Although the plate bypass capacitors will have low reactances at the operating frequency of the amplifier, the attenuation through the feedback path shown in Fig. 7.122 will, in general, not be sufficient unless the decoupling inductance L_d is inserted

Fig. 7.122. Feedback path through plate supply.

(a) SERIES DECOUPLING   (b) PARALLEL DECOUPLING
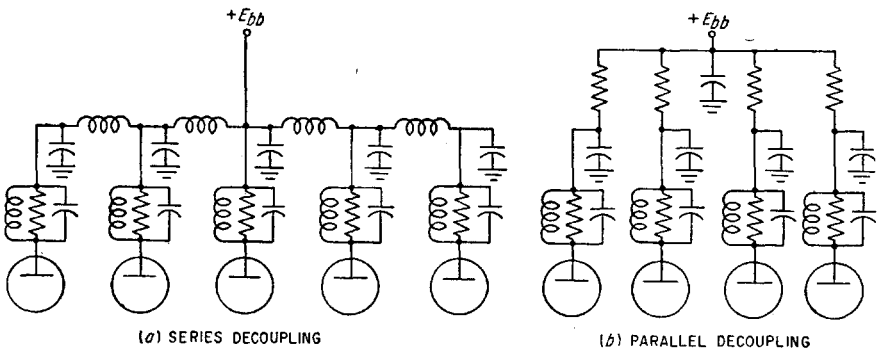
Fig. 7.123. Plate decoupling circuits.

as shown. A decoupling resistor can be used in place of the inductance if the voltage drop across the resistor is not objectionable. Although series decoupling, as shown in Fig. 7.123a, is usually used, parallel decoupling as shown in Fig. 7.123b can also be used. If series decoupling is used, the attenuation in one section of the filter should be approximately 20 db greater than the grid-to-plate stage gain in decibels. In parallel decoupling, the decoupling between the input and output stages should be 20 db greater than the gain in decibels through that portion of the amplifier. The value of inductive reactance which will provide the necessary attenuation for the series-decoupling case is given approximately by

$$X_L \geq 100 g_m X_a{}^2 \qquad (7.253)$$

where X_L = reactance of decoupling inductance at frequency of operation or resistance of decoupling resistor, ohms

X_c = reactance of bypass capacitor in ohms at frequency of operation

g_m = transconductance of amplifier tube

When a series-decoupling inductance is used it is desirable to make the inductance self-resonant at the operating frequency to achieve the highest possible value of series impedance.   If a separate screen supply is utilized, the same considerations apply to the decoupling of the screen circuits except that the grid-screen transconductance is used in Eq. (7.253).   Normally, the screen is either connected directly to the plate bypass as shown in Fig. 7.122, or it is separately bypassed and returned through a series resistor to the plate supply voltage.

7.13b. Grid Circuit Decoupling.   In i-f and r-f amplifiers where a gain control bias voltage is applied to the grids of several stages, similar decoupling considerations apply as discussed in Sec. 7.13a. The feedback path is shown in Fig. 7.124.   The bias control decoupling networks can be placed in series or in parallel just as in the plate decoupling case, and the same considerations in determining the required decoupling apply equally to the bias circuits.   The minimum value of either series decoupling reactances or resistances is given by Eq. (7.253).   Although series-decoupling resistors are frequently used instead of inductances, the time constant of the series-decoupling resistance and the bypass capacitance must be considered in evaluating the response time of the bias control circuit.

7.13c. Bypass Capacitors.   In high gain i-f and r-f amplifiers considerable attention

FIG. 7.124. Feedback path through bias circuit.

must be given to the type of the bypass capacitor used and the manner in which it is mounted, especially at frequencies above a few megacycles.   Any physical capacitor consists of a series combination of a capacitance and an inductance because of the length of the lead from the capacitor to the point in the circuit which is being bypassed. If the value of the lead inductance is large enough to resonate with the bypass capacitance at a frequency somewhat below the frequency of operation of the amplifier, the capacitor may have a high effective reactance to ground, causing instability in the amplifier.   At 10 Mc, a 1,000-$\mu\mu$f bypass capacitor will be series-resonant with a 0.25-μh inductance.

Very effective bypassing can be achieved by making the bypass capacitor series resonant with its lead inductance at the operating frequency.   The effective Q_e of such a circuit can be determined from

$$\frac{1}{Q_e} = \frac{R_L}{\omega_o L} + \frac{1}{\omega_o R_c C} \qquad (7.254)$$

where R_L = lead resistance

R_c = capacitor shunt leakage resistance

L = lead inductance

To achieve good bypassing over a wide bandwidth, i.e., have a low Q circuit, the capacitance should be large and the inductance small. The equivalent series resistance R_s due to the finite Q of a bypass capacitor is given by

$$R_s = \frac{X_c}{Q} \qquad (7.255)$$

where X_c = capacitive reactance at frequency of interest

Q = Q of capacitor neglecting lead inductance

The total series resistance at the series-resonant frequency will be $R_s + R_L$. This will be considerably smaller than X_c.

The self-resonant frequency of a bypass capacitor can be determined by cutting the lead lengths to be equal to those used in the actual circuit and shorting the leads together to form a resonant circuit. A grid-dip meter is then loosely coupled to the circuit, and the frequency of resonance is measured.
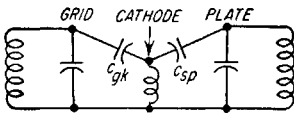
FIG. 7.125. Feedback circuit for a pentode-tuned amplifier having the suppressor tied internally to the cathode.
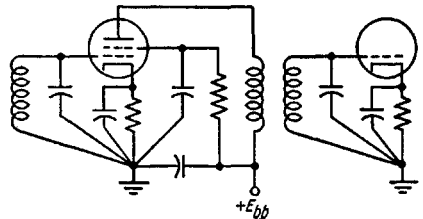
FIG. 7.126. Optimum grounding of the circuit elements in an r-f amplifier to minimize ground-current coupling between stages.

To adequately bypass the elements of an r-f or i-f amplifier circuit which should be at ground potential, the impedance of the bypass capacitor should be in the order of 5 ohms or less at the frequency of operation.

In tubes having a separate suppressor connection, the suppressor should be grounded to minimize the plate-to-cathode capacitance. This is especially true at high frequencies where the cathode lead inductance may have considerable reactance. In tubes where the suppressor is tied internally to the cathode and where the cathode lead has appreciable inductance, the net reactance between grid and plate may appear inductive and, if the stage has sufficient gain, oscillation may occur. The equivalent feedback circuit is shown in Fig. 7.125. Since the stage gain required for oscillation is proportional to the ratio c_{gk}/c_{pk} (see Sec. 6.2), a tube having a separate suppressor ground, so as to minimize c_{pk}, is often required.

7.13d. Grounding. In high gain r-f and i-f amplifiers, ground currents from different stages flowing through a common path in the chassis can introduce sufficient feedback to cause oscillation. Consequently, all of the ground returns which carry signal currents for a particular stage should be returned to a common ground point. This is illustrated in Fig. 7.126. The heater circuit, which is not part of the signal circuit, should be grounded separately to prevent the signal currents of the other tube electrodes from inducing voltages into the filament circuit through a common ground impedance. The electrostatic coupling between the cathode and the heater within the tube will induce small currents into the heater circuit. Amplifier instability caused by feedback through the tube heater circuit is minimized by the use of decoupling inductances between the heater connections of the various heater stages in a manner analogous to the plate circuit decoupling discussed in Sec. 7.13a. Series decoupling is used almost exclusively in filament circuits.

The resonant circuit inductances should be oriented so as to minimize inductive coupling between the inductances and the chassis ground currents.

A multistage amplifier should be constructed in a straight line to achieve maximum isolation between the ground currents of the successive stages.

The outer conductor of a coaxial cable which might introduce a signal into an amplifier should be grounded on the outside of the chassis to eliminate the effects of any undesired currents induced in the outer conductor.

7.13e. Shielding. In high-gain i-f and r-f amplifiers at frequencies above approximately 10 Mc, it is often necessary to completely shield the amplifier to eliminate the introduction of spurious signals and to ensure freedom from regeneration due to coupling around several stages. Also, an amplifier which is completely stable without
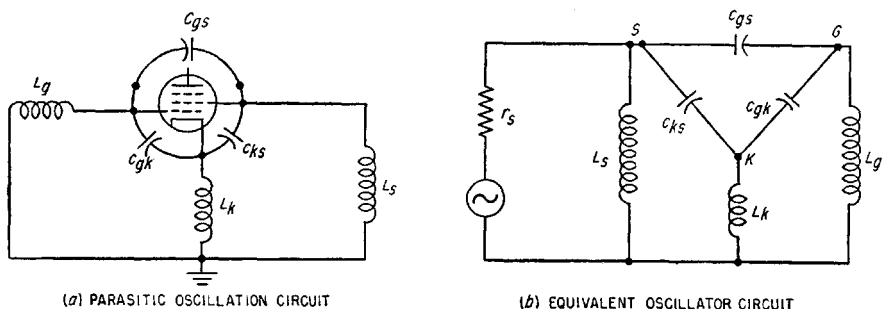
(a) PARASITIC OSCILLATION CIRCUIT   (b) EQUIVALENT OSCILLATOR CIRCUIT

FIG. 7.127. Circuit for parasitic oscillation of a tuned-pentode amplifier.

shielding at low-gain settings is often unstable when the gain is increased to near maximum unless the amplifier is completely shielded. This is usually due to small amounts of inductive coupling between stages which is minimized by the presence of an enclosing shield.

When a shield cover is used on the chassis, the interior of the box can be considered as a waveguide which is capable of propagating an electromagnetic wave from one end to the other. For any practical chassis and shield dimension, the waveguide thus formed has a lower cutoff frequency which is many octaves above the frequency of the amplifier. For this condition and assuming the wave to be the lowest mode possible, the wave will be attenuated approximately 27 db in traveling a length in the box equal to the largest width dimension. Thus, an r-f wave propagated down a box having a cross section of 1 by 2 in. is attenuated approximately 27 db in each 2-in. length of the box. For these particular shield dimensions, the spacing between successive stages should be made large enough so that the gain of the amplifier is considerably less than 27 db for each 2 in. of length.

7.13f. Parasitic Oscillations. Frequently in tuned amplifiers, the effective reactances of the various circuit elements and the inductive reactance presented by the leads within the amplifier tube at high frequencies combine in such a manner as to satisfy the conditions for oscillation at some frequency far removed from the desired frequency of operation. Such spurious oscillations are known as parasitic oscillations (see Sec. 4.4a). Their effect upon the amplifier is the introduction of distortion and nonlinearity in the amplifier gain characteristics and a general tendency to cause the gain and passband characteristics of the amplifier to be particularly sensitive to the locations of external grounds, proximity of adjacent parts, etc.

The presence and frequency of parasitic oscillations can be determined by loosely coupling a grid-dip meter successively to each stage of an r-f or i-f amplifier and searching the frequency range of the grid-dip meter for an indication of oscillation.

The frequency band checked should extend up to about 600 Mc for miniature and subminiature receiving tubes.

In pentode-tuned amplifiers, parasitic oscillations almost always occur at frequencies much higher than the frequency of operation and are usually caused by screen lead inductance. The equivalent circuit is that of a tuned-plate tuned-grid oscillator (see Sec. 6.2c) with the screen serving as the effective plate as shown in Fig. 7.127. Parasitic oscillations of this type can be eliminated by a small resistor in series with the screen-grid lead located between the screen connection and the screen bypass capacitor. The resistance reduces the Q of the parasitic circuit to a value so low that there is insufficient loop gain to sustain oscillation. A series resistor of 10 ohms usually is sufficient to suppress the parasitic oscillation and has a negligible effect upon the operation of the tuned amplifier. The use of a screen bypass capacitor having a low Q also tends to suppress a parasitic oscillation arising from this cause.

Multivibrators

8.1. Introduction. A gate, in the terminology of pulse circuitry, is a positive or negative voltage waveform of essentially constant amplitude and of any arbitrary time duration. Gates are used to operate electronic switch circuits, actuate sawtooth generators or counters, and in numerous other applications where an on-off type of signal is desired.

The most frequently used type of gate generating circuit is the multivibrator. It consists of two tubes interconnected in such a manner that the plate current of one tube is at a maximum when the plate current of the other tube is cut off. At regular intervals or when properly triggered, the previously conducting tube is cut off and the cutoff tube is made to conduct, thereby producing gate voltages at the plates of the tubes. The switching of the tube from one state to the other occurs very rapidly because of the regenerative feedback between the tubes. Multivibrator circuits can be divided into three classes: (1) The *bistable multivibrator*, or *flip-flop*, which remains in one state with either tube conducting and the other cut off unless triggered to initiate the switching action, (2) the *monostable*, or *driven*, *multivibrator* which main-

FIG. 8.1. Bistable multivibrator.

tains current flow in one tube unless triggered, at which time the other tube is made to conduct for a predetermined length of time and then is automatically switched back to its original state, and (3) The *astable*, or *free-running*, multivibrator in which the tubes are switched from one state to the other at regular time intervals without any triggering or actuating voltage.

8.2. Bistable Multivibrators. The circuit of a bistable multivibrator is shown in Fig. 8.1. When plate voltage is applied to the circuit, the tubes conduct almost equally. However, some unbalance is always present in a circuit due to slight differences in tube characteristics, resistor values, tube noise currents, etc., and one tube will start to conduct more than the other. A regenerative switching action then takes place which causes the tube initially conducting more than the other to conduct more heavily and cuts off the plate current of the other tube.

Assume that after the application of plate voltage, $V1$ is conducting and $V2$ is cut off. The plate current of $V1$ flowing through R_1 causes the grid voltage of $V2$ to be held below the cutoff bias by the voltage divider formed by R_2 and R_3. Since $V2$ is cut off, the grid voltage of $V1$ is determined entirely by the voltage divider formed by R_4, R_5, and R_6 between the plate supply voltage E_{bb} and the grid bias supply E_{cc}. The divider is adjusted so that the grid voltage of $V1$ is clamped at approximately 0 volts by grid conduction when $V2$ is cut off. The tubes will remain in this state indefinitely or until a trigger is applied to the circuit. A negative trigger applied to the cathodes of the diodes $V3$ and $V4$ is coupled to the plate of $V2$ through $V4$, but does not appear at the plate of $V1$ since the large reverse bias across $V3$, caused by the volt-

8-2

age drop across R_1, prevents $V3$ from conducting except on triggers large enough to overcome this bias. The negative trigger appearing at the plate of $V2$ is coupled to the grid of $V1$ through C_2. The negative trigger at the grid of $V1$ reduces the plate current through $V1$ causing the plate voltage of $V1$ to increase. This increase is coupled to the grid of $V2$ through C_1 and R_2, causing plate current to flow through $V2$ and further lowering the plate voltage on $V2$. This regenerative switching action causes $V1$ to be cut off and $V2$ to conduct. The tubes will remain in this state until
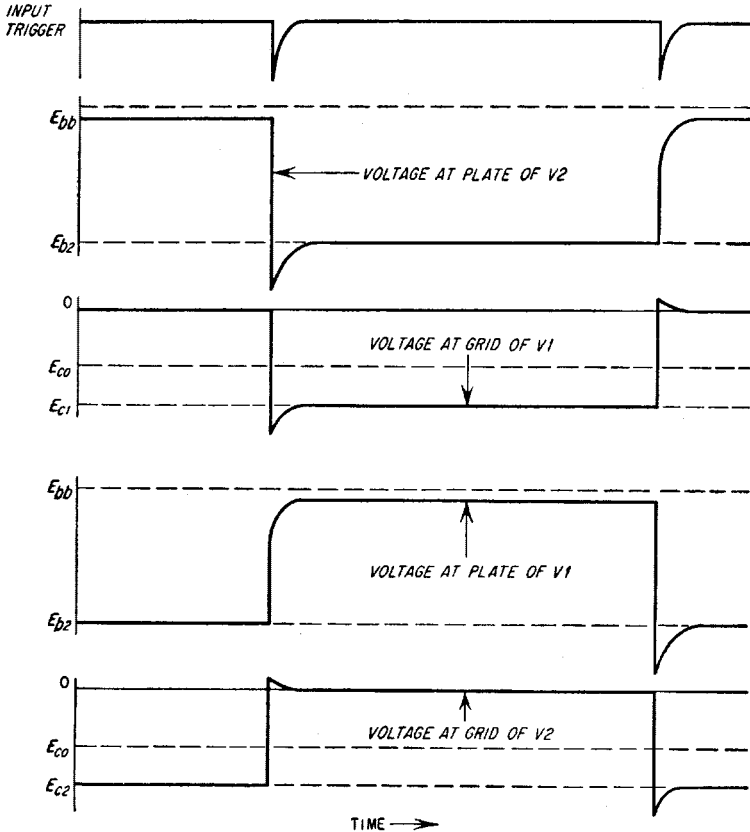
FIG. 8.2. Waveforms for a bistable multivibrator.

a trigger is again applied to the circuit. The waveforms associated with the switching of the plate currents of the two tubes are shown in Fig. 8.2.

Coupling capacitors C_1 and C_2 speed up the regenerative switching action by coupling the full trigger voltage appearing at the plate of the cutoff tube to the grid of the conducting tube. Because of these coupling capacitors, the grid voltage of the tube being switched off is initially driven more negative than its quiescent cutoff value, as illustrated in Fig. 8.2, and the grid of the tube being switched on is initially driven further positive than its quiescent value by the charging current of the coupling capacitor connected to it.

The rise of the waveform at the plate of the tube which is being cut off is slowed because of the charging current of the coupling capacitor connected to the plate (see Fig. 8.2). To obtain the fastest rise of the plate waveform, the coupling capacitor

should be made no larger than necessary for satisfactory triggering of the multivibrator. The coupling capacitance should be made sufficiently larger than the shunt capacitance present at the grid of the conducting tube so that the triggering pulse is not appreciably attenuated by the voltage divider formed by these capacitances. Typical values of coupling capacitors are between 20 and 100 $\mu\mu f$.

The presence of the coupling capacitor between the plate of the tube which is made to conduct and the grid of the tube which is being cut off slows down the fall of the plate waveform of the conducting tube. However, the grid of the conducting tube is initially driven positive by the charging current of the coupling capacitor connected to it, thereby causing the tube to draw plate current in excess of its quiescent value. This results in a dip and an exponential rise in the plate voltage of the conducting tube which masks the effects of the rounding of the waveform due to the discharge current of the coupling capacitor connected to the plate. The input trigger is also coupled through the series diode to the plate of the tube which is being made to conduct and may appear as a negative pulse at the leading edge of the plate waveform if it is of sufficient amplitude. The initial drop in voltage at the plate of the tube being made to conduct is also coupled to the grid of the cutoff tube, thereby increasing the amplitude of the initial dip occurring at this grid when the tube is cut off.

The rise time of the leading edge and the fall time of the trailing edge of each grid and plate waveform is further modified by the shunt capacitances in the circuit (see Sec. 8.6).

In the design of bistable multivibrators, the factors which are usually specified are the amplitude of the voltage change at each plate and the rise time of the waveform at the plate of either tube when that tube is cut off.

The quiescent grid and plate voltages for the two tubes are determined by constructing the d-c load lines for each tube on the E_b-I_b characteristic curves. If it is assumed that the plate-load resistance of either tube is much smaller than the coupling resistor connected to it, the d-c load resistances for the tubes are given by

$$R_{b1} \simeq R_1 \qquad (V1) \tag{8.1}$$
$$R_{b2} \simeq R_4 \qquad (V2) \tag{8.2}$$

where R_1 = plate-load resistance of $V1$
$\qquad R_4$ = plate-load resistance of $V2$

Ordinarily, this assumption is of sufficient accuracy for most applications, and its validity is assumed in the remaining material. In those instances where the decrease in plate voltage due to the current through the coupling resistors is important, Thévenin's theorem can be applied to determine the effective d-c plate-load resistance and plate supply voltage.

The construction of load lines for a typical bistable multivibrator is shown in Fig. 8.3. It is common practice to make the components of a bistable multivibrator symmetrical, that is, R_1 equal to R_4, R_2 equal to R_5, R_6 equal to R_3, and $V1$ and $V2$ the halves of a twin triode. This is assumed in Fig. 8.3.

The voltage at the grid of the conducting tube determined by the resistive voltage divider between the plate of the cutoff tube and the bias supply is always made somewhat positive to ensure that the grid of the conducting tube is conducting and clamps the grid at essentially zero bias. This is done so that precision resistors are not required in the voltage divider network to hold the grid bias sufficiently constant for stable operation without selected parts and tubes. The voltage at the grid of the cutoff tube, as determined by the resistive voltage divider between the plate of the conducting tube and the negative supply voltage, must be below the cutoff voltage for the tube. These conditions impose the following restrictions on R_2, R_3, R_5, and R_6, assuming both R_1 and R_4 to be much smaller than $R_2 + R_3$ and $R_5 + R_6$.

$$\frac{R_6}{R_4 + R_5 + R_6} (E_{bb} - E_{cc}) + E_{cc} \geq 0 \qquad (8.3)$$

$$\frac{R_3}{R_1 + R_2 + R_3} (E_{bb} - E_{cc}) + E_{cc} \geq 0 \qquad (8.4)$$

$$E_{c1} = \frac{R_6}{R_5 + R_6} (E_{b2} - E_{cc}) + E_{cc} \leq E_{co} \qquad (8.5)$$

$$E_{c2} = \frac{R_3}{R_2 + R_3} (E_{b1} - E_{cc}) + E_{cc} \leq E_{co} \qquad (8.6)$$

where E_{b1} = plate voltage of $V1$ when conducting
E_{b2} = plate voltage of $V2$ when conducting
E_{c1} = quiescent grid voltage of $V1$ when cut off
E_{c2} = quiescent grid voltage of $V2$ when cut off
When the above conditions are met, the quiescent plate voltage of the conducting tube is determined by point A in Fig. 8.3.
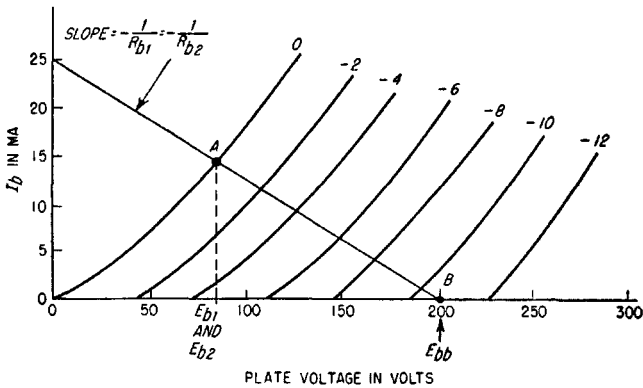
FIG. 8.3. Plate load line for a bistable multivibrator.

The instantaneous plate voltages e_{b1} and e_{b2} for the tubes during the intervals that they are not conducting are given approximately by Eqs. (8.7a) and (8.7b) provided that the grid of the other tube in each instance is positive and assuming no shunt capacitances and that R_1 and R_4 are much smaller than R_2 and R_5, respectively.

$$e_{b1} = E_{bb} - (E_{bb} - E_{b1} + E_{c2})e^{-t_1/R_1C_1} \qquad (V1) \qquad (8.7a)$$
$$e_{b2} = E_{bb} - (E_{bb} - E_{b2} + E_{c1})e^{-t_2/R_4C_2} \qquad (V2) \qquad (8.7b)$$

where t_1 = time in seconds from instant that $V1$ is cut off
t_2 = time in seconds from instant that $V2$ is cut off
The rise time in seconds of the plate voltage waveforms between the zero and 90 per cent points are given approximately by

$$T_1 = R_1C_1 \log_e 10 \left(\frac{E_{bb} + E_{c2} - E_{b1}}{E_{bb} - E_{b1}} \right) \qquad (V1) \qquad (8.8a)$$

$$T_2 = R_4C_2 \log_e 10 \left(\frac{E_{bb} + E_{c1} - E_{b2}}{E_{bb} - E_{b2}} \right) \qquad (V2) \qquad (8.8b)$$

The rise time of the plate waveforms can also be determined graphically from Fig. 8.16.

The requirement for a negative supply voltage can be eliminated by utilizing a common cathode resistor as illustrated in Fig. 8.4. The design procedure is identical

with that described in the preceding material except that the cathode voltage is taken as the zero reference and ground is taken as the bias supply voltage. Assuming a symmetrical circuit, the zero bias plate current of either tube multiplied by the value of the cathode resistance gives the cathode potential. In practice, a bypass capacitor may be required across R_k to stabilize the cathode potential during the switching intervals.
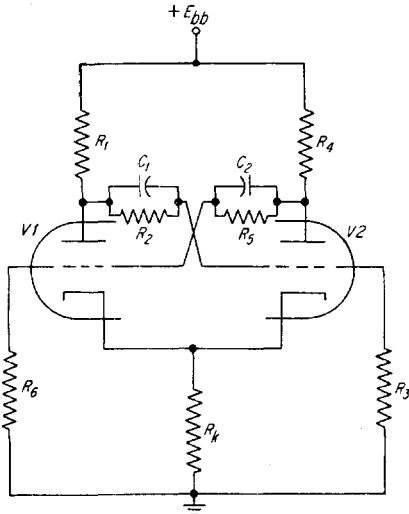
FIG. 8.4. Bistable multivibrator employing cathode bias.

Example 8.1

Design a symmetrical bistable multivibrator utilizing a 12AU7 tube operating from a 250-volt plate supply and a −150-volt grid bias supply. The amplitudes of the plate waveforms should be approximately 100 volts. The E_b-I_b characteristics of the 12AU7 are shown in Fig. 8.5.

Solution

1. Determine the plate-load resistances R_1 and R_4.

Assume that $R_1 \ll R_2$ and $R_4 \ll R_5$. Since the plate waveforms should be about 100 volts in amplitude, construct a vertical line on the E_b-I_b characteristics at 150 volts, that is, 100 volts less than the supply voltage. The maximum plate dissipation for one-half of a 12AU7 is 2.75 watts. The intersections of plate voltage and current values corresponding to 2.75 watts plate dissipation are drawn on the E_b-I_b curves to establish a maximum plate power contour. The plate load line drawn from E_{bb} must cross the zero
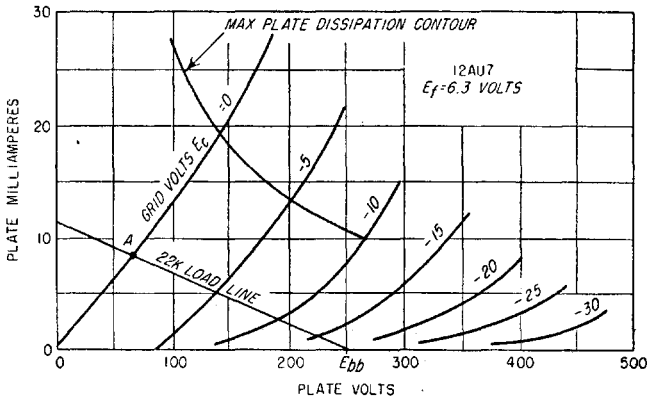
FIG. 8.5. Plate load line for a symmetrical plate-coupled bistable multivibrator.

grid bias line below the maximum plate dissipation contour and to the left of the +150-volt plate voltage line. Any load resistance larger than 5,600 ohms will satisfy this condition. The smaller the value of the plate-load resistance, the faster the rise time of the plate waveform. When the rise time of the plate waveform is of prime importance, the minimum allowable value of the plate load should be used. For this example, assume a plate-load resistance of 22 kilohms and construct the plate load line.

2. Determine the plate voltage of the tubes when conducting.

From Fig. 8.5, the plate voltage at the operating point A is

$$E_{b1} = E_{b2} = +67 \text{ volts}$$

3. Determine the values of R_2 and R_5.

If possible, the coupling resistors should be at least ten times the plate-load resistors. Therefore, let

$$R_2 = R_5 = 470 \text{ kilohms}$$

4. Determine the values of R_6 and R_3.

From Eq. (8.3) for the grid of the conducting tube to be positive,

$$\frac{R_6}{R_4 + R_5 + R_6} (E_{bb} - E_{cc}) + E_{cc} \geq 0$$

Therefore

$$\frac{R_6}{22,000 + 470,000 + R_6} (250 + 150) - 150 \geq 0$$

$$R_6 = R_3 \geq 295 \text{ kilohms}$$

From Eq. (8.5), for the grid of the cutoff tube to be below cutoff bias, i.e., below -22 volts,

$$\frac{R_6}{R_5 + R_6} (E_{b2} - E_{cc}) + E_{cc} \leq E_{co}$$

$$\frac{R_6}{470,000 + R_6} (67 + 150) - 150 \leq -22$$

Therefore

$$R_6 = R_3 \leq 675 \text{ kilohms}$$

Thus R_6 and R_3 can have any value between 295 and 675 kilohms. Let

$$R_6 = R_3 = 470 \text{ kilohms}$$

5. Determine the quiescent plate voltage E of the cutoff tube.

The coupling resistor and the plate-load resistor of the cutoff tube are in series between ground and $+250$ volts. The quiescent plate voltage will, therefore, be

$$E = \frac{470,000}{470,000 + 22,000} \times 250$$

$$= 239 \text{ volts}$$

Since the drop in voltage across the plate-load resistor of the cutoff tube appears as a bias across the diode which couples the negative input trigger to the tube plate, the input trigger amplitude must be made large enough to overcome this voltage drop before any portion of the trigger is coupled to the multivibrator plate.

The complete circuit including 100-$\mu\mu$f coupling capacitors is shown in Fig. 8.6. The minimum trigger amplitude which will cause the multivibrator to switch is that which will cause the grid voltage of the cutoff tube to rise above E_{co} and initiate plate current conduction in the tube (see Sec. 8.7).
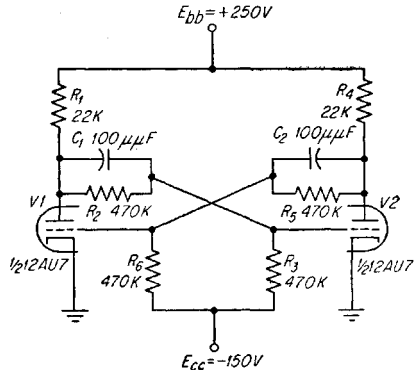
Fig. 8.6. Complete circuit for Example 8.1.

8.3. Plate-coupled Monostable Multivibrators. The circuit of one type of plate-coupled monostable multivibrator is shown in Fig. 8.7. It is similar to the bistable multivibrator previously discussed except that the grid of the normally conducting tube is only capacitively coupled rather than d-c coupled to the plate of the other tube, and the grid resistor of the conducting tube is returned to a potential above cutoff for the tube. As a result of these modifications, $V2$ in Fig. 8.7 conducts and $V1$ is cut off until a negative trigger is applied through the diode $V3$ to the plate of $V1$. A negative trigger of sufficient amplitude at the plate of $V1$ is coupled through C_1 to the grid of $V2$ and causes it to be cut off. As $V2$ is cut off, the plate voltage of $V2$

rises and $V1$ is brought into conduction by the voltage divider action of C_2, R_5 and R_6. The flow of plate current through $V1$ drops the plate voltage of $V1$ to a value deter-

FIG. 8.7. Plate-coupled monostable multivibrator.

mined by the plate load and grid bias. When $V1$ conducts, the plate voltage drop of $V1$ is coupled through C_1 to the grid of $V2$ causing $V2$ to be cut off. The voltage at the grid of $V2$ immediately starts to increase exponentially toward the supply voltage E_{bb}. As soon as cut-off bias E_{co} is reached, $V2$ again begins to conduct and the plate voltage of $V2$ drops and the regenerative switching action takes place. As a result, $V1$ is cut off and $V2$ conducts. This condition is maintained until another negative trigger is applied to the plate of $V1$. The waveforms for this type of plate-coupled multivibrator are illustrated in Fig. 8.8.
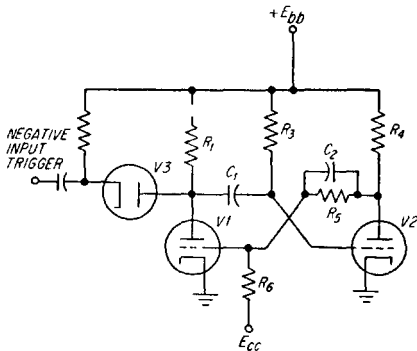
The grid of the normally cutoff tube ($V1$ in Fig. 8.7), need not be resistively coupled to the plate of $V2$ if C_2 is made large enough that the grid voltage of $V1$ does not
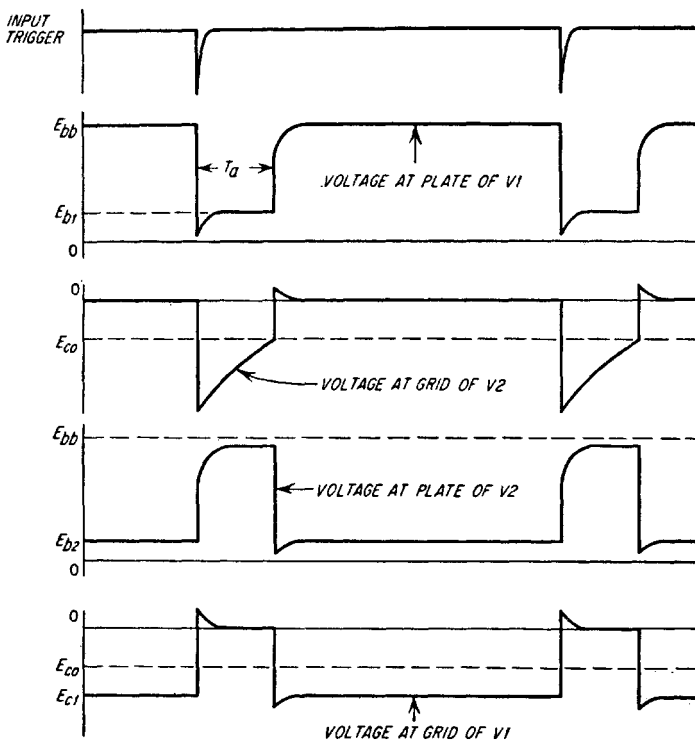
FIG. 8.8. Waveforms for a plate-coupled monostable multivibrator.

drop appreciably during the conducting interval of $V1$. The circuit of such a multivibrator is shown in Fig. 8.9, and the waveforms for this type of plate-coupled monostable multivibrator are shown in Fig. 8.10.

The grid resistor of the normally conducting tube is shown returned to the plate supply voltage E_{bb} in Figs. 8.7 and 8.9. Actually, the grid only needs to be returned to a voltage above cutoff so that the tube conducts until a trigger is applied. Therefore R_3 could have been returned to ground. However, returning the grid of the normally conducting tube to a positive voltage has a twofold benefit. It increases the steepness of the slope of the grid waveform of $V2$ at the cutoff bias voltage point as it is rising toward the grid return voltage after having been cut off by an input trigger. A steep slope at this point is desirable to minimize time jitter of the switching action due to variations in supply voltages as illustrated in Fig. 8.11. Returning the grid of $V2$ to a positive voltage also stabilizes the quiescent grid voltage at nearly

FIG. 8.9. Plate-coupled monostable multivibrator with a-c coupling between both stages.

0 volts. If the grid of $V2$ is not returned to either ground or a positive voltage, a precision voltage divider is required to establish accurately the quiescent grid bias.
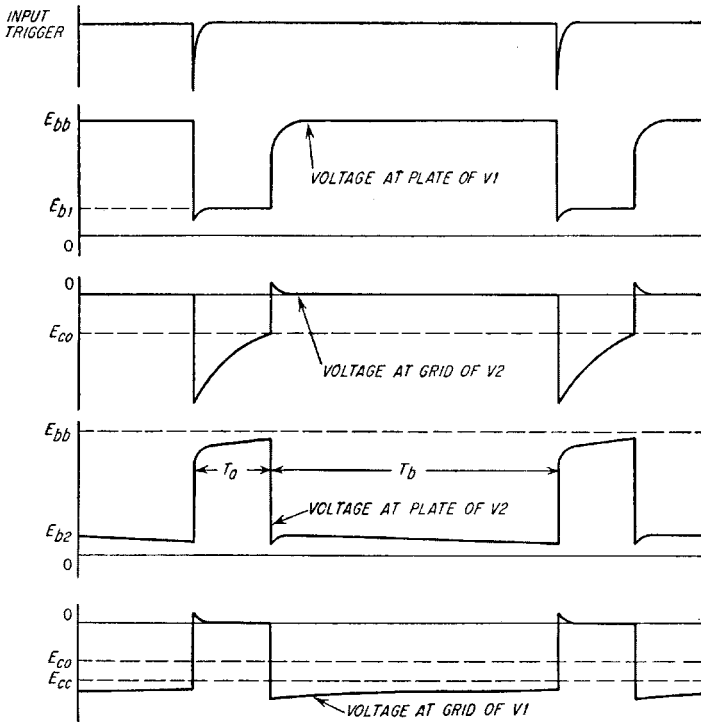
FIG. 8.10. Waveforms for the plate-coupled monostable multivibrator shown in Fig. 8.9.

8.3a. Plate-coupled Monostable Multivibrator with D-C Coupling between the Plate of the Normally Conducting Tube and the Grid of the Normally Cut-off Tube (see Fig.
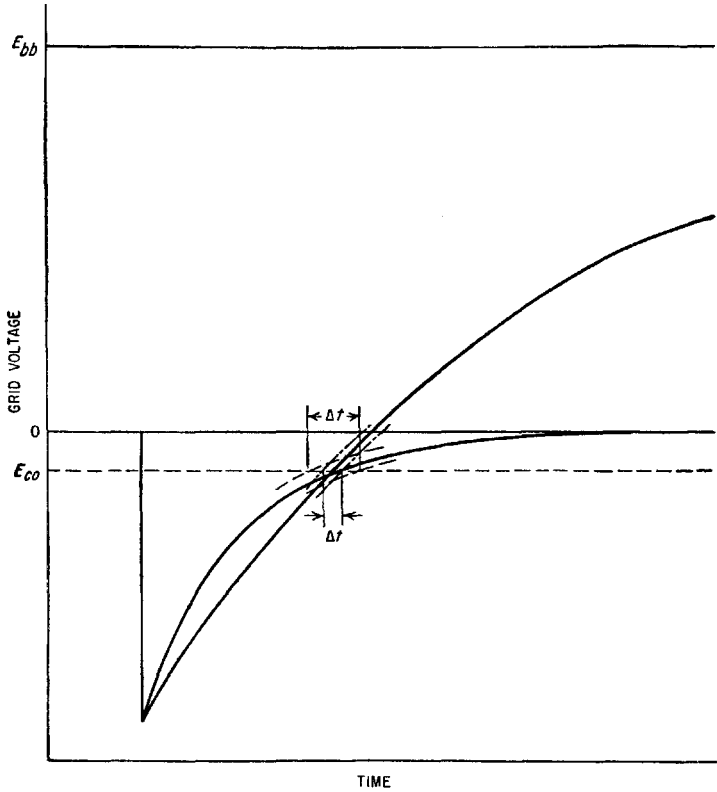
Fig. 8.11. Variation Δt in the period that the normally conducting tube is cut off as a function of the grid return voltage for a fixed amplitude displacement of the timing waveform.

8.7). In this analysis, it is assumed that $V2$ is the normally conducting tube, that the coupling resistor R_5 is much larger than R_4 and that the grids of $V2$ and $V1$ are clamped at 0 volts whenever either tube is conducting. Assuming that R_4 is much less than $R_5 + R_6$, the d-c load resistances for each tube are
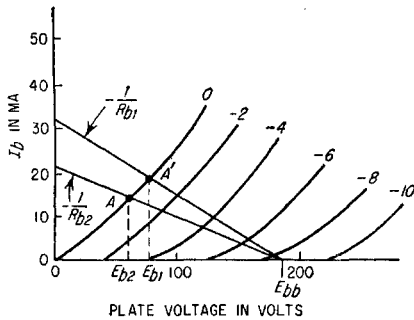
Fig. 8.12. Construction of d-c load lines for a plate-coupled monostable multivibrator.

$$R_{b1} \simeq R_1 \qquad (V1) \qquad (8.9)$$
$$R_{b2} \simeq R_4 \qquad (V2) \qquad (8.10)$$

The construction of a typical load line is shown in Fig. 8.12. The quiescent plate voltage E_{b2} of $V2$ when conducting is determined by the intersection of the d-c load line and the zero bias line on the E_b-I_b characteristic curves for the tube. This is shown as point A in Fig. 8.12. The plate voltage E_{b1} of $V1$ when conducting is likewise determined by the intersection of the d-c load line and the zero bias line. This is shown as point A' in Fig. 8.12.

The instantaneous grid voltage of $V2$ when cut off by the switching action is determined from the equivalent circuit of Fig. 8.13. The instantaneous grid voltage e_{c2} as a function of time t after the switching action has cut off $V2$ is given by

$$e_{c2} = \frac{R_3}{R_3 + R_1 r_{p1}/(R_1 + r_{p1})} (E_{b1} - E_R - E_{bb})e^{-t/T_x} + E_R \qquad (V2) \qquad (8.11)$$

where R_3 = grid resistor of $V2$

r_{p1} = dynamic plate resistance of $V1$ evaluated at operating point A' of E_b-I_b characteristic curves (see Fig. 8.12)

E_{b1} = plate voltage of $V1$ when conducting

E_R = bias supply voltage for grid of $V2$

T_x = time constant in seconds of discharge circuit of C_1

$$= C_1 \left(R_3 + \frac{R_1 r_{p1}}{R_1 + r_{p1}} \right)$$

In most practical circuits, R_3 will be between 10 and 100 times the parallel combination of R_1 and r_{p1}. For this condition, Eq. (8.11) can be approximated by

$$e_{c2} = (E_{b1} - E_R - E_{bb})e^{-t/R_3 C_1} + E_R \tag{8.12}$$

As a function of time, e_{c2} rises exponentially toward E_R. When e_{c2} reaches the cutoff bias E_{co} for $V2$, current flows in $V2$ and the resulting switching action switches $V1$ off and $V2$ on. The length of time that $V2$ is cut off is determined, therefore, by the time T_a required for e_{c2} to reach E_{co} after $V2$ is initially switched off. This period is given by

FIG. 8.13. Approximate equivalent circuit and initial conditions for determining e_{c2} when $V2$ is switched off.

$$T_a = T_x \log_e \frac{R_3(E_{b1} - E_R - E_{bb})}{(E_{co} - E_R)[R_3 + R_1 r_{p1}/(R_1 + r_{p1})]} \tag{8.13}$$

For the condition,

$$R_3 \gg \frac{R_1 r_{p1}}{R_1 + r_{p1}}$$

Eq. (8.13) reduces to

$$T_a = R_3 C_1 \log_e \left(\frac{E_{b1} - E_R - E_{bb}}{E_{co} - E_R} \right) \tag{8.14}$$

The value of time delay T_a in Eq. (8.14) is readily determined graphically from Fig. 8.14 where RC is equal to $R_3 C_1$ and T is equal to T_a.

When e_{c2} reaches the cutoff bias for $V2$, the resultant switching action causes $V2$ to conduct. The equivalent circuit for the charging of C_1 is shown in Fig. 8.15. With $V1$ cut off, C_1 charges from the potential $E_{b1} - E_{co}$, which it had at the instant of switching, toward the difference between the plate supply voltage E_{bb} and the grid voltage E_{c2} for $V2$ when conducting. Because the dynamic grid resistance r_{g2} of $V2$ in the positive grid region is usually quite small compared to R_1, the instantaneous plate voltage e_{b1} of $V1$ at time t after cutoff is given closely by Eq. (8.15).

$$e_{b1} = E_{bb} - (E_{bb} + E_{co} - E_{b1})e^{-t/R_1 C_1} \tag{8.15}$$

The rise time T_1 of e_{b1} from 0 to the 90 per cent point is given approximately by

$$T_1 = R_1 C_1 \log_e 10 \left(\frac{E_{bb} + E_{co} - E_{b1}}{E_{bb} - E_{b1}} \right) \tag{8.16}$$

This is plotted in Fig. 8.16. The presence of shunt capacitances will modify Eqs. (8.15) and (8.16) as discussed in Sec. 8.6. The finite value of r_{g2} in the positive grid

FOR MONOSTABLE PLATE-
COUPLED MULTIVIBRATORS
(SEE FIGS. 8.7 AND 8.9)
$$\frac{T_a}{R_3 C_1} = LOG_e A \quad WHERE \ A = \left(\frac{E_{b1} - E_R - E_{bb}}{E_{co} - E_R}\right)$$

FOR A MONOSTABLE CATHODE-
COUPLED MULTIVIBRATOR
(SEE FIG. 8.21)
$$\frac{T_a}{R_3 C_1} = LOG_e A \quad WHERE \ A = \left(\frac{E_{b1} + E_{k2} - 2E_{bb}}{E_{k1} + E_{co} - E_{bb}}\right)$$

FOR AN ASTABLE PLATE-
COUPLED MULTIVIBRATOR
(SEE FIG. 8.27)
$$\frac{T_a}{R_3 C_1} = LOG_e A \quad WHERE \ A = \left(\frac{E_{b1} - E_{bb} - E_R}{E_{co} - E_R}\right)$$
$$\frac{T_b}{R_6 C_2} = LOG_e A \quad WHERE \ A = \left(\frac{E_{b2} - E_{bb} - E_R}{E_{co} - E_R}\right)$$
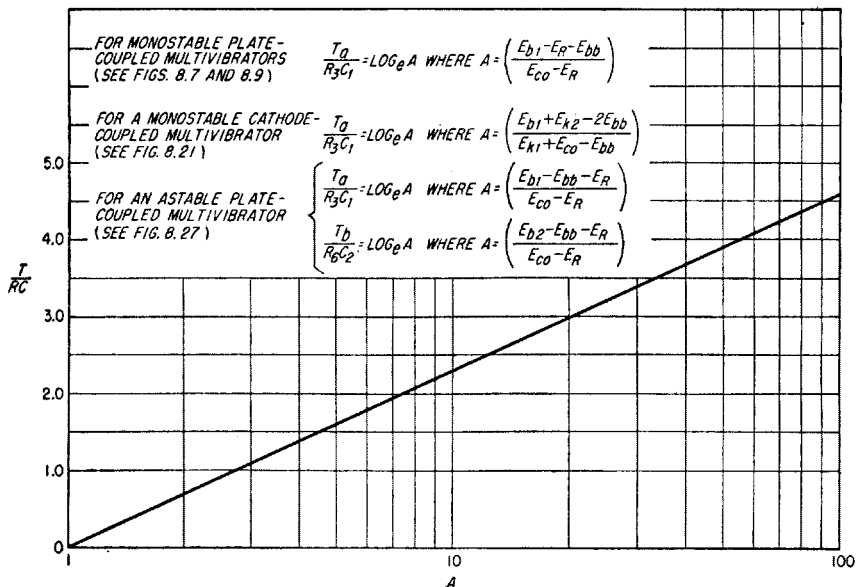
FIG. 8-14. Multivibrator gate lengths.

region causes a small pip in the grid voltage waveform at the beginning of the conduction period of $V2$ as shown in Fig. 8.8. This waveform causes a corresponding dip in the plate waveform of $V2$ and a decrease in the initial value of the grid exponential

FIG. 8.15. Approximate equivalent circuit and initial conditions for charging of C_1 when $V2$ is switched on.

of $V1$. A corresponding pip also occurs at the grid of $V1$ when it is switched on, causing similar dips in the plate waveform of $V1$ and the grid waveform of $V2$.

The value of C_2 is determined on the basis that it should be large enough to ensure reliable switching and as small as possible to minimize the rise time of the plate voltage waveform of $V2$ at the time that $V2$ is switched off. It is ordinarily made several times larger than the input capacitance of $V1$ to ensure that most of the negative voltage waveform at the plate of $V2$ will appear at the grid of $V1$ so as to cut off $V1$.

The values of R_5 and R_6 are determined from Eqs. (8.3) and (8.5) to ensure that the grid of $V1$ is conducting when $V1$ is switched on and that $V1$ is cut off when $V2$ is conducting.

The equivalent circuit for the charging of C_2 when $V2$ is cut off is shown in Fig. 8.17. The instantaneous plate voltage e_{b2} of $V2$ at time t after cutoff, assuming R_4 to be much larger than r_{g1} and much smaller than R_5, is given by

$$e_{b2} = E_{bb} - (E_{bb} - E_{b2} + E_{c1})e^{-t/R_4 C_2} \qquad (8.17)$$

The rise time T_2 of the plate waveform of $V2$ from 0 to the 90 per cent point is

given approximately by

$$T_2 = R_4C_2 \log_e 10 \left(\frac{E_{bb} - E_{b2} + E_{c1}}{E_{bb} - E_{b2}} \right) \tag{8.18}$$

This is plotted in Fig. 8.16.   From Eq. (8.18) and Fig. 8.16, it can be seen that the rise time of the plate waveform of $V2$ is minimized by making E_{c1} as large (negative)

FOR A CATHODE-COUPLED MONOSTABLE MULTIVIBRATOR (SEE FIG. 8.21): $(V1)\frac{T_1}{R_1C_1} = LOG_e\,10A$ WHERE $A = \frac{E_{bb}-E_{b1}-E_{k2}+E_{k1}+E_{co}}{E_{bb}-E_{b1}}$

FOR PLATE-COUPLED MONOSTABLE MULTIVIBRATORS WITH D-C COUPLING (SEE FIG. 8.7):
$(V1)\frac{T_1}{R_1C_1} = LOG_e\,10A$ WHERE $A = \frac{E_{bb}+E_{co}-E_{b1}}{E_{bb}-E_{b1}}$
$(V2)\frac{T_2}{R_4C_2} = LOG_e\,10A$ WHERE $A = \frac{E_{bb}+E_{c1}-E_{b2}}{E_{bb}-E_{b2}}$

FOR A BISTABLE PLATE-COUPLED MULTIVIBRATOR (SEE FIG. 8.1):
$(V1)\frac{T_1}{R_1C_1} = LOG_e\,10A$ WHERE $A = \frac{E_{bb}+E_{c2}-E_{b1}}{E_{bb}-E_{b1}}$
$(V2)\frac{T_2}{R_4C_2} = LOG_e\,10A$ WHERE $A = \frac{E_{bb}+E_{c1}-E_{b2}}{E_{bb}-E_{b2}}$

FOR PLATE-COUPLED ASTABLE MULTIVIBRATORS (SEE FIG. 8.27):
$(V1)\frac{T_1}{R_1C_1} = LOG_e\,10A$ WHERE $A = \frac{E_{bb}+E_{co}-E_{b1}}{E_{bb}-E_{b1}}$
$(V2)\frac{T_2}{R_4C_2} = LOG_e\,10A$ WHERE $A = \frac{E_{bb}+E_{co}-E_{b2}}{E_{bb}-E_{b2}}$

(Vertical axis: $\frac{T}{RC}$, values 0, 0.1, 0.2, 0.3. Horizontal axis: A, values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0.)

FIG. 8.16. Rise time between 0 and 90 per cent points of plate waveform of a multivibrator due to coupling capacitor.

as possible.   The effects of shunt and stray capacitances further modify the rise time as discussed in Sec. 8.6.   Since C_2 is ordinarily considerably smaller than C_1 and since E_{c1} can ordinarily be somewhat larger than E_{co}, the rise time of the plate waveform e_{b2} of $V2$ can usually be made considerably shorter than the rise time of e_{b1}.   However, the negative-going waveforms at the plates of both tubes have shorter response times than the positive-going waveforms at either tube because the plate resistance of each tube appears in parallel with the load resistances and because the positive pip at the grid of the conducting tube sharpens up the drop in plate voltage for that tube.   For these reasons, a negative-going plate waveform should be utilized from a plate-coupled multivibrator, if possible, when the minimum "switching on" time is desired.

FIG. 8.17. Approximate equivalent circuit and initial conditions for determining the instantaneous plate voltage e_{b2} when $V2$ is cut off.

8.3b. Plate-coupled Monostable Multivibrator with Capacitive Coupling between the Plate of the Normally Conducting Tube and the Grid of the Normally Cut-off Tube.   The circuit of this multivibrator is shown in Fig. 8.9.   It is identical in operation to the monostable multivibrator previously described except that the grid of $V1$ is capacitively coupled to the plate of $V2$.   As a consequence, the grid of $V1$ is held above cutoff

during the period that $V1$ conducts only by the charging current of C_2 flowing through R_6. The charging time constant for C_2 must be very much larger than the period that $V1$ conducts if the grid voltage of $V1$ when conducting and the plate voltage of $V2$ when cut off are to remain essentially constant. Normally, the charging current changes by several per cent during the on period of $V1$, causing slightly tilted waveforms as shown in Fig. 8.10. The equivalent circuits for the charge and discharge of C_2 are shown in Fig. 8.18a and b, respectively. When the multivibrator is triggered at a constant rate, the voltage across C_2 adjusts itself so that the charge lost during the period that $V2$ is conducting is just regained in the period that $V1$ is conducting. The time constants of the charging and discharging circuits are usually made equal to many times the periods T_a and T_b, respectively, in order that the top of the plate waveform of $V2$ may be nearly flat (see Fig. 8.10). If this is done, the voltage across

(a) EQUIVALENT CIRCUIT FOR CHARGING C_2 (b) EQUIVALENT CIRCUIT FOR DISCHARGE OF C_2
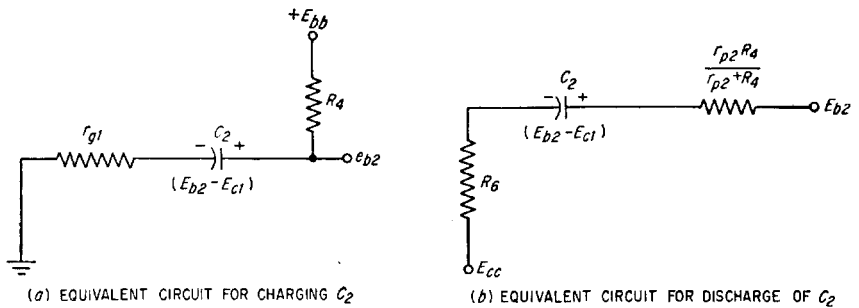
FIG. 8.18. Approximate equivalent circuits and initial conditions for charge and discharge of C_2.

the capacitor will change by only a small amount during the time that $V1$ conducts and the capacitor voltage E_c may be approximated by

$$E_c \simeq \frac{(E_{b2} - E_{cc})T_b/R_b + E_{bb}T_a/R_a}{T_b/R_b + T_a/R_a} \qquad (8.19)$$

where T_a = period in seconds that $V1$ conducts [see Eq. (8.14)]

T_b = period in seconds that $V2$ conducts

$R_a = r_{g1} + R_4$

$R_b = R_6 + \dfrac{R_4 r_{p2}}{R_4 + r_{p2}}$

The voltage across C_2 is not constant since there is a slight increase in the capacitor voltage during the interval T_a and a slight decrease in the capacitor voltage during the interval T_b. From Eq. (8.19), it can be seen that when T_a/R_a is large compared to T_b/R_b the voltage across C_2 approaches E_{bb}, and if T_b/R_b is large compared to T_a/R_a, the voltage across C_2 approaches $E_{b2} - E_{cc}$.

The d-c load lines for $V1$ and $V2$ are constructed on the E_b-I_b characteristic curves for the tubes using R_1 and R_4 as the load resistances and E_{bb} as the supply voltage. The instantaneous grid voltage e_{c2} of $V2$ at time t after cutoff is given by Eq. (8.11) or (8.12), and the duration of the period that $V2$ is cut off is given by Eq. (8.13) or (8.14) and may be determined from Fig. 8.14.

The plate voltage e_{b1} of $V1$ at time t after cutoff, assuming $R_1 \gg r_{p2}$, is given by Eq. (8.15), and the rise time T_1 of the plate waveform is given by Eq. (8.16) and Fig. 8.16. The instantaneous plate voltage e_{b2} of $V2$ at time t after cutoff, assuming $R_4 \gg r_{g1}$, is given approximately by

$$e_{b2} = E_{bb} - (E_{bb} - E_c)e^{-t/R_a C_2} \qquad (8.20)$$

The voltage e_{b2} approaches E_{bb} and has a nearly flat top if the following conditions are met:

$$R_a C_2 \geq 5T_a$$
$$R_b C_2 \geq 5T_b$$
$$T_a/R_a \geq 5T_b/R_b$$

When these conditions exist, E_c is nearly equal to E_{bb} and the exponential term in Eq. (8.20) approaches zero. The rise time T_2 of the waveform e_{b2} is then limited only by the shunt capacitances in parallel with R_4. The necessity for a relatively large value of C_2, however, increases somewhat the minimum shunt capacitance over that which can be achieved in the direct-coupled circuit.

Example 8.2

Design a plate-coupled monostable multivibrator having direct coupling between the plate of the normally conducting tube and the grid of the normally cut-off tube. The positive waveform at the plate of $V2$ should have a minimum amplitude of 150 volts and a duration of 400 μsec. The period between triggering pulses is 1,000 μsec. Use a 12AU7 tube, a 250-volt plate supply, and a −100-volt bias supply. The E_b-I_b characteristic curves are given in Fig. 8.19.

FIG. 8.19. Characteristic curves for Example 8.2.

Solution

1. Determine the plate resistor R_4 for $V2$.
The value of R_4 is a compromise between the amplitude and the rise time requirements of the gate at the plate of $V2$. In general, R_4 is made just large enough to provide the required gate amplitude, thus securing the best rise time possible under the conditions imposed. Since the current flow between E_{bb} and ground through R_4, R_5, and r_{g1} causes the quiescent voltage at the plate of $V2$ when nonconducting to be slightly less than E_{bb}, the quiescent plate voltage of $V2$ when conducting should be somewhat less than 100 volts to ensure a minimum gate amplitude of 150 volts. From Fig. 8.19, it can be seen that a plate load resistance of 25 kilohms causes the plate voltage of $V2$ when conducting to be equal to 65 volts, which should permit a 150-volt positive gate.

2. Determine R_6 and R_5.
In general, R_5 should be made as large as possible to minimize the current flowing between E_{bb} and ground through R_4, R_5, and r_{g1} as described in step 1. Since R_5 is largest for large values of R_6, assume

$$R_6 = 1 \text{ megohm}$$

Then, from Eq. (8.3)

$$\frac{10^6}{25,000 + 10^6 + R_5} (250 + 100) - 100 \geq 0$$

$$R_5 \leq 2.47 \times 10^6 \text{ ohms}$$

From Eq. (8.5)

$$\frac{10^6}{R_5 + 10^6} (65 + 100) - 100 \leq -25$$

$$R_5 \geq 1.2 \times 10^6 \text{ ohms}$$

Therefore, let $R_5 = 1.5 \times 10^6$ ohms.
When $V2$ is cut off, the voltage at the plate of $V2$ is very nearly equal to $E_{bb}R_5/(R_4 + R_5)$, or 246 volts. Therefore, the amplitude of the waveform at the plate of $V2$ is 246 − 65, or 181 volts.

3. Determine the value of R_1.
If the plate waveform of $V1$ is to be utilized, the same considerations as to rise time apply as for R_4. Otherwise, the value of R_1 is not critical except that a large amplitude waveform is desirable at the plate of $V1$ to minimize timing jitter. Let $R_1 = 100,000$ ohms. Construct the d-c load line for $V1$ on the E_b-I_b characteristic curves. Therefore,

$$E_{b1} = 15 \text{ volts}$$

4. Determine E_R.

For maximum timing accuracy, the grid of $V2$ should be returned to a high positive voltage. Therefore, let $E_R = E_{bb}$.

5. Determine R_3 and C_1.

The time constant R_3C_1 is determined from Fig. 8.14.   The ratio A is equal to

$$A = \frac{15 - 250 - 250}{-25 - 250} = 1.76$$

Therefore

$$\frac{T}{R_3C_1} = 0.57$$

$$R_3C_1 = \frac{400}{0.567}$$

$$= 705 \ \mu sec$$

The grid resistance R_3 should be large to minimize the grid current of $V2$ when conducting. Therefore, let $R_3 = 1$ megohm.   The value of C_1 is then

$$C_1 = 705 \ \mu\mu f$$

6. Determine C_2.

The considerations in determining C_2 are discussed in the text.   A value of 47 $\mu\mu f$ is satisfactory.   The complete circuit is shown in Fig. 8.20.
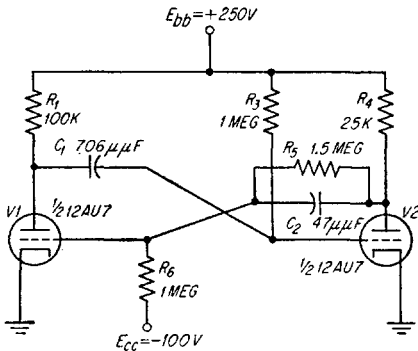
FIG. 8.20. Complete circuit for Example 8.2.   FIG. 8.21. Cathode-coupled monostable multivibrator.
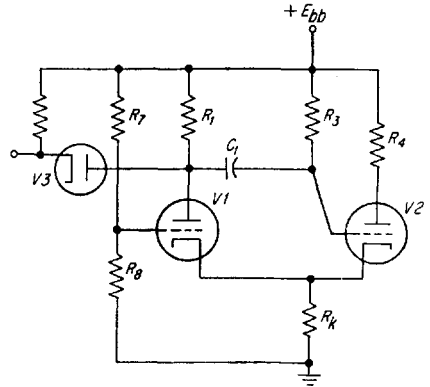
8.4. Cathode-coupled Monostable Multivibrators.   The circuit of a cathode-coupled monostable multivibrator is shown in Fig. 8.21.   The cathodes of both tubes are connected to a common cathode resistor.   The grid of the normally conducting tube $V2$ is returned to a positive voltage through a large resistance so that the grid-cathode voltage of $V2$ is essentially zero.   The grid of $V1$ is connected to a potential sufficiently less positive than the common cathodes of both tubes so as to cut off $V1$.   A negative trigger applied to the plate of $V1$ is coupled to the grid of $V2$ through C_1, causing the grid and cathode voltages of $V2$ to drop.   When the cathode voltage drops to the level where $V1$ is no longer cut off, $V1$ begins to conduct, further reducing the plate voltage of $V1$.   The ensuing regenerative switching action causes $V2$ to be cut off and $V1$ to conduct until C_1 discharges sufficiently so that $V2$ again conducts, which in turn switches $V1$ off.   The circuit remains in this state until another trigger is applied. The waveshapes of a cathode-coupled monostable multivibrator are illustrated in Fig. 8.22.

The d-c load resistance for $V1$ is $R_1 + R_k$ and the d-c load resistance for $V2$ is $R_4 + R_k$.   Load lines for a typical cathode-coupled multivibrator are shown in Fig.

8.23. The grid of the normally conducting tube $V2$ is usually returned through a large resistance to the supply voltage E_{bb}. The grid-cathode voltage of $V2$ is, therefore, very nearly equal to 0 volts, and the quiescent level of the grid and cathode voltages of $V2$ is

$$E_{c2} = E_{k2} = I_2 R_k \qquad (8.21)$$

where $I_2 =$ quiescent current through $V2$ when $V2$ is conducting

= current at point A on load line of $V2$ (see Fig. 8.23)

$E_{k2} =$ cathode voltage of $V1$ and $V2$ when $V2$ is conducting

The quiescent plate-to-ground voltage E_{b2} of $V2$ when conducting is given by

$$E_{b2} = E_{bb} - I_2 R_4 \qquad (8.22)$$

The voltage E_{c1} to which the grid of $V1$ must be returned to maintain $V1$ at cutoff is given by

$$E_{c1} \le E_{k2} + E_{co} \qquad (8.23)$$

Fig. 8.22. Waveforms of a typical cathode-coupled monostable multivibrator.

where $E_{co} =$ cutoff bias for $V1$ for a plate voltage equal to $E_{bb} - E_{k2}$

The grid voltage of $V1$ should be adjusted so that the grid of $V1$ is clamped to the cathode voltage by grid conduction when $V1$ is conducting. This requires that

$$E_{c1} \ge I_1 R_k \qquad (8.24)$$

where $I_1 =$ plate current of $V1$ at zero bias

The plate current I_1 is determined by the intersection of the zero-bias characteristic curve with the tube load line as shown by point B in Fig. 8.23.

Fig. 8.23. Typical load lines for a cathode-coupled monostable multivibrator.

Neglecting tube and wiring capacitances, the voltage at the plate of $V2$ would rise to E_{bb} at the instant $V2$ is cut off without any exponential rise or overshoot. The cathode voltage E_{k2} likewise would drop to E_{k1} without any exponential decay or overshoot as rapidly as the switching action cuts off $V2$.

The equivalent circuit for the discharge of C_1 when $V2$ is cut off is shown in Fig. 8.24a. The instantaneous grid voltage e_{c2} of $V2$ at time t after cutoff, assuming R_3 to be much larger than the parallel combination of R_1 and $r_{p1} + \mu R_k$, is given approximately by

$$e_{c2} = E_{bb} + (E_{b1} + E_{k2} - 2E_{bb})e^{-t/R_3C_1} \qquad (8.25)$$

After e_{c2} exponentially increases to a value equal to E_{k1} less the cutoff voltage for $V2$, $V2$ conducts and the regenerative switching action cuts off $V1$. The duration of the

NOTE: $E_{k2} = E_{c2}$

(a) APPROXIMATE EQUIVALENT CIRCUIT
AND INITIAL CONDITIONS FOR DISCHARGE
OF C_1 WHEN $V2$ IS CUTOFF

(b) APPROXIMATE EQUIVALENT CIRCUIT
AND INITIAL CONDITIONS FOR CHARGE
OF C_1 WHEN $V2$ IS SWITCHED ON
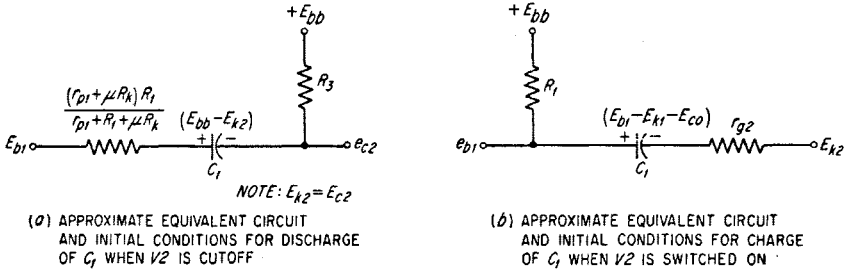
FIG. 8.24. Equivalent circuits for discharge and charge of C_1.

time interval T_a that $V1$ conducts is given by

$$T_a = R_3C_1 \log_e \left(\frac{E_{b1} + E_{k2} - 2E_{bb}}{E_{k1} + E_{co} - E_{bb}}\right) \qquad (8.26)$$

The time interval T_a can be determined graphically from Fig. 8.14 by making A equal to the logarithmic term in Eq. (8.26).

The equivalent charging circuit for C_1 when $V2$ is switched on is shown in Fig. 8.24b. The instantaneous plate voltage e_{b1} of $V1$ at time t after cutoff, assuming $R_1 \gg r_{g2}$, is given approximately by

$$e_{b1} = E_{bb} - (E_{bb} + E_{k1} - E_{k2} - E_{b1} + E_{co})e^{-t/R_1C_1} \qquad (8.27)$$

where t = time in seconds from the instant $V1$ is switched off

The rise time T_1 of this waveform from the 0 to the 90 per cent points is given by

$$T_1 = R_1C_1 \log_e 10 \left(\frac{E_{bb} + E_{k1} - E_{k2} - E_{b1} + E_{co}}{E_{bb} - E_{b1}}\right) \qquad (8.28)$$

This is plotted in Fig. 8.16. The positive peak of e_{c2} at the time $V2$ is switched on due to the charging of C_1 through r_{g2} causes a corresponding peak in the cathode voltage E_{k1} and a slight dip in the plate voltage E_{b2} at the same instant as shown in Fig. 8.22. The cathode-coupled monostable multivibrator possesses the advantages over the plate-coupled types of (1) having one plate which is completely "free," i.e., no feedback for the multivibrator is obtained from this plate and only shunt capacitance limits the rise time of the waveform at the plate, and (2) having a gate duration which can be made very nearly a linear function of a d-c control voltage (see Sec. 9.6). It has the disadvantage that, when triggered by a negative pulse coupled to the plate of $V1$ as shown in Fig. 8.21, the trigger appearing between grid and cathode of $V1$ to initiate conduction is not amplified by $V2$ as in the plate-coupled multivibrator. As a result, a considerably larger trigger amplitude is required at the plate of $V1$ than for a plate-coupled multivibrator.

Example 8.3

Design a cathode-coupled monostable multivibrator using a type 6J6 tube with the requirement that a gate waveform having a minimum amplitude of 100 volts be present

at the plate of $V2$.   Let the period that $V2$ remains nonconducting when triggered be equal
to 400 μsec.   Assume a 250-volt supply.   The E_b-I_b characteristic curves for the 6J6 are
given in Fig. 8.25.

Solution

1. Determine R_4 and R_k.
The values of R_4 and R_k are determined from the following considerations:
(1) $I_2R_4 \geq 100$ volts.
(2) $I_2R_k \geq E_{co} = 10$ volts for a 6J6 with $E_{bb} = 250$ volts.   (To permit a slight factor of
safety, E_{co} has been assumed to be equal to -12 volts in the following calculations.)
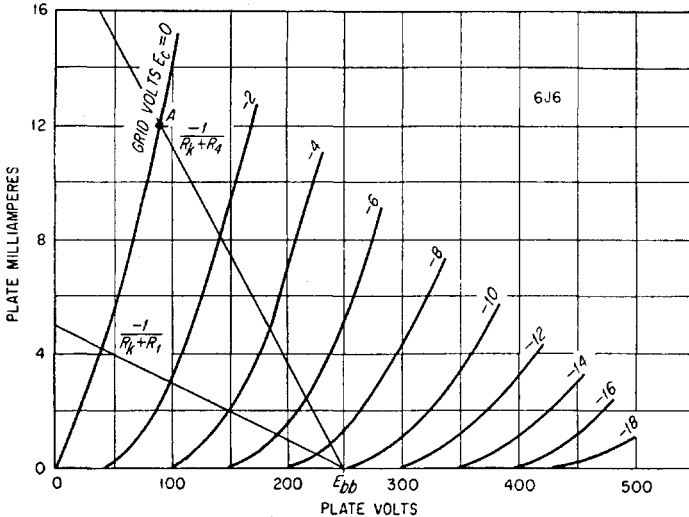
FIG. 8.25. Plate load lines for Example 8.3.

(3) For the fastest rise time of e_{b2}, R_4 should be as small as possible to minimize the
effects of tube and wiring capacitances.
(4) The smallest value of $R_4 + R_k$ is determined by the allowable plate dissipation for $V2$.
Assume that rise-time considerations limit R_4 to a maximum value of 10 kilohms.   For a
plate waveform amplitude of 100 volts minimum, the tube current at zero bias must be
10 ma or more.   Allowing a safety factor of 20 volts, a zero bias plate current of 12 ma is
required for the 10 kilohms plate-load resistance.   Construct a d-c load line on the E_b-I_b
characteristic curves satisfying these conditions.   This corresponds to a plate plus cathode
load resistance of

$$R_4 + R_k = \frac{250 - 90}{12 \times 10^{-3}} = 13.3 \times 10^3 \text{ ohms}$$

Therefore $R_k = 3,300$ ohms, and $E_{k2} = I_2R_k$, or 40 volts.
2. Determine R_1.
When $V1$ is made to conduct, the grid voltage e_{c2} of $V2$ must initially drop to a value less
than $E_{k1} + E_{co}$ if $V2$ is to be cut off.   Therefore, the gate amplitude at the plate of $V1$
must be somewhat greater than $E_{k1} - (E_{k1} + E_{co})$ for $V2$ to be cut off.   This condition
requires that

$$I_1(R_1 + R_k) \geq E_{k2} - E_{co} = 40 - (-12) = 52 \text{ volts}$$

Thus, the voltage drop across V1 must be somewhat less than 198 volts when $V1$ is conduct-
ing.   For maximum stability of the gate length, e_{c2} should initially drop to as low a value
as possible.   This is accomplished by making R_1 large.   Therefore, let R_1 equal 47 kilohms.
3. Construct the d-c load line for $V1$.
The d-c load resistance for $V1$ is

$$R_1 + R_k = 47,000 + 3,300$$
$$= 50,300 \text{ ohms}$$

This is constructed on the E_b-I_b characteristic curves in Fig. 8.25. The current through $V1$ at zero bias is approximately 4.25 ma.

4. Determine E_{c1}.

For the maximum stability in gate amplitudes and gate length without resorting to precision resistors for the voltage divider determining E_{c1}, $V1$ should be operated at zero bias when conducting, and the voltage divider determining E_{c1} should have high resistance. However, the potential to which the grid of $V1$ is returned must be kept low enough to ensure that $V1$ is nonconducting when $V2$ is conducting. The allowable limits on E_{c1} to satisfy these conditions are

$$E_{c1} \leq E_{k2} + E_{co} = 28 \text{ volts}$$
$$E_{c1} \geq I_1R_k = 4.25 \times 10^3 \times 3.3 \times 10^3$$
$$= 14 \text{ volts}$$

The median value for E_{c1} as determined by the voltage divider R_7 and R_8 would therefore be approximately 21 volts.

5. Determine R_7 and R_8.

Since R_7 and R_8 should be very high compared to r_{g1}, let R_8 equal 100 kilohms. Then

$$\frac{100}{R_7 + 100} (250) = 21 \text{ volts}$$

$$R_7 = 1.09 \text{ megohms}$$

R_7 and R_8 can be 10 per cent tolerance resistors and still maintain E_{c1} within the allowable range of voltages.

6. Determine R_3.

The grid resistor R_3 for $V2$ should be high to minimize the grid current of $V2$ when conducting. The maximum value of R_3 is limited by the desired gate length and the size of C_1. The minimum size of C_1 should be somewhat larger than the input capacitance of $V2$ so that nearly the full voltage drop which appears at the plate of $V1$ will appear at the grid of $V2$. Assume R_3 equal to 1.0 megohm initially and determine the required value of C_1 for the specified 400-μsec gate length. If the value of C_1 is too small, assume a new value of R_3 and redetermine C_1.

7. Determine C_1 (see Fig. 8.14).

From Fig. 8.14

$$A = \frac{E_{b1} + E_{k2} - 2E_{bb}}{E_{k1} + E_{co} - E_{bb}}$$

where $E_{b1} = E_{bb} - I_1R_1$

$$A = \frac{50 + 40 - 500}{14 - 12 - 250} = 1.65$$

Therefore, from Fig. 8.14, T_a/R_3C_1 is equal to 0.5. The required value of C_1 is

$$C_1 = \frac{T_a}{0.5R_3}$$
$$= \frac{400 \times 10^{-6}}{0.5 \times 10^6}$$
$$= 800 \ \mu\mu f$$

This value of C_1 is many times larger than the input capacitance of $V2$. Therefore the selected value of R_3 is satisfactory. The complete circuit of the cathode-coupled multivibrator is shown in Fig. 8.26.

8.5. Astable Multivibrators. In the astable multivibrator each tube alternately conducts and is cut off for a prescribed interval of time without the use of external triggers. The circuit of a plate-coupled astable multivibrator is shown in Fig. 8.27. The plate of each tube is capacitively coupled to the grid of the opposite tube, so that each tube remains cut off only for the interval required for the charge on its grid coupling capacitor to be reduced to where the tube again conducts. Typical grid and plate waveforms for a plate-coupled astable multivibrator are given in Fig. 8.28.

The quiescent grid and plate voltages for each tube are determined by constructing the d-c load line for each tube on the E_b-I_b characteristic curves for the tube. The
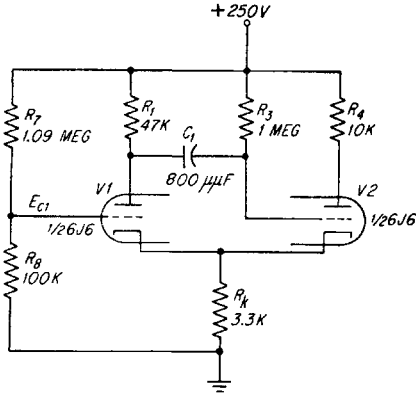
FIG. 8.26. Cathode-coupled monostable multivibrator for Example 8.3.
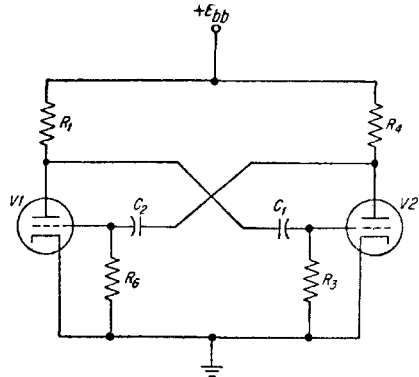
FIG. 8.27. Plate-coupled astable multivibrator.
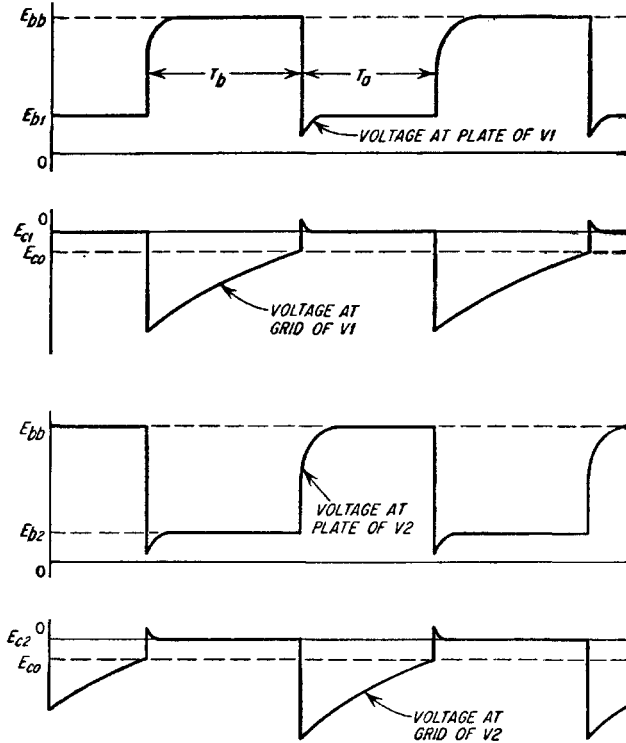
FIG. 8.28. Waveforms for a plate-coupled astable multivibrator.

quiescent plate voltage E_{b1} of $V1$ when conducting and the quiescent plate voltage E_{b2} of $V2$ when conducting are determined from the intersection of the respective load lines with the zero grid bias line.

The equivalent circuit for the discharge of C_1 when $V1$ is switched on is given in Fig. 8.29a, and the equivalent circuit for the discharge of C_2 when $V2$ is switched on is given in Fig. 8.29b.

The instantaneous voltage e_{b1} at the grid of $V1$ at time t after cutoff, assuming that R_6 is much larger than R_4 in parallel with r_{p2}, is given approximately by

$$e_{c1} = E_R + (E_{b2} - E_{bb} - E_R)e^{-t/R_6C_2} \tag{8.29}$$

(a) APPROXIMATE EQUIVALENT CIRCUIT AND INITIAL CONDITIONS FOR THE DISCHARGE OF C_1

(b) APPROXIMATE EQUIVALENT CIRCUIT AND INITIAL CONDITIONS FOR THE DISCHARGE OF C_2
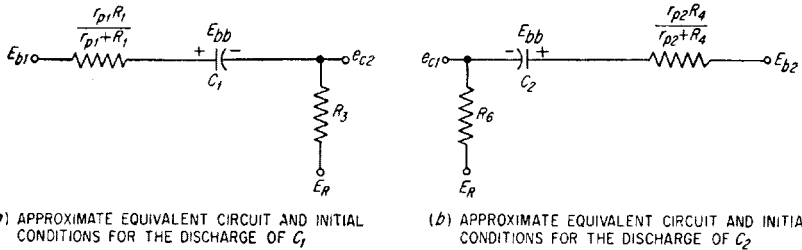
FIG. 8.29. Equivalent circuits for the discharge of the coupling capacitors in the plate-coupled astable multivibrator.

The instantaneous voltage e_{c2} at the grid of $V2$ at time t after cutoff, assuming that R_3 is much larger than R_1 in parallel with r_{p1}, is given approximately by

$$e_{c2} = E_R + (E_{b1} - E_{bb} - E_R)e^{-t/R_3C_1} \tag{8.30}$$

The time T_b that $V1$ will remain cut off is given by

$$T_b = R_6C_2 \log_e \left(\frac{E_{b2} - E_{bb} - E_R}{E_{co} - E_R} \right) \tag{8.31}$$

where E_{co} = cutoff bias voltage for $V1$
The time T_a that $V2$ will remain cut off is given by

$$T_a = R_3C_1 \log_e \left(\frac{E_{b1} - E_{bb} - E_R}{E_{co} - E_R} \right) \tag{8.32}$$

where E_{co} = cutoff bias voltage for $V1$
Equations (8.31) and (8.32) can be solved graphically from Fig. 8.14 by letting the factor A in the figure equal the logarithmic term in the equations.

(a) EQUIVALENT CIRCUIT FOR THE CHARGING OF C_1 (b) EQUIVALENT CIRCUIT FOR THE CHARGING OF C_2
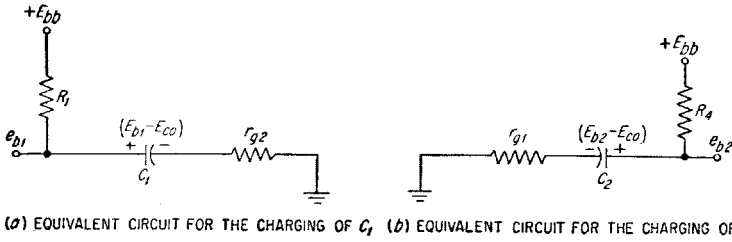
FIG. 8.30. Equivalent circuits for the charging of the coupling capacitors of a plate-coupled astable multivibrator.

The equivalent circuit for the charging of C_1 when $V1$ is cut off is given in Fig. 8.30a, and the equivalent circuit for the charging of C_2 when $V2$ is cut off is given in Fig. 8.30b. The instantaneous plate voltage e_{b1} of $V1$ at time t after cutoff, assuming R_1 to be much larger than r_{g2}, is given approximately by

$$e_{b1} = E_{bb} - (E_{bb} + E_{co} - E_{b1})e^{-t/R_1C_1} \tag{8.33}$$

The instantaneous plate voltage e_{b2} of $V2$ at time t after cutoff, assuming R_4 to be much larger than r_{g1}, is given approximately by

$$e_{b2} = E_{bb} - (E_{bb} + E_{co} - E_{b2})e^{-t/R_4C_2} \qquad (8.34)$$

The rise times T_1 and T_2 from the 0 to the 90 per cent points of the waveforms at the plates of $V1$ and $V2$ when cut off are given by Eqs. (8.35) and (8.36), respectively.

$$T_1 = R_1C_1 \log_e 10 \left(\frac{E_{bb} + E_{co} - E_{b1}}{E_{bb} - E_{b1}} \right) \qquad (8.35)$$

$$T_2 = R_4C_2 \log_e 10 \left(\frac{E_{bb} + E_{co} - E_{b2}}{E_{bb} - E_{b2}} \right) \qquad (8.36)$$

The rise time can also be obtained from Fig. 8.16.

The positive exponential in the grid waveform of each tube due to the charging current for the coupling capacitor connected to that grid flowing through the grid-cathode resistance is illustrated in Fig. 8.28. The dip in the plate waveform caused by
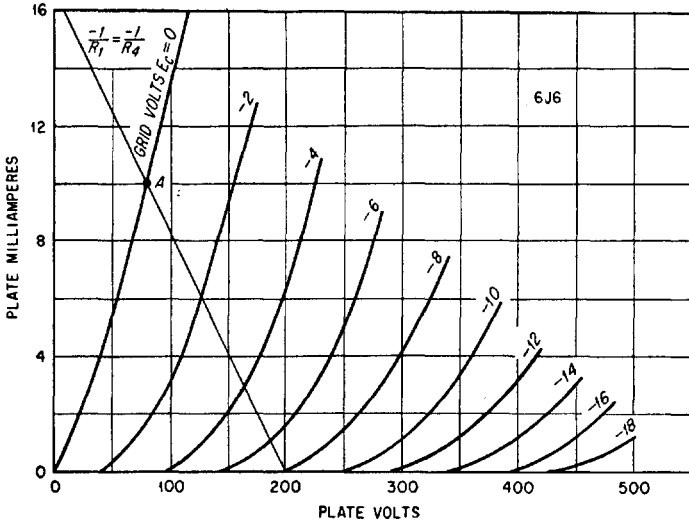
Fig. 8.31. Plate load lines for Example 8.4.

the positive exponential at the grid of the conducting tube results in a slight increase in the initial negative voltage at the grid of the cutoff tube.

Example 8.4

Design a symmetrical astable multivibrator having gate lengths of 50 μsec using a 6J6 tube. Assume a supply voltage of 200 volts. A 100-volt minimum gate amplitude is desired. The E_b-I_b characteristic curves for the 6J6 are given in Fig. 8.31.

Solution

1. Determine R_1 and R_4.
The load resistors R_1 and R_4 are determined from the considerations that (1) the gate amplitudes at the plates must have minimum values of 100 volts and (2) the larger the plate-load resistors the longer the rise times of the plate waveforms. Therefore, the load resistors are normally selected to have the smallest values which will give the necessary gate amplitudes with some safety factor for the tube selected. Assuming a design center gate amplitude of 120 volts, the plate-load resistors are determined by constructing a load line on the E_b-I_b characteristics for the 6J6 which provides an 80-volt tube drop at the zero bias

line. This is shown in Fig. 8.31. The plate-load resistances are given by

$$R_1 = R_4 = \frac{200 - 80}{0.01}$$
$$= 12,000 \text{ ohms}$$

2. Determine R_6 and R_3.

The grid resistors R_6 and R_3 should be as large as possible so that the coupling capacitors can be as small as possible and still achieve the required gate length. This is desirable since the rise time of the plate waveforms are direct functions of the values of the coupling capacitors [see Eqs. (8.35) and (8.36)]. However, the coupling capacitors should be several times larger than the input capacitance of each tube in order to couple the plate waveform of one tube to the grid of the other tube without excessive attenuation. Therefore, let $R_6 = R_3 = 1$ megohm and determine C_1 and C_2. If the values of C_1 and C_2 are less than about 30 $\mu\mu$f, R_6 and R_3 should be lowered. To minimize the time jitter of the gate lengths, R_6 and R_3 should be returned to E_{bb}.

3. Determine C_1 and C_2.

The duration of the time intervals T_b and T_a that $V1$ and $V2$ are nonconducting can be determined from either Eqs. (8.31) and (8.32) or from Fig. 8.14.

From Fig. 8.14

$$A = \frac{E_{b1} - E_{bb} - E_R}{E_{co} - E_R}$$
$$= \frac{80 - 200 - 200}{-8 - 200}$$
$$= 1.54$$

Therefore, from Fig. 8.14

$$\frac{T_a}{R_3 C_1} = \frac{T_b}{R_6 C_2} = 0.43$$

$$C_1 = C_2 = \frac{50 \times 10^{-6}}{0.43 \times 10^6}$$
$$= 116 \ \mu\mu\text{f}$$

The complete circuit is shown in Fig. 8.32.
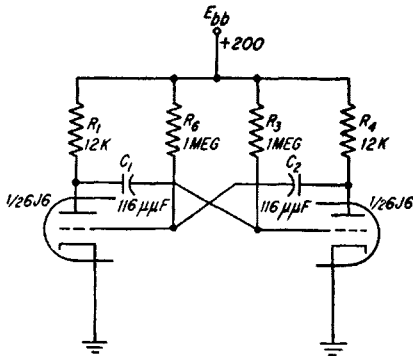
FIG. 8.32. Astable multivibrator of Example 8.4.
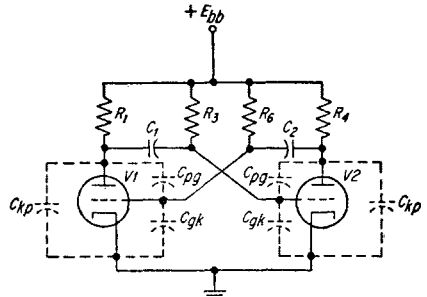
FIG. 8.33. Tube capacitances in a typical multivibrator circuit.

8.6. Effects of Tube and Stray Capacitances in Multivibrator Circuits. The tube capacitances present in a typical multivibrator circuit are shown in Fig. 8.33. Socket and wiring capacitances may effectively parallel any of these capacitances and increase the actual value from that of the tube alone. The input capacity of a resistively loaded triode is given by

$$c_i = c_{gk} + c_{gp}(1 + |A|) \tag{8.37}$$

where A = gain of stage at frequency of interest

The output impedance of a tube having the input circuit shown in Fig. 8.34a is shown in Fig. 8.34b.   The resistance R_c includes the output resistance of the previous stage in parallel with the grid resistor of the stage under consideration.   At the frequencies of interest, the value of R_c is usually considerably smaller than the capacitive reactance

(a) INPUT CIRCUIT FOR WHICH EQUIVALENT OUT-
PUT IMPEDANCE IS GIVEN

(b) APPROXIMATE OUTPUT IMPEDANCE OF TUBE
SHOWN IN (a)
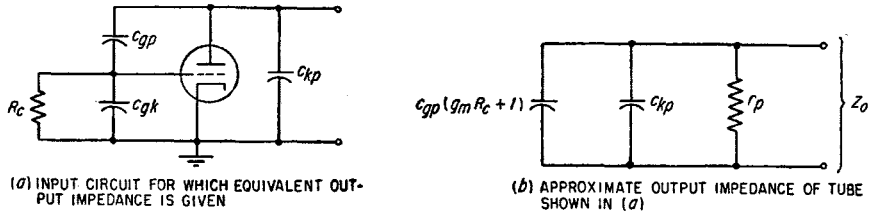
FIG. 8.34. Output impedance of a vacuum tube.

of c_{gk} which includes the output capacitance of the previous stage.   For this condition, the output admittance Y_o is given approximately by

$$Y_o \simeq \frac{1}{r_p} + j\omega(c_{gp}g_m R_c + c_{kp} + c_{gp}) \tag{8.38}$$

Since the absolute magnitude of the stage gain $|A|$ is approximately equal to $g_m R_c$, provided the following tube is of the same type, the output capacitance can be expressed approximately by

$$c_o \simeq c_{kp} + c_{gp}(1 + |A|) \tag{8.39}$$

Therefore, the shunt capacitance C_a at the plate of $V1$ and the grid of $V2$ is given approximately by

$$C_a \simeq c_{kp1} + c_{gp1}(1 + |A_1|) + c_{gk2} + c_{gp2}(1 + |A_2|) \tag{8.40}$$

Similarly, the shunt capacitance C_b at the plate of $V2$ and the grid of $V1$ is given approximately by

$$C_b \simeq c_{kp2} + c_{gp2}(1 + |A_2|) + c_{gk1} + c_{gp1}(1 + |A_1|) \tag{8.41}$$

During the switching interval, a multivibrator may be considered a two-stage amplifier having positive feedback.   The effective shunt capacitances in the circuit can be determined approximately from Eqs. (8.40) and (8.41) once the gains of each stage are known.   Since the μ, g_m, and r_p of each stage vary rapidly during the switching cycle, the shunt capacitances in the circuit also change rapidly during the switching period.   However, an approximate answer can be obtained by assuming an average value of voltage gain for each stage from the E_b-I_b characteristic curves.

In the periods from the end of the switching cycle to the points where essentially quiescent conditions exist, the rise time of the plate waveforms are determined principally by the charging of the coupling capacitors.   The rise and fall times given in Secs. 8.2 to 8.5 are modified by the presence of shunt capacitances and, therefore, are only approximations to the exact solution.   However, in most cases the results are sufficiently accurate to provide a satisfactory initial design.   In some cases, modification of the circuit values may be required to achieve the desired performance due to the effects of shunt capacitances.   In general, the same considerations apply to multivibrators as apply to video amplifiers when the generation of short-duration gates with rapid rise and fall times is required.

8.7. Triggering.   Bistable and monostable multivibrators can be triggered by applying a negative pulse to the grid of the conducting tube or a positive pulse to the grid of the nonconducting tube.   Except for the cathode-coupled monostable multi-

vibrator, the triggering pulses to either grid can be applied to the plate of the opposite tube and are coupled to the grid through the coupling capacitor between the plate and grid. The trigger amplitude and duration must be sufficient to ensure that both tubes have started the switching cycle if the regenerative switching is to continue after the triggering pulse is removed.

The best trigger source is one which is disconnected from the multivibrator as soon as the leading edge of the pulse has initiated the switching cycle. The trailing edge of the triggering pulse then does not have a tendency to reverse the switching cycle, and the trigger source does not load down the multivibrator. One method of accomplishing this is shown in Fig. 8.1. The negative triggers are coupled to the grid of the conducting tube through either diode $V3$ or $V4$. If $V1$ is conducting, the plate of diode $V3$ will be much less positive than the cathode, and the trigger will not cause $V3$ to conduct. However, the plate of $V4$ will be approximately at E_{bb}, and the trigger will couple through $V4$ and C_2 to the grid of $V1$, initiating the switching cycle. This same technique is shown applied to a monostable multivibrator in Fig. 8.7.

The minimum trigger amplitude which will cause a multivibrator to switch states is that which causes the grid voltage of the cutoff tube to rise above E_{co} and initiate plate current conduction in the tube. If a negative trigger is coupled to the grid of the conducting tube, the required minimum trigger amplitude E_t is given by

$$E_t = \frac{E_{co} - E_{cb}}{|A|} \qquad (8.42)$$

where E_{co} = cutoff bias for cutoff tube
  E_{cb} = quiescent grid voltage of cutoff tube
  A = voltage gain of conducting tube near zero bias
The trigger duration must be long enough to allow the regenerative action to be initiated in both tubes before the end of the trigger pulse. When the trigger is coupled
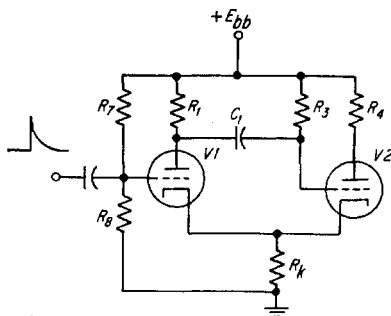
Fig. 8.35. Alternate method of triggering a cathode-coupled monostable multivibrator.

through the plate diodes as shown in Fig. 8.1, the voltage drop across the plate-load resistors of the multivibrator due to the quiescent current through the coupling resistors must be added directly to Eq. (8.42) to determine the minimum trigger amplitude which will cause the multivibrator to switch states.

Monostable multivibrators can also be triggered by a positive pulse applied to the grid of the normally cutoff tube as shown in Fig. 8.35. Whenever the trigger source is coupled to a multivibrator without an isolating diode, the source should be coupled through as small a capacitor as possible if the point in the circuit to which the trigger source is connected is one which moves rapidly during the switching period since the additional capacitance will slow down both the transition and the recovery period.

Triggers inserted in this manner should, if possible, be shaped as shown in Fig. 8.36 to minimize the negative overshoot produced when the trigger is passed through an *RC* coupling circuit having a short time constant.

FIG. 8.36. Ideal waveform for direct triggering of a multivibrator.

8.8. Synchronization. Monostable multivibrators can be used as count-down circuits or frequency dividers as shown in Fig. 8.37. The normally conducting tube is switched off by the first negative pulse applied to the grid after it has switched on because of the discharge of the voltage across the timing capacitor. All pulses applied in the interval that the normally conducting tube is cut off have no effect on the circuit, and the count-down ratio n is determined by adjusting the period of the multivibrator so that every nth pulse switches the conducting tube off. The maximum

GRID WAVEFORM OF NORMALLY CONDUCTING TUBE

INPUT TRIGGER

PLATE WAVEFORM OF NORMALLY CUTOFF TUBE
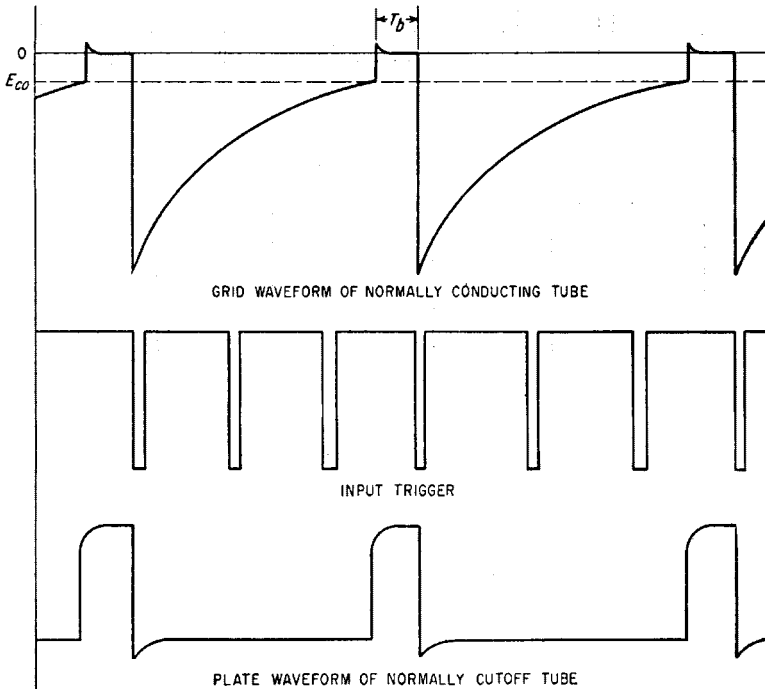
FIG. 8.37. Monostable multivibrator used as a four-to-one count-down.

count-down ratio is limited by the requirement that the timing capacitor be completely recharged in the interval T_b (see Fig. 8.37), which must be less than the interpulse period of the initiating triggers.

An astable multivibrator can be synchronized with a pulse train by applying either positive or negative pulses to the grid or plate of either tube. If negative synchronizing pulses are applied to the plate of $V1$ or the grid of $V2$, the leading edge of the positive waveform at the plate of $V2$ will be synchronized with the triggers. By adjustment of the period that $V1$ conducts, the multivibrator may be used as a count-down as previously described for the monostable multivibrator (see Fig. 8.37).

By applying positive triggers to the plate of *V*1 or the grid of *V*2, the rising edge of the waveform at the plate of *V*1 and the falling edge of the waveform at the plate of *V*2

GRID WAVEFORM OF TUBE BEING TRIGGERED

INPUT TRIGGER

PLATE WAVEFORM OF TUBE BEING TRIGGERED
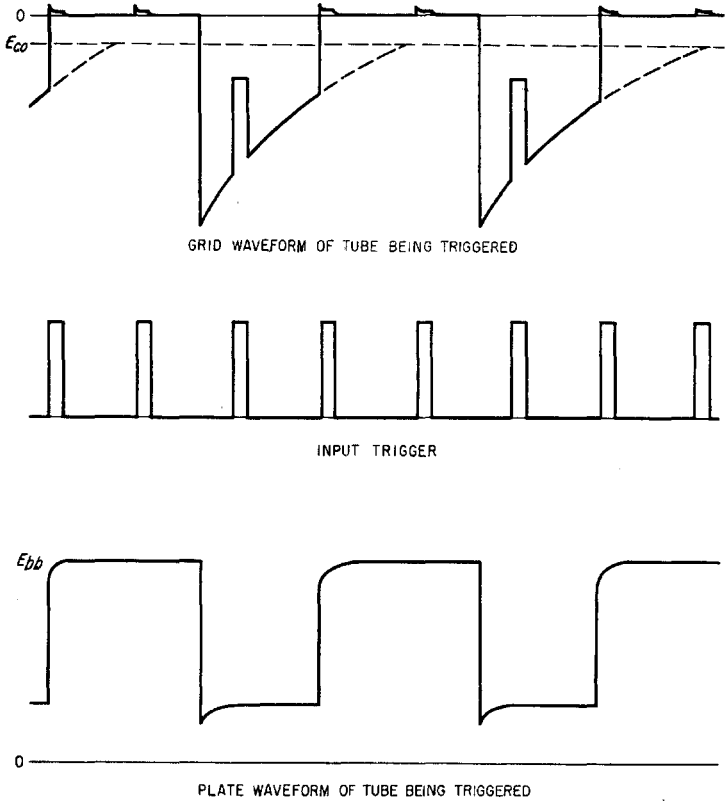
FIG. 8.38. Synchronization of an astable multivibrator as a three-to-one count-down utilizing positive triggers applied to the grid of one of the tubes.

will be synchronized with the triggers.   By adjustment of the periods that *V*1 and *V*2 conduct without triggering, the multivibrator can be used as a count-down multivibrator.   This is illustrated in Fig. 8.38.

Variable Delay Circuits

9.1. Introduction. A delay circuit is a circuit having an output gate or trigger which is delayed by a prescribed time interval from the initiating gate or trigger applied to the circuit. In variable delay circuits, the time delay between input signal and output signal is adjustable within limits by the variation of one or more circuit parameters. In general, the variation in time delay is accomplished by the variation in the magnitude of a resistance, a capacitance, or a voltage.

Any of the monostable multivibrator circuits discussed in Sec. 8.3 or 8.4 can be used as variable delay circuits by utilizing the delay period between the instant the input trigger cuts off the normally conducting tube and the instant the tube is switched back on. These periods are given for the two multivibrator circuits by Eqs. (8.14) and (8.26). In each case, the referenced intervals are directly proportional to the magnitude of the coupling capacitor between the plate of the normally cut-off tube and the grid of the normally conducting tube and to the equivalent resistance in series with the capacitor during the period that it is being discharged. The variation of the magnitude of either of these circuit components will produce a linear change in the period that the normally conducting tube is cut off.

In many instances, it is desirable to have a variable time delay which is adjustable by means of the variation of a voltage rather than a circuit component. Equations (8.14) and (8.26) indicate that the gate length of a monostable multivibrator can be changed by varying any of the reference voltages. However, with the exception of the grid reference voltage of the normally cut-off tube in a cathode-coupled monostable multivibrator, such voltage variations in a monostable multivibrator produce gate-length changes which vary linearly with the voltage changes only when the resulting variations in gate length are small percentages of the total delay and when the total delay is much less than the RC coupling time constant.

Variable delay circuits which provide a delay time very linear with respect to a control voltage are treated in Secs. 9.2 to 9.6. With the exception of the cathode-coupled monostable multivibrator, all the circuits discussed generate their time delay by means of a linear sweep circuit and a regenerative pick-off circuit.

9.2. The Screen-coupled Phantastron. In the screen-coupled phantastron circuit, a pentode is employed as a Miller sweep generator (see Sec. 11.1c). The phanta-

Fig. 9.1. Typical screen-coupled monostable phantastron.

stron has regenerative switching at the end of the linear sweep as a result of the coupling between the screen and the suppressor. The circuit of a screen-coupled phantastron is shown in Fig. 9.1. In the quiescent state, the control grid of $V1$ is very nearly at 0 volts, and the screen conducts heavily. As a result of the resistive coupling between the screen and the suppressor, the suppressor bias is held below the value required for plate current cutoff whenever the screen voltage is below some critical value,

as it is when the control grid is at zero bias. When a positive trigger is applied to the suppressor of $V1$ through $V3$, the plate current flows and the plate voltage drops. The drop in plate voltage causes a corresponding drop in control grid voltage because of the coupling capacitor C_1 between plate and grid. The lowering of the grid voltage reduces the cathode current and thereby causes a reduction in the screen current and a rise in the screen voltage. The rise in screen voltage is coupled to the suppressor through C_2 and R_2. As a result of this regenerative action, the suppressor voltage is increased until it is clamped to the potential existing at the cathode of $V4$. The plate voltage drops only a few volts below E_{bb} before the grid voltage is lowered almost to cutoff. All of the preceding takes place at time a as shown in Fig. 9.2.
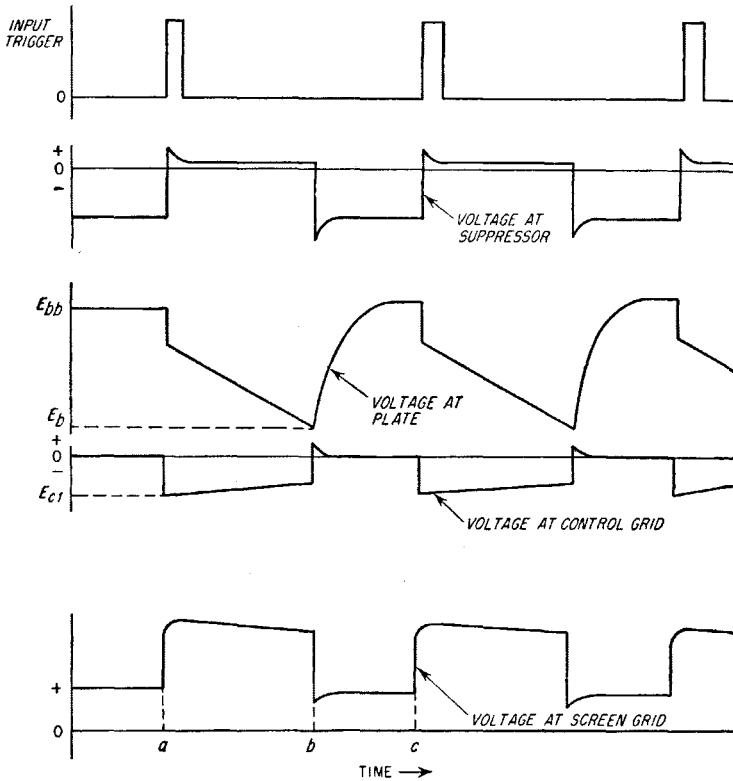
FIG. 9.2. Typical waveforms for a screen-coupled phantastron.

In the next portion of the cycle, a linear sawtooth timing waveform is generated. As a result of the regenerative action which occurred when the trigger was applied, the control grid of $V1$ was driven below zero bias; consequently, no grid current flows and the current through R_c charges C_1 toward E_{bb}. As the voltage at the grid of $V1$ rises because of the charging of C_1, the plate voltage of the tube is reduced by $A \, \Delta e_{c1}$ volts, where A is the voltage gain of $V1$ and Δe_{c1} is the magnitude of the sawtooth at the grid of the tube. As a result of the negative feedback between the plate and grid, the capacitor C_1 has an effective value, as observed at the grid, which is equal to $C_1(1 + |A|)$. Because of this effect, the effective time constant of the grid charging circuit is equal to $R_cC_1(1 + |A|)$ and can be made very large compared to the desired time delay. For this reason, the current through R_c remains very nearly constant

and provides a nearly constant rate of charge of C_1 and a very linear decrease in plate voltage as a function of time. This action continues until the tube operating point reaches the knee of the E_b-I_b characteristic curves for the tube and the plate voltage can no longer decrease with increasing control grid voltage. When this occurs, the effective time constant of the grid circuit is suddenly reduced to R_cC_1. C_1 then charges at a new rate which is faster than the original rate by the factor $|A| + 1$. The result is a sudden increase in screen current, a drop in the screen voltage, and a drop in the suppressor voltage which reduces the plate current and causes the plate voltage to rise. The rise in plate voltage is coupled through C_1 to the control grid in a regenerative switching cycle which returns the tube to its original quiescent state. The switching cycle is initiated at time b as shown in Fig. 9.2. The circuit remains in this state until another trigger is applied to the suppressor.
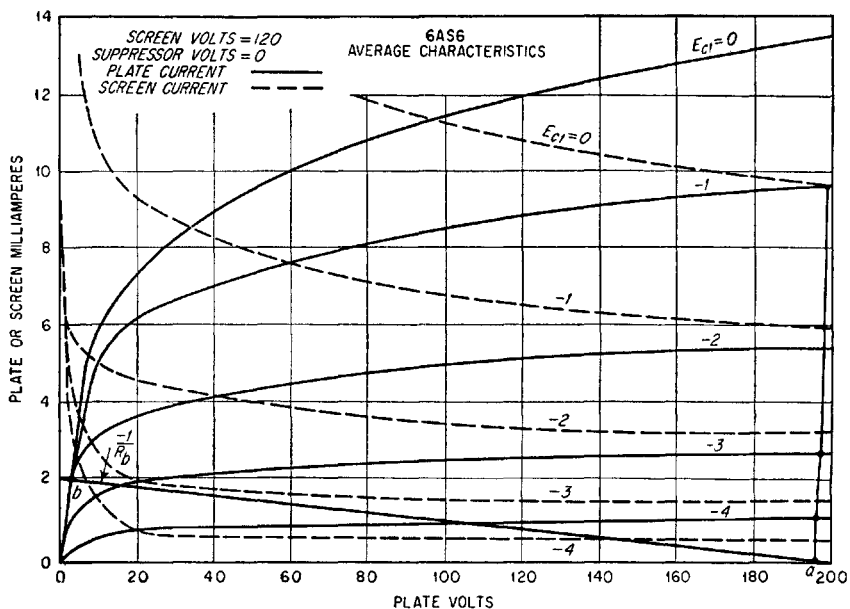
Fig. 9.3. Typical E_b-I_b characteristic curves for a pentode tube.

If the suppressor and screen voltages are established for the interval that the plate conducts, the grid and plate waveforms can be calculated from the E_b-I_b characteristic curves. With reference to Fig. 9.3, the plate-load resistor should be large enough that the load line intersects the maximum plate current curve (see Sec. 12.2d) before zero bias is reached to permit the sudden rise in grid voltage which causes the increase in screen current that initiates the regenerative switching-off action. The plate-load resistor should be small so as to allow a short recharging time for C_1 after the tube has been turned off.

When the circuit is quiescent, the grid voltage is very nearly equal to 0 volts and the plate voltage is equal to E_{bb}. At the instant the suppressor is switched on by the input trigger, both the plate and control grid voltages drop by an amount equal to E_{c1}.

$$E_{c1} = i_b R_b \qquad (9.1)$$

where E_{c1} = grid bias at which plate current i_b will flow with a plate-load resistor equal to R_t

This condition is satisfied at point a in Fig. 9.3. As C_1 is discharged, the operating point moves along the plate load line from point a to point b. The equivalent circuit for the linear discharge of C_1 is given in Fig. 9.4. The initial current i_c through R_c after the suppressor is switched on is given by

$$i_c = \frac{E_{bb} - E_{c1}}{R_c} \qquad (9.2)$$

FIG. 9.4. Approximate equivalent circuit and initial conditions for the linear discharge of C_1.

As this current flows into C_1, the instantaneous grid voltage tends to rise toward E_{bb}. However, an increase in the grid voltage equal to Δe_{c1} causes a decrease Δe_b in the plate voltage equal to

$$\Delta e_b = -A \, \Delta e_{c1} \qquad (9.3)$$

Thus, an increase in e_{c1} results in a much larger decrease in e_b, which tends to increase the current i_c into C_1 and lower e_{c1}. The result is that the magnitude of the capacitor C_1 appears to be increased by the factor $1 + |A|$. The instantaneous plate voltage e_b is approximately given by

$$e_b = E_{bb} + E_{c1} - |A|(E_{bb} - E_{c1})(1 - e^{-t/R_c C_1(|A|+1)}) \qquad (9.4)$$

where $A \simeq \dfrac{-\mu R_b}{r_p + R_b}$

$\simeq -g_m R_b$

t = time in seconds from instant that plate current is switched on

Since E_{bb} is usually much larger than E_{c1} and since A is usually much larger than unity, the slope of the plate waveform is very nearly equal to $-E_{bb}/R_c C_1$ volts/sec until plate bottoming occurs. For typical circuit values, the linearity of the negative sawtooth can be expected to be in the order of 0.1 per cent.
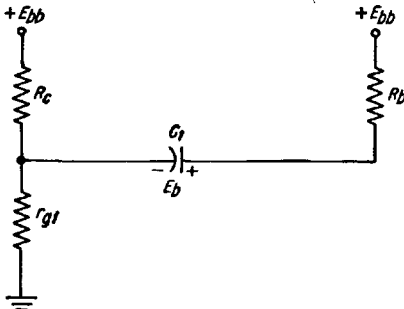
After the regenerative switching cycle has cut the plate current off at the end of the plate sawtooth, the plate voltage rises exponentially toward E_{bb} as C_1 is recharged. The equivalent circuit for the charging of C_1 is shown in Fig. 9.5. The instantaneous plate voltage e_b from time b to time c in Fig. 9.2 is approximately given by Eq. (9.5), assuming that the control grid-to-cathode resistance r_{g1} of $V1$ when the grid is slightly positive is much smaller than either R_c or R_b.

FIG. 9.5. Equivalent circuit for the charging of C_1.

$$e_b = E_{bb} - (E_{bb} - E_b)e^{-t/R_b C_1} \qquad (9.5)$$

where E_b = plate voltage at termination of plate sawtooth waveform

The value of E_b is approximately equal to the plate voltage at the intersection of the plate load line and the zero grid bias curve. This is illustrated as point b in Fig. 9.3. The recovery time T_r required for the plate voltage to rise within 1 per cent of E_{bb} is given by

$$T_r \simeq 4.6 R_b C_1 \qquad (9\,6)$$
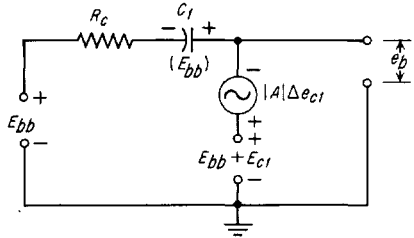
The grid-voltage waveform during the recovery of the plate is given very closely by

$$e_{c1} \simeq \frac{r_{g1}}{R_b + r_{g1}} (E_{bb} - E_b)e^{-t/R_bC_1} \tag{9.7}$$

where t = time in seconds from instant that plate current is cut off

The duration of the time interval T_a between a and b in Fig. 9.2 is determined by the time required for the plate voltage rundown from point a to point b in Fig. 9.3. If the potential to which the plate is returned is made equal to some reference potential E_r by means of a diode as shown in Fig. 9.1, the plate rundown will start from E_r instead of E_{bb} and the duration of the period T_a that the plate conducts will be given by

$$T_a = R_cC_1 \frac{E_r - E_b}{E_{bb}} \tag{9.8}$$

The length of the interval that the plate conducts is thus directly proportional to the reference voltage E_r minus a constant and, consequently, provides a convenient means
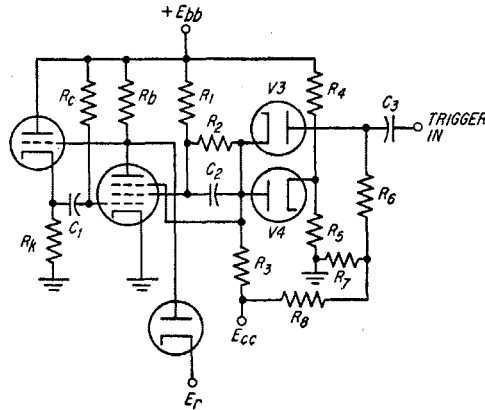
FIG. 9.6. Screen-coupled phantastron utilizing a cathode follower for rapid charging of C_1.

of obtaining a time-delay gate which varies linearly with a d-c control voltage. The time-delay waveform can be taken from the screen or suppressor if a positive gate is desired or from the grid if a small negative gate is desired.

The recovery time of a phantastron can be greatly reduced by charging C_1 from a cathode follower whose grid is directly coupled to the plate of $V1$ as shown in Fig. 9.6. By this method, the recovery time can be reduced to a few microseconds even though the plate rundown period is of several milliseconds duration.

During the interval of plate conduction, the instantaneous grid voltage e_{c1} is given by

$$e_{c1} = E_{c1} + (E_{bb} - E_{c1})(1 - e^{-t/R_cC_1(|A|+1)}) \tag{9.9}$$

This waveform is shown in Fig. 9.2.

For the regenerative switching action at the beginning and end of the plate conduction period to be as rapid as possible, it is necessary to place a small coupling capacitor C_2 between the screen and suppressor to cause the suppressor to follow the rapid changes in screen voltage. C_2 should be five to ten times the value of the shunt capacitance at the suppressor to ensure that essentially the full change in the screen voltage will appear at the suppressor. The coupling capacitor should not be so large that it cannot be completely discharged during the recovery period of the phantastron. Typical values of C_2 are usually between 22 $\mu\mu$f and 100 $\mu\mu$f.

The quiescent suppressor voltage should be at or below the value required for plate current cutoff. The quiescent screen voltage should be set at a value low enough that the suppressor voltage can be varied by the coupling between the screen and the suppressor from below cutoff to above the desired reference voltage when the screen voltage is changed from its quiescent value to its value during the period of plate conduction. The suppressor voltage during plate conduction is clamped at the desired reference level by the diode $V4$ and the voltage divider formed by R_4 and R_5 in Fig. 9.1. By establishing the suppressor voltage by this means rather than by the voltage divider formed by R_1, R_2, and R_3, the variations in the suppressor voltage from tube to tube due to variations in screen characteristics can be minimized. For ease of design, the screen and suppressor voltages are usually set at values for which the tube manufacturers have published characteristic curves. However, these voltages are not critical.

Example 9.1

Design a voltage-variable time-delay circuit of the screen-coupled phantastron type utilizing a 6AS6 pentode. The time delay should be variable from 50 to 800 μsec and the interpulse period is to be 1,000 μsec. Assume that a plate supply of 200 volts and a bias supply of -150 volts are available. The plate characteristics of the 6AS6 are given in Fig. 9.3.

Solution

1. Determine the screen and suppressor voltages.
To use the manufacturer's tube data conveniently, let the screen voltage be equal to 120 volts during the interval that plate current flows. With a screen voltage of $+120$ volts, cutoff for the suppressor of a 6AS6 is about -15 volts. Therefore let the suppressor voltage be equal to -20 volts during the interval that the plate current is cut off. Also, for convenience, let the suppressor voltage be equal to 0 volts when plate current flows. The cathode of $V4$ can then be grounded and the resistors R_4 and R_5 in Fig. 9.1 eliminated. The screen voltage during plate current cutoff should be made equal to or less than 90 volts so that the suppressor may be varied by the required 20 volts.
2. Select a tentative value of R_b.
From Fig. 9.3 it is seen that a plate-load resistance of 100 kilohms will provide a sharp rise in screen current when the plate voltage bottoms at the zero grid bias curve. Therefore, assume a tentative value for R_b equal to 100 kilohms.
3. Determine the value of initial grid bias when the suppressor is raised to 0 volts.
The curve of $E_{bb} - |E_{c1}|$ versus E_{c1} is plotted on the E_b-I_b characteristic curves in Fig. 9.3. The intersection of this curve and the plate load line at point a is the initial operating point during plate conduction. From Fig. 9.3, the value of E_{c1} is -4.5 volts.
4. Determine R_c and C_1 from Eq. (9.8).
At the maximum time delay of 800 μsec, the reference voltage E_r is set equal to E_{bb}. From Fig. 9.3, the voltage E_b is found at point b to be approximately 4 volts. Therefore, from Eq. (9.8),

$$R_c C_1 = \frac{T_a E_{bb}}{E_r - E_b}$$
$$= \frac{800 \times 10^{-6} \times 200}{200 - 4}$$
$$= 816 \ \mu\text{sec}$$

The time required to very nearly recharge C_1 when the plate current is cut off is given by Eq. (9.6).

$$T_r \simeq 4.6 \ R_b C_1$$

The available recovery time is 200 μsec, therefore

$$R_b C_1 \leq \frac{200 \times 10^{-6}}{4.6}$$
$$\leq 43.5 \times 10^{-6}$$

If R_b is 100 kilohms (from step 2),

$$C_1 \leq \frac{43.5 \times 10^{-6}}{10^5}$$
$$\leq 435 \ \mu\mu f$$

Let $R_c = 2.2$ megohms; then

$$C_1 = \frac{816 \times 10^{-6}}{2.2 \times 10^6}$$
$$= 371 \ \mu\mu f$$

5. Determine the required voltage range for E_r.

When the time delay equals 800 μsec, $E_r = 200$ volts. The minimum time delay is 50 μsec. From Eq. (9.8)

$$E_r = E_b + \frac{T_a E_{bb}}{R_c C_1}$$
$$= 4.0 + \frac{50 \times 10^{-6} \times 200}{816 \times 10^{-6}}$$
$$= 16.3 \text{ volts}$$

Therefore, the reference voltage E_r must be adjustable from 16.3 volts to 200 volts.

6. Determine R_1, R_2, and R_3.

Since the published tube characteristic curves are usually not complete enough to allow a determination of the screen voltage and current which will exist when the plate current is cut off, the values of R_1, R_2, and R_3 are most readily determined experimentally. R_2 and R_3 should be adjusted so that the suppressor voltage is somewhat greater than zero when the screen voltage is equal to 120 volts. R_1 should have a value such that the variation in screen current between the two states produces a sufficiently large change in the suppressor voltage to ensure plate current cutoff during the quiescent condition of operation.

Summary

The complete circuit is shown in Fig. 9.7. The time-delay gate can be taken from the screen or the suppressor. C_2 should have a value of approximately 50 $\mu\mu f$. The bias on
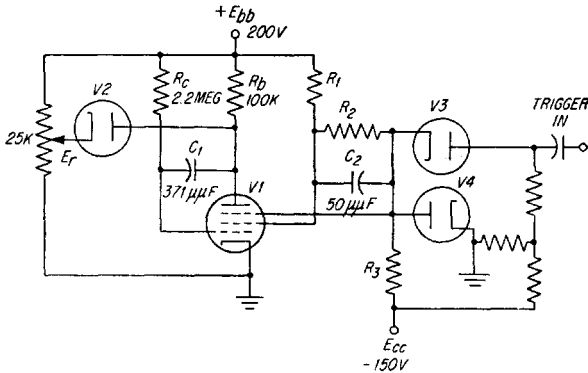
Fig. 9.7. Circuit of the screen-coupled phantastron of Example 9.1.

the plate of $V3$ should be -20 volts or slightly higher so that the suppressor will remain below cutoff prior to a trigger being applied to the plate of $V3$. A trigger amplitude of about 20 volts will trigger the circuit satisfactorily.

9.3. Cathode-coupled Phantastron. In the screen-coupled phantastron, the regenerative action is supplied by the RC coupling between the suppressor grid and screen grid. The same regenerative action can be obtained by inserting a resistance in the cathode circuit. The circuit of a typical cathode-coupled phantastron is shown in Fig. 9.8. In the quiescent state, the control grid is clamped to a positive reference voltage through diode $V2$, and the suppressor is returned to a voltage lower than the cathode by an amount sufficient to cut off the plate current. The screen current is

relatively high because the grid is near zero bias and the suppressor is biased beyond cutoff. A positive trigger applied to the suppressor will cause plate current to flow. The resultant drop in plate voltage is coupled to the control grid through C_1, causing the control grid voltage to drop. The drop in control grid voltage causes a decrease in total space current, which causes a drop in cathode voltage, which in turn causes the suppressor voltage to further increase the plate current, completing the regenerative switching action. The plate rundown as described for the screen-coupled phantastron in Sec. 9.2 then occurs. It is terminated when the plate voltage bottoms at the knee of the bias curves of the tube characteristics. When the plate voltage can no longer decrease with increasing grid voltage, the grid voltage rises at a much higher rate. This increases the space current and hence the cathode voltage, resulting in an increase in the suppressor bias and a reduction in plate current. The increase in plate voltage is coupled to the control grid through C_1, completing the regenerative switching action which cuts off the suppressor and returns the tube to its initial quiescent condition. The duration of the plate rundown is controlled by the initial plate voltage E_r which is established by the diode $V3$. Adjustment of the reference voltage E_r to which $V3$ is returned

FIG. 9.8. Typical cathode-coupled phantastron.

provides a linear adjustment of the duration of the plate rundown. The waveforms for the cathode-coupled phantastron are illustrated in Fig. 9.9. The screen provides a positive output gate which is free from coupling to other tube electrodes. A negative gate can be obtained from the cathode.

If R_c is much larger than R_1, as is normally the case, the quiescent tube voltages are determined as follows. The quiescent control grid voltage E_{c1} is given by

$$E_{c1} = \frac{R_2 + R_3}{R_1 + R_2 + R_3} E_{bb} \tag{9.10}$$

The quiescent suppressor voltage E_{c3} is given by

$$E_{c3} = \frac{R_3}{R_1 + R_2 + R_3} E_{bb} \tag{9.11}$$

The quiescent cathode voltage E_1 is given by

$$E_k = I_{c2}R_k \tag{9.12}$$

where I_{c2} is that value of screen current which flows with the established values of grid voltage, cathode resistance, and screen resistance when the plate current is cut off. Sufficient tube data are normally not supplied by tube manufacturers to allow the determination of this current, and it must be found experimentally. The cathode voltage must be sufficiently higher than the suppressor voltage given in Eq. (9.11) that the plate current is cut off and slightly higher than E_{c1} as determined from Eq. (9.10).

The screen dropping resistor R_s must be large enough to limit the screen dissipation in the quiescent condition to less than the maximum permissible value.
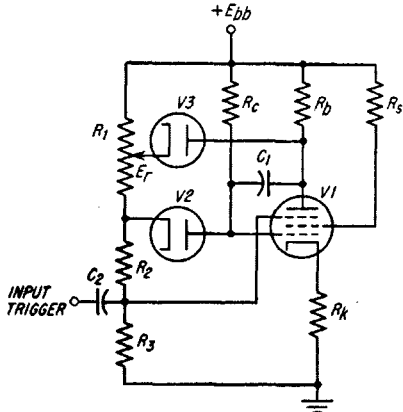
When a positive trigger is applied to the suppressor and the regenerative switching action occurs, the plate, cathode, and control grid voltages simultaneously drop until the plate current which flows is just sustained by the resulting grid bias. The published tube characteristics do not ordinarily contain sufficient information to enable the calculation of this voltage drop. For the 6AS6 pentode and plate-load resistances over approximately 50,000 ohms, the initial grid-voltage drop will typically be to a level between 4 and 5 volts below the cathode voltage.
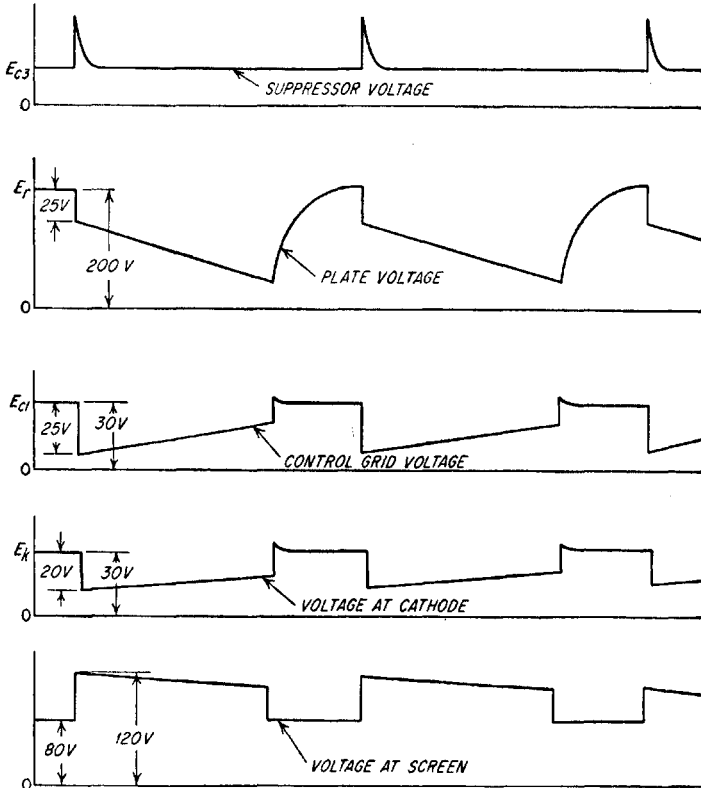
Fig. 9.9. Waveforms of a typical cathode-coupled phantastron.

As the grid voltage drops to its initial level at the start of the plate conduction interval, the total cathode current must drop to a level which will allow the suppressor-to-cathode voltage to be above cutoff. For stabilization of the suppressor-cathode voltage, the cathode voltage during the plate conduction interval should drop to a level less than the suppressor voltage as given by Eq. (9.11) so that the suppressor will be clamped to the cathode by suppressor conduction.

During the plate conduction period, the plate voltage decreases as the grid voltage rises, resulting in the linear reduction in plate voltage with time as discussed for the screen-coupled phantastron. The effect of the cathode resistor is to reduce the effective gain of the stage and thereby slightly reduce the linearity of the plate waveform. The instantaneous plate voltage at time t after the initiation of plate conduction is given approximately by

$$e_b = E_{bb} - I_b R_b - |A|(E_{bb} - E_{c1} + I_b R_b)(1 - e^{-t/R_c C_1(|A|+1)}) \qquad (9.13)$$

where I_b = initial value of plate current

When A is much larger than unity and E_{bb} is much larger than $(E_{c1} - I_b R_b)$, the plate waveform is very nearly equal to $-E_{bb}/R_c C_1$ volts/sec for the full plate-voltage rundown. The minimum value of the plate-to-cathode voltage for the 6AS6 tube with a plate-load resistance of 50,000 ohms or greater is approximately 10 volts or less.

At the end of the plate rundown, the instantaneous plate-voltage rise toward E_{bb} (until $e_b = E_r$) at time t after plate current cutoff is given by

$$e_b = E_{bb} - \frac{R_b}{R_b + R_x} (E_{bb} - E_b)e^{-t/(R_b + R_x)C_1} \qquad (9.14)$$

where $R_x = \dfrac{R_1(R_2 + R_3)}{R_1 + R_2 + R_3}$

E_b = plate-to-ground voltage at end of plate voltage rundown, i.e., plate-to-ground voltage at plate bottoming

Assuming that $E_{bb} \gg (E_{c1} - I_b R_b)$, the time duration T_a of the plate rundown is given by

$$T_a = R_c C_1 \left(\frac{E_r - E_b}{E_{bb}} \right) \qquad (9.15)$$

The time interval T_a can be varied linearly as a function of the change in reference voltage E_r to which the plate is clamped by diode $V3$.

The advantages of the cathode-coupled phantastron are that (1) no bias supply is required, (2) no connections are made to the screen, thus allowing better rise time on the screen waveform, and (3) a negative gate may be taken from the cathode. The principal disadvantage is the increased design complexity.

9.4. The Sanatron. The phantastron circuits previously described are usually limited in producing very fast plate-voltage rundowns and, therefore, very short duration gates because of the screen current limitation of the tubes, such as the 6AS6, which are applicable to such a circuit. The reduction of R_c to achieve a larger discharge current for C_1 in the screen-coupled phantastron circuit results in an increase in the quiescent screen dissipation because of the significant increase in screen current with an increase in the positive grid bias. In the sanatron circuit, tubes capable of higher screen dissipations, such as the 6AG7, but having poorer suppressor control characteristics than the 6AS6, can be used to produce very short gate waveforms which have time durations which are a linear function of a reference voltage.

The circuit of a sanatron is shown in Fig. 9.10. The tube $V1$ is essentially the same as the amplifier in a screen-coupled phantastron. The grid of $V1$, however, is directly coupled to the grid of $V2$. Tube $V2$ acts as an amplifier to drive the suppressor of $V1$ from plate-current cutoff to zero bias or to a positive value of suppressor voltage. The plate of $V2$ is directly coupled to the suppressor of $V1$, and waveforms of 50 to 100 volts in amplitude can be applied to the suppressor in this fashion. $V2$ conducts

FIG. 9.10. Circuit of a sanatron.

during the quiescent state of $V1$. However, $V2$ is usually cut off during the plate rundown period of $V1$ by having a lower screen voltage for $V2$ or by making $V2$ a tube which has a less negative grid cutoff voltage and which is, therefore, cut off by the initial drop in grid voltage at the start of the plate rundown period.

The sanatron is triggered by a negative trigger applied to the suppressor of $V2$. A positive output waveform can be taken from the plate of $V2$ at a relatively high amplitude. Since the principal application of the sanatron is in the generation of short-duration variable delays, shunt capacities in the plate circuit of $V2$ and suppressor circuit of $V1$ must be minimized, and the circuit must be designed with a small value of load resistance for $V2$ if the gate rise time is to be short.

9.5. The Sanaphant. The sanaphant is another linear time-delay circuit and is very similar to the sanatron previously described. The circuit of a sanaphant is illustrated in Fig. 9.11. The sanaphant is basically different from the sanatron in the manner in which the grid waveform for $V2$ is obtained. In the sanaphant, the grid waveform for $V2$ is taken from the cathode of $V1$ instead of the grid as in the sanatron.
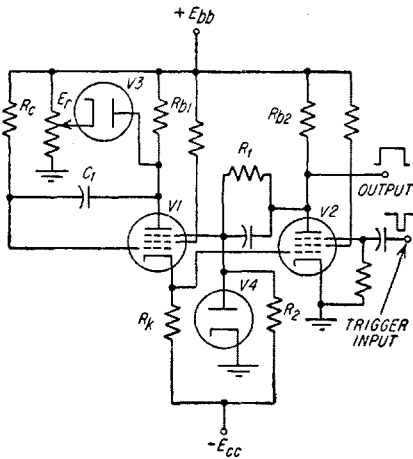
Fig. 9.11. Sanaphant variable delay circuit.
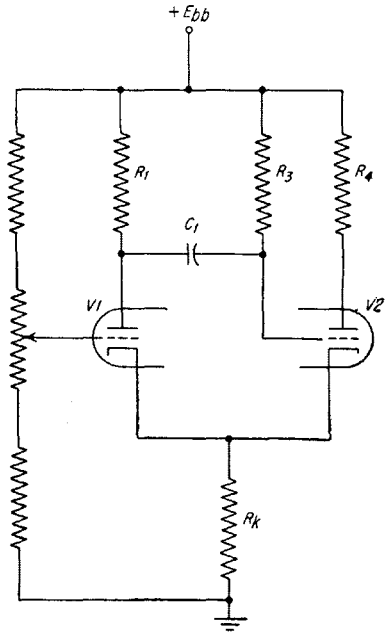
Fig. 9.12. Cathode-coupled monostable multivibrator.

As is the case for the sanatron, the sanaphant is used to produce variable short time delays which cannot be produced satisfactorily by the phantastron. The sanaphant possesses the advantage over the sanatron of having a low impedance source to drive the grid of $V2$, thus minimizing the effect of shunt capacitances at the grid of $V2$.

9.6. Cathode-coupled Monostable Multivibrator. A nearly linear relationship exists between the gate length of the cathode-coupled monostable multivibrator described in Sec. 8.4 and the reference potential at the grid of the normally cut-off tube $V1$. As a result, the circuit is very useful as a voltage-variable time delay.

The circuit of a cathode-coupled monostable multivibrator is given in Fig. 9.12. The period T_a that the normally cut-off tube conducts is approximately given by

$$T_a = (R + R_3)C_1 \log_e \left[\left(\frac{E_{b1} + E_{k2} - E_r - E_{bb}}{E_{k1} + E_{co} - E_r} \right) \frac{R_3}{R + R_3} \right] \qquad (9.16)$$

where T_a = time, sec

E_{b1} = quiescent voltage at plate of $V1$ when conducting
   = $E_{bb} - I_1R_1$

E_{k2} = cathode voltage when $V2$ is conducting

E_r = potential to which R_3 is returned

E_{k1} = cathode voltage when $V1$ is conducting

E_{co} = cutoff bias for $V2$

E_{bb} = plate supply voltage

I_1 = plate current through $V1$ when conducting

R_3 = grid resistor of $V2$

$$R = \frac{R_1(r_{p1} + \mu R_k)}{R_1 + r_{p1} + \mu R_k}$$

The period T_a that the normally cut-off tube $V1$ conducts is shown graphically in Fig. 9.13. For simplicity, R has been assumed to be negligible compared to R_3. If it is also assumed that the plate current I_1 through $V1$ when conducting varies linearly as a function of the grid-cathode voltage e_{gk1} of $V1$, the desired linear relationship between gate length T_a and the grid-to-ground voltage can be achieved by causing T_a to vary linearly as a function of I_1. Examination of Fig. 9.13 shows that both the voltage E_x on the grid of $V2$ at the start of the timing period and the voltage E_y at the end of the timing period are functions of I_1. Because the voltage rise at the grid of $V2$ is exponential, the gate length T_a does not increase linearly as E_y alone is increased. However, by causing E_x to become more negative by the correct amount as E_y is increased, the effect of this nonlinearity on the gate length T_a can be almost exactly compensated for and T_a can be made to vary almost linearly with I_1. Good compensation can be achieved for gate lengths up to approximately $0.1(R_3 + R)C_1$. The relationships for achieving the optimum com-

FIG. 9.13. Waveform at the grid of the normally conducting tube of a cathode-coupled monostable multivibrator when the grid is cut off.

pensation are determined by finding the conditions for which T_a varies linearly with changes in I_1 in Eq. (9.16). Therefore, if Eq. (9.16) is rewritten as a linear function of I_1,

$$\frac{T_a}{(R + R_3)C_1} = KI_1 + W = \log_e \left(\frac{A + I_1 R_1}{B - I_1 R_k}\right) k_1 \qquad (9.17)$$

where $A = E_r - E_{k2}$ (9.18)

$B = E_r - E_{co}$ (9.19)

K = constant representing slope of $\dfrac{T_a}{(R + R_3)C_1}$ versus I_1

W = theoretical initial value of $\dfrac{T_a}{(R + R_3)C_1}$ when $I_1 = 0$

I_1 = current through $V1$ when conducting

$k_1 = \dfrac{R_3}{R_3 + R}$

Equation (9.17) can be rewritten as

$$e^{KI_1 + W} = e^{KI_1} \cdot e^W = \frac{A + I_1 R_1}{B - I_1 R_k} k_1 \qquad (9.20)$$

$$e^{KI_1} = \frac{A + I_1 R_1}{B - I_1 R_k} k_1 Z \qquad (9.21)$$

where $Z = e^{-W}$

By expanding e^{KI_1} in a series and utilizing the first three terms in place of e^{KI_1} in Eq. (9.21), the value of R_1 which provides the best approximation to a linear time delay as a function of I_1 can be obtained for any given values of A, B, and R_k. The resulting equations are

$$R_1 = \frac{AR_k}{B} \qquad (9.22)$$

$$K = \frac{2R_k}{B} \qquad (9.23)$$

$$W = \log_e \frac{Ak_1}{B} \qquad (9.24)$$

The time delay is given approximately by

$$T_a \simeq (R_3 + R)(KI_1 + W)C_1 \qquad (9.25)$$

The value of T_a is negative when I_1 is zero since $k_1 A/B$ is always less than unity. Therefore, there is a particular value of I_1 which satisfies the condition for T_a equal to zero. The value of I_1 for which T_a is equal to zero is given by

$$I_{1(min)} = \frac{B - Ak_1}{R_1 k_1 + R_k} \qquad (9.26)$$

For a maximum gate length T_a equal to $0.1(R + R_3)C_1$, the maximum plate current $I_{1(max)}$ needed through $V1$ is given by

$$I_{1(max)} = I_{1(min)} + \frac{0.1}{K} \qquad (9.27)$$

The maximum available plate current I_1 and, consequently, the maximum delay T_a that can be achieved is determined by the d-c plate load line for $V1$ together with Eq. (9.28).

$$E_{c1(max)} = E_{k2} + E_{co} \qquad (9.28)$$

where $E_{c1(max)}$ = maximum grid-to-ground voltage that can be placed on grid of $V1$ and still hold it at cutoff with $V2$ conducting

E_{k2} = cathode voltage with $V2$ conducting

    = $I_2 R_k$

E_{co} = cutoff bias for $V1$

In practice, it is found that deviations as great as 2 to 1 from the optimum value of R_1 as determined from Eq. (9.22) do not appreciably deteriorate the linearity of the time delay versus I_1.

Example 9.2

Design a linear voltage-variable time-delay circuit using a 6J6 in a cathode-coupled multivibrator circuit. The delay should be variable from 0 to 200 μsec. Assume a +250-volt plate supply. The plate characteristics of a 6J6 are given in Fig. 9.14.

Solution

1. Let $R_4 = 10$ kilohms

    $R_k = 10$ kilohms

2. Determine the cathode voltage E_{k2}.

Construct a 20-kilohm load line on the E_b-I_b characteristic curves of the tube as shown in Fig. 9.14. Assuming R_3 to be large, the grid bias of $V2$ when conducting will be essentially 0 volts. The plate current through $V2$ is, therefore, 9.4 ma, and the cathode voltage E_{k2} is

$$E_{k2} = 9.4 \times 10^{-3} \times 10^4 = 94 \text{ volts}$$

3. Determine A.

From Eq. (9.18)

$$A = E_r - E_{k2}$$

Assuming that R_3 is returned to the plate supply,

$$A = 250 - 94 = 156 \text{ volts}$$

4. Determine B.
From Eq. (9.19)

$$B = E_r - E_{co}$$

From Fig. 9.14, E_{co} is seen to be -10 volts. Therefore,

$$B = 250 + 10 = 260 \text{ volts}$$
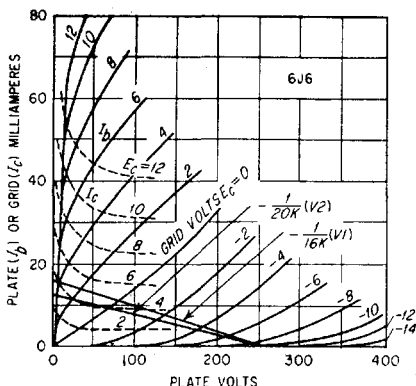
Fig. 9.14. Tube characteristics for Example 9.2.

5. Determine K.
From Eq. (9.23)

$$K = \frac{2R_k}{B}$$

$$= \frac{2 \times 10^4}{260}$$

$$= 76.9$$

6. Determine R_1.
From Eq. (9.22)

$$R_1 = \frac{AR_k}{B}$$

$$= \frac{156 \times 10^4}{260}$$

$$= 6{,}000 \text{ ohms}$$

7. Determine R_3 and C_1.
The time constant R_3C_1 should be equal to or greater than ten times the maximum desired gate length. Therefore

$$R_3C_1 \geq 10 \times 200$$
$$\geq 2{,}000 \ \mu\text{sec}$$

Let $R_3 = 1$ megohm
  $C_1 = 2{,}000 \ \mu\mu\text{f}$
8. Determine $I_{1(min)}$.
From Eq. (9.26)

$$I_{1(min)} = \frac{B - Ak_1}{R_1k_1 + R_k}$$

and

$$k_1 = \frac{R_3}{R_3 + R}$$

where R is defined in Eq. (9.16) as

$$R = \frac{R_1(r_{p1} + \mu R_k)}{R_1 + r_{p1} + \mu R_k}$$

Since $R_3 \gg R_1$,

$$k_1 \simeq 1$$

Therefore,

$$I_{1(min)} \simeq \frac{260 - 156}{6{,}000 + 10{,}000}$$

$$\simeq 6.50 \text{ ma}$$

9. Determine $I_{1(max)}$.

From Eq. (9.27)

$$I_{1(max)} = I_{1(min)} + \frac{0.1}{K}$$

$$= 6.50 \times 10^{-3} + \frac{0.1}{76.9}$$

$$= 6.50 \times 10^{-3} + 1.3 \times 10^{-3}$$

$$= 7.8 \text{ ma}$$

10. Check to ensure that, when the grid voltage of $V1$ is adjusted to provide the current $I_{1(max)}$ when $V1$ is conducting, the tube $V1$ is cut off when $V2$ is conducting.

Construct the load line for $V1$ as shown in Fig. 9.14. The grid bias required to achieve a plate current of 7.80 ma (for maximum delay) in $V1$ is seen from Fig. 9.14 to be approximately −2.0 volts. Therefore, the required grid-to-ground voltage E_{c1}' is

$$E_{c1}' = -2.0 + 7.80 \times 10^{-3} \times 10^4$$

$$= 76.0 \text{ volts}$$

From Eq. (9.28), the maximum value of the grid-to-ground voltage $E_{c1(max)}$ is

$$E_{c1(max)} = E_{k2} + E_{co}$$

$$= 94 - 10$$

$$= 84 \text{ volts}$$

Therefore, the value of $I_{1(max)}$ determined in step 9 is satisfactory.

11. Determine the range of E_{c1} required to achieve the gate length variation.

From step 10,

$$E_{c1}' = 76.0 \text{ volts}$$

From Fig. 9.14, a plate current of 6.50 ma (zero time delay) is achieved with a grid bias of approximately −2.6 volts. Therefore, the minimum required grid-to-ground voltage E_{c1}'' is

$$E_{c1}'' = -2.6 + 6.50 \times 10^{-3} \times 10^4$$

$$= 62.4 \text{ volts}$$

12. Check the linearity of T_a versus E_{c1}.

Since the range of plate current variation is quite small, the linearity of I_1 versus e_{gk1} should be very good. By assuming this relationship to be exactly linear, the linearity of T_a versus E_{c1} can be determined by calculating T_a versus I_1 from Eq. (9.17) and plotting
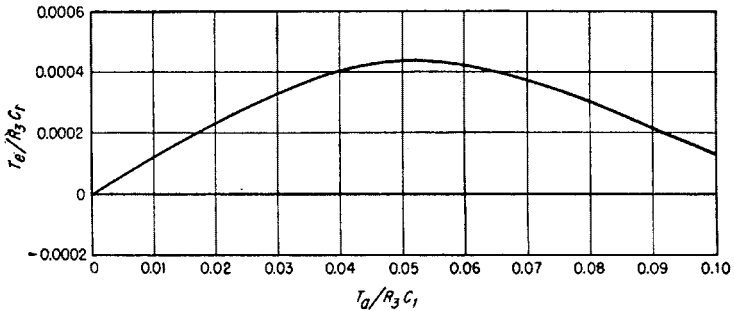
FIG. 9.15. Deviation T_e of time delay from perfect linearity for circuit of Example 9.2.

the deviation T_e from exact linearity as a function I_1. This is plotted in Fig. 9.15, assuming $k_1 = 1$ and $R_3 \gg R$. The maximum deviation from linearity, expressed as a per cent of the maximum delay period, is found from Fig. 9.15 to be

$$\% \text{ deviation} = \frac{T_{e(max)}/R_3 C_1}{T_{a(max)}/R_3 C_1} \times 100$$

$$= \frac{0.00044}{0.1}$$

$$= 0.44\%$$

NOTE: The linearity could be improved by making T_a smaller with respect to $(R + R_3)C_1$

Trigger Circuits

10.1. Introduction. Trigger circuits are used to initiate action in other circuits, e.g., to start a sawtooth sweep, to initiate a multivibrator gate, etc. The important waveform and circuit characteristics are the trigger amplitude, rise time, duration, temporal stability, and the circuit output impedance. Trigger generators can be either self-initiating, i.e., recurrent at a specified repetition frequency, or they may require an initiating waveform.

10.2. *RC* Differentiation. The simplest and one of the most frequently used methods of generating a trigger from a gate or other rectangular waveform is by the
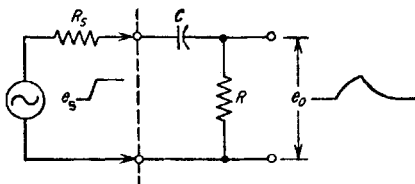
Fig. 10.1. *RC* differentiator.
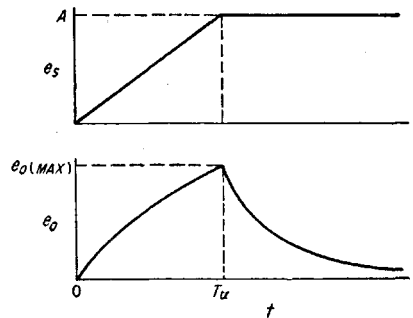
Fig. 10.2. Waveforms for the *RC* differentiator.

use of an *RC* differentiator. The circuit is shown in Fig. 10.1. The leading edge of the output waveform e_o in Fig. 10.2 is given by

$$e_o = \frac{ARC}{T_a} \left(1 - e^{-t/(R+R_s)C}\right) \qquad \text{for } 0 < t < T_a \qquad (10.1)$$

where T_a = rise time of input waveform, sec
      R_s = source resistance, ohms
      R = value of differentiating circuit resistor, ohms
      A = amplitude of input waveform, volts (see Fig. 10.2)
      C = value of differentiating circuit capacitor, farads

The peak of the output waveform $e_{o(max)}$ occurs when t is equal to T_a and is given by Eq. (10.1). The maximum possible output amplitude is $AR/(R_s + R)$, and this value is obtained when T_a is small compared to $(R_s + R)C$. The peak output-voltage amplitude $e_{o(max)}$ will be within 5 per cent of the input-voltage amplitude A if T_a is equal to or less than $0.1(R_s + R)C$. When R_s is small compared to R, $e_{o(max)}$ can be determined from Fig. 10.3. The trailing edge of the output waveform decays exponentially from the voltage $e_{o(max)}$ to zero with a time constant of $(R_s + R)C$.

The width of the differentiated waveform at the 50 per cent amplitude level is given by

$$\frac{T_{50\%}}{T_a} = \frac{(R + R_s)C}{T_a} \log_e \left(1 + e^{T_a/(R+R_s)C}\right) \qquad (10.2)$$

The ratio $T_{50\%}/T_a$ is plotted as a function of $(R + R_s)C/T_a$ in Fig. 10.4.

10-2
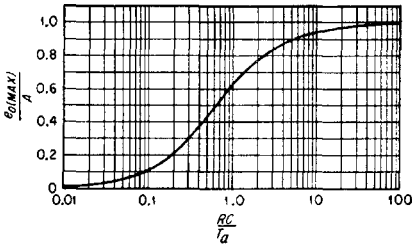
FIG. 10.3. Peak amplitude of differentiated waveform as a function of RC/T_a assuming $R \gg R_s$.
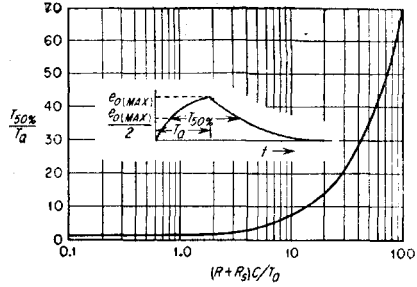
FIG. 10.4. Width of differentiated waveform at 50 per cent amplitude as a function of $(R + R_s)C/T_a$.

Example 10.1

Design an RC differentiating network to develop a trigger with the maximum possible amplitude and having a width of 3 μsec at the 50 per cent amplitude level. This trigger is to be developed from a gate having an amplitude of 100 volts and a rise time of 1 μsec. The gate has a source impedance of 1,000 ohms.

Solution

1. Determine the discharge time constant $(R + R_s)C$.
From Fig. 10.4 for $T_{50\%}/T_a = 3$,

$$\frac{(R + R_s)C}{T_a} \simeq 3.8$$

Therefore,

$$(R + R_s)C \simeq 3.8 \ \mu\text{sec}$$

To achieve the maximum amplitude possible, R should be much larger than R_s, i.e., much larger than 1,000 ohms. Since the values of R and C are not critical, let

$$R = 10,000 \text{ ohms}$$
$$C = 380 \ \mu\mu\text{f}$$

2. Determine the peak amplitude of the differentiated waveform.
From Fig. 10.3 where $(R + R_s)C/T_a = 3.8$,

$$\frac{e_{o(\text{max})}}{A} \simeq 0.87$$

Therefore,

$$e_{o(\text{max})} \simeq 87 \text{ volts}$$

10.3. RLC Peaking. Ideally, a circuit containing only inductance and resistance could be used to produce triggers in a manner similar to that described for the RC

(a) VOLTAGE GENERATOR   (b) CURRENT GENERATOR
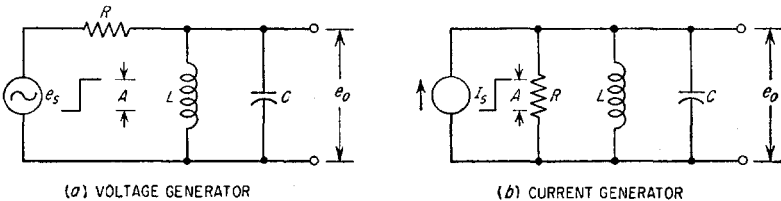
FIG. 10.5. Basic RLC peaking circuits.

differentiator in Sec. 10.2. However, shunt capacitance is always present across any physical inductance or resistance, and in reality any RL differentiator is an RLC peaking circuit. Triggers of relatively high amplitude and short duration can be

developed from the leading or trailing edge of a rectangular gate by applying the gate to an RLC peaking circuit. The basic RLC circuits are shown in Fig. 10.5. For the circuit of Fig. 10.5a, where the LC circuit is driven by a voltage generator having an internal resistance R, the output voltage e_o is given by Eq. (10.3) when the input is a positive step.

$$\frac{e_o}{A} = \frac{2\zeta e^{-\zeta \omega_n t}}{\sqrt{1 - \zeta^2}} \sin\ (\sqrt{1 - \zeta^2}\ \omega_n t) \qquad (10.3)$$

where A = amplitude of generator voltage step

$$\zeta = \text{damping ratio} = \frac{\sqrt{L/C}}{2R}$$

$$\omega_n = \frac{1}{\sqrt{LC}}$$

Equation (10.3) is plotted in Fig. 10.6 for various values of the damping ratio ζ. Increasing the damping increases the peak pulse amplitude and lengthens the pulse. Increasing ω_n decreases the pulse width. A negative-voltage step input will produce an output identical to that of Eq. (10.3) and Fig. 10.6 except that the polarity of the output will be negative.
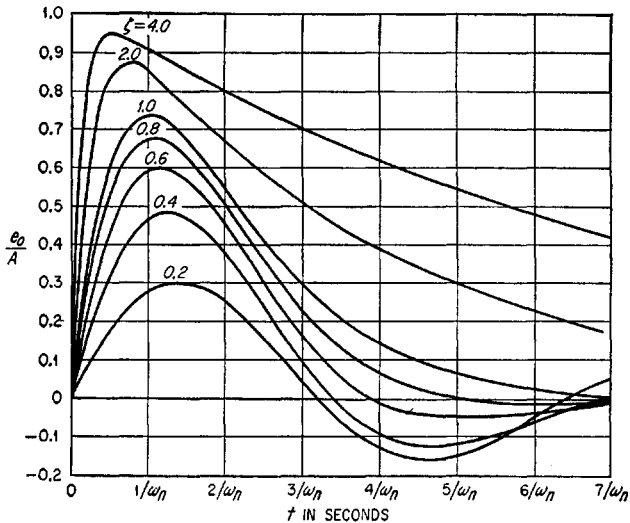
FIG. 10.6. Output waveform of RLC peaking circuit of Fig. 10.5a for a positive-voltage step input of amplitude A.

When a positive-current step input is applied to an RLC circuit as shown in Fig. 10.5b, the output voltage e_o is given by

$$\frac{e_o}{A\sqrt{L/C}} = \frac{e^{-\zeta \omega_n t}}{\sqrt{1 - \zeta^2}} \sin\ (\sqrt{1 - \zeta^2}\ \omega_n t) \qquad (10.4)$$

where ζ and ω_n have the same values as in Eq. (10.3).

Equation (10.4) is plotted in Fig. 10.7 for various values of the damping ratio ζ. When ζ is equal to unity, the circuit is critically damped and the pulse has no overshoot. Smaller values of damping increase the peak pulse amplitude and narrow the pulse length. However, the overshoot increases as the damping is decreased. The overshoot can be reduced to a negligible value by connecting across the RLC circuit a diode which conducts when the output voltage reverses polarity.

In either RLC circuit, the resonant frequency ω_n should be high to provide a fast pulse rise time. In Fig. 10.5a, the ratio L/C should be as large as possible so that
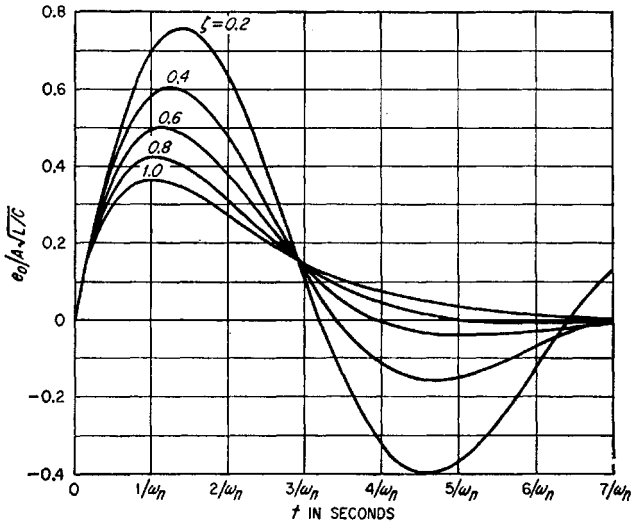
FIG. 10.7. Output waveform of RLC peaking circuit of Fig. 10.5b for a positive-current step of amplitude A.

the desired value of damping can be obtained with a reasonable value of source resistance. The ratio L/C should be small for the current-generator circuit of Fig. 10.5b to achieve the highest possible output-voltage amplitude.

Example 10.2

Design an RLC peaking circuit to produce a negative pulse having a rise time of 0.1 μsec when driven from the plate of a 5840 pentode having a plate supply of 100 volts and a 10-volt positive gate applied to the grid. The tube can be considered a constant current source.

Solution

1. General circuit considerations.
For maximum pulse amplitude, the damping ratio should be low and a damping diode should be incorporated to minimize the overshoot. The circuit is shown in Fig. 10.8. Assume that a 5829 subminiature diode is used as the damping diode.

2. Determine C.
For a fast rise time, C should be as small as possible. The minimum value of C is determined by

FIG. 10.8. Circuit of RLC peaking circuit for Example 10.2.

$$C_{\min} = C_o + C_p + C_L + C_s$$

where C_o = output capacitance of $V1$
$\quad\quad$ = 3.4 $\mu\mu$f
$\quad C_p$ = capacitance of diode $V2$ from plate to all other electrodes
$\quad\quad$ = 2.6 $\mu\mu$f
$\quad C_L$ = distributed capacitance of inductance L
$\quad C_s$ = socket and stray circuit capacitance
$\quad\quad \simeq$ 5 $\mu\mu$f

Therefore

$$C_{\min} = 11 \ \mu\mu f + C_L$$

3. Determine the resonant frequency ω_n.

Assume a minimum damping ratio ζ equal to 0.2. From Fig. 10.7 a rise time T equal to 0.1 μsec between the 10 per cent and 90 per cent points of the pulse amplitude requires that

$$0.1 \times 10^{-6} = \frac{0.96 - 0.08}{\omega_n}$$

$$\omega_n = \frac{0.88}{0.1 \times 10^{-6}}$$

$$= 8.8 \times 10^6 \text{ radians/sec}$$

or
$$f_n = 1.4 \text{ Mc/sec}$$

4. Determine L.

L is determined on the basis of the required resonant frequency ω_n and the capacitance $C_{\min}$. However, $C_{\min}$ is effected by the distributed capacitance of the inductance L. Therefore, determine L on the basis of zero distributed capacitance and then make a correction for $C_{\min}$ on the basis of the expected capacitance for the determined value of inductance.

$$L = \frac{1}{\omega_n{}^2 C_{\min}}$$

$$= \frac{1}{(8.8 \times 10^6)^2 \times 11 \times 10^{-12}}$$

$$= \frac{1}{77.4 \times 10^{12} \times 11 \times 10^{-12}}$$

$$= 1.175 \text{ mh}$$

The distributed capacitance of a coil of this size is about 10 $\mu\mu$f. Therefore, assume $C_{\min}$ to be equal to 20 $\mu\mu$f. Based on this corrected value of $C_{\min}$, the required value of L is equal to 646 μh. The exact value of L should be established experimentally.

5. Determine the maximum plate current and the value of R_k.

Since the maximum plate dissipation rating is 1.1 watts,

$$I_{b(\max)} = \frac{1.1}{100}$$

$$= 11 \text{ ma}$$

Therefore, from Fig. 10.9, it can be seen that the smallest value of grid bias is limited to approximately -1.0 volts. Self-bias protection for the tube can be obtained by making the value of R_k such that the grid-cathode bias is -1 volt when the grid voltage is 0 volts. The required value of R_k is given by

$$R_k = \frac{1.0}{I_b + I_{c2}}$$

$$= \frac{1.0}{10.6 \times 10^{-3} + 3.4 \times 10^{-3}}$$

$$= 71.5 \text{ ohms}$$

The maximum amplitude of the gate applied to the grid of $V1$ is clamped at ground potential by utilizing the other half of the diode $V2$ as shown in Fig. 10.8. The fixed bias of -8 volts as shown in Fig. 10.8 will keep $V1$ cut off in the absence of the positive gate and will ensure that the grid-cathode voltage reaches approximately 0 volts when the 10-volt gate is applied.

6. Determine the value of the damping resistor R.

From Eq. (10.3)

$$R = \frac{\sqrt{L/C}}{2\zeta}$$

$$= \frac{\sqrt{646 \times 10^{-6}/(20 \times 10^{-12})}}{2 \times 0.2}$$

$$= 14{,}200 \text{ ohms}$$
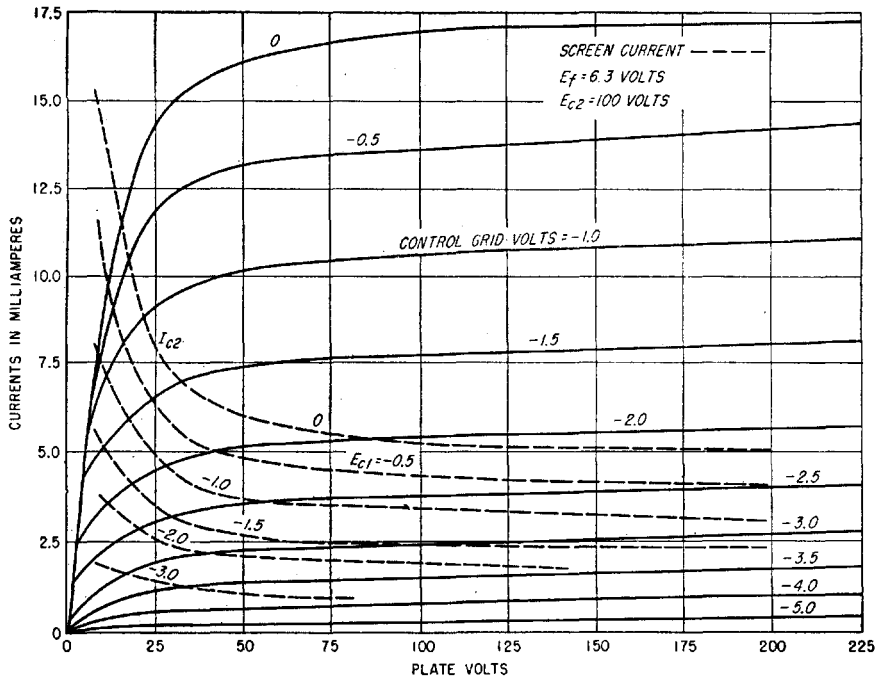
FIG. 10.9. Plate characteristics for 5840 pentode.

7. Determine the peak amplitude of the pulse.
From Fig. 10.7 and $\zeta = 0.2$,

$$\frac{e_{o(\text{max})}}{A \sqrt{L/C}} = 0.76 \qquad \text{at } t = 1.4/\omega_n$$

Therefore,

$$e_{o(\text{max})} = 0.76 \times 10.6 \times 10^{-3} \times \sqrt{\frac{646 \times 10^{-6}}{20 \times 10^{-12}}}$$

$$= 45.8 \text{ volts}$$

8. Determine the width of the pulse at the half-amplitude level.
From Fig. 10.7

$$\tau = \frac{2.52 - 0.43}{\omega_n}$$

$$= \frac{2.09}{8.8 \times 10^6}$$

$$= 0.238 \ \mu\text{sec}$$

10.4. Blocking Oscillators. A blocking oscillator is a transformer-coupled oscillator having a broadband feedback path. Such a circuit is capable of producing large amplitude pulses having widths of approximately 0.1 to 25 μsec.

Blocking oscillators can be either monostable, in which a triggering signal is required to initiate the pulse generation, or astable, in which the circuit is free-running and produces pulses at a fixed repetition frequency. A blocking-oscillator circuit possesses the desirable characteristic of having the tube nonconducting except during the period of pulse generation, thus allowing high-peak power output at a low value of average power.

The basic circuit of a blocking oscillator is shown in Fig. 10.10. The transformer primary is in the plate circuit, and the transformer secondary is coupled into the grid circuit in such a manner that increasing plate current causes a positive voltage to be applied to the grid. As a result of this positive feedback, an initiating trigger which causes plate current to flow in $V1$ will cause the plate current and grid voltage to increase rapidly until the grid is driven considerably positive. When equilibrium has been reached, the grid voltage, grid current, and plate voltage remain essentially constant. As is later discussed, the plate current increases with time because of increasing transformer magnetizing current, resulting in a slight decrease
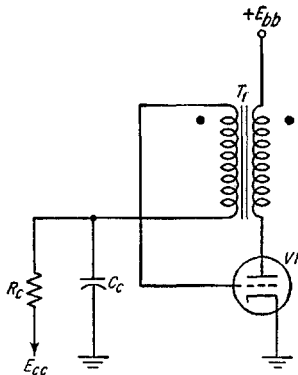
FIG. 10.10. Basic circuit of a blocking oscillator.
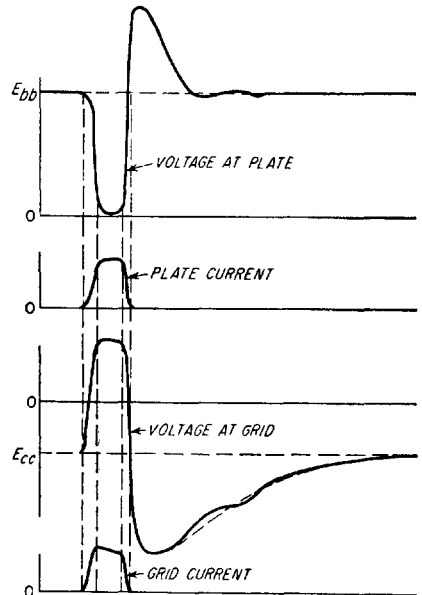
FIG. 10.11. Typical waveforms for an astable blocking oscillator.

in the grid voltage. This action continues until the grid voltage cannot support the increased plate current. When this point is reached, the plate current and grid voltage decrease rapidly in a regenerative manner until the tube is cut off. Typical plate and grid current and voltage waveforms are shown in Fig. 10.11.

Transformers used in blocking oscillators must be capable of passing the band of frequencies required to achieve the desired pulse shape. The high-frequency response determines the maximum pulse rise and fall rates, and the low-frequency response determines the maximum pulse length. The desirable characteristics of a blocking-oscillator pulse transformer are:

1. Low leakage inductance
2. Low winding capacitances
3. High open-circuit inductance (determined by requirements of pulse length)
4. High core permeability

For greater detail on pulse-transformer design, see Sec. 14.7.

The desired turns ratio between primary and secondary is a function of the tube type and the choice between maximum tube plate current and maximum plate-voltage swing. Maximum plate current is obtained when the plate resistance reflected to the grid circuit is equal to the grid resistance at the top of the pulse. Turns ratios from 1:1 to 3:1 are commonly used.

The path of operation of the tube in a blocking-oscillator circuit is very difficult to describe accurately because of the effects of circuit capacitances, transformer leakage inductance, and the lack of tube characteristics covering the operation of the tube

at high grid voltages and currents. As a result, the design of a blocking oscillator having certain prescribed characteristics is normally based on empirical data, and the exact pulse widths, rise times, and amplitudes are determined experimentally. However, a number of general considerations in the design of blocking oscillators are very useful in determining the initial choice of tube type, transformer, and RC network.

By making several simplifying assumptions, the approximate path of operation of a typical blocking oscillator can be described to illustrate the factors determining pulse amplitude and pulse width. With reference to the circuit shown in Fig. 10.12a, let it be assumed that C_c is charged to a potential of -50 volts at quiescence and that C_c is large enough that this potential does not change during the pulse. At the initiation of the pulse by an input trigger, the magnetizing current is zero. Therefore, if it is assumed that the rise time is zero, the initial plate current, as modified by the transformer turns ratio, flows entirely through the static grid resistance R_g.

(a) CIRCUIT OF A BLOCKING OSCILLATOR

(b) EQUIVALENT GRID CIRCUIT, GRID CONDUCTING

(C) EQUIVALENT PLATE CIRCUIT WHEN GRID CURRENT IS ZERO
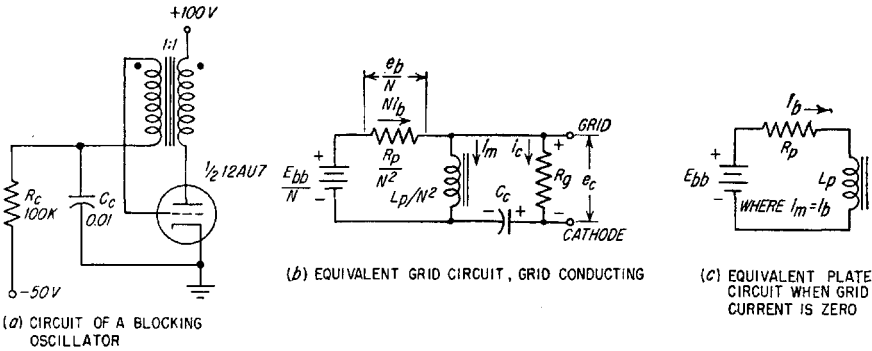
FIG. 10.12. Typical blocking-oscillator circuit.

The approximate equivalent circuit referred to the grid when the grid is conducting, neglecting shunt capacitances and leakage inductance, is given in Fig. 10.12b. The open-circuit transformer primary inductance is designated as L_p. The initial regenerative action is terminated when Eqs. (10.5) and (10.6) are satisfied.

$$i_c = N i_b \qquad (10.5)$$

$$\frac{E_{bb}}{N} = e_c - E_c + \frac{e_b}{N} \qquad (10.6)$$

where N = transformer turns ratio = $\dfrac{\text{turns in plate winding}}{\text{turns in grid winding}}$

i_b = instantaneous plate current

i_c = instantaneous grid current

e_c = instantaneous grid-to-ground voltage

   = $i_c R_g$

R_g = static grid resistance

e_b = instantaneous plate-to-ground voltage

   = $i_b R_p$

R_p = static plate resistance

E_{bb} = plate supply voltage

E_c = capacitor voltage

For the circuit shown in Fig. 10.12a, the dotted line $DABC$ in Fig. 10.13 shows the locus of all points satisfying Eq. (10.6), i.e., all points at which the sum of the grid and plate voltages is equal to 50 volts. The grid current curve for a control grid bias of

+30.6 volts labeled X intersects the $DABC$ line at point A and satisfies both Eqs. (10.5) and (10.6) at point A.   It should not be inferred that either the curve X or the line $DABC$ describes the path on the tube characteristic for the leading edge of the pulse.   Instead, they are used only to establish the operating point at the termination of the initial regenerative action.   Assuming zero rise time, the magnetizing current i_m begins to flow when point A is reached and increases essentially linearly with time (initial slope of an exponential curve) and causes the plate current to increase to a value above that satisfying Eq. (10.5).   The increasing plate current causes e_b to rise slightly because of the increased voltage drop across the tube and causes e_c to
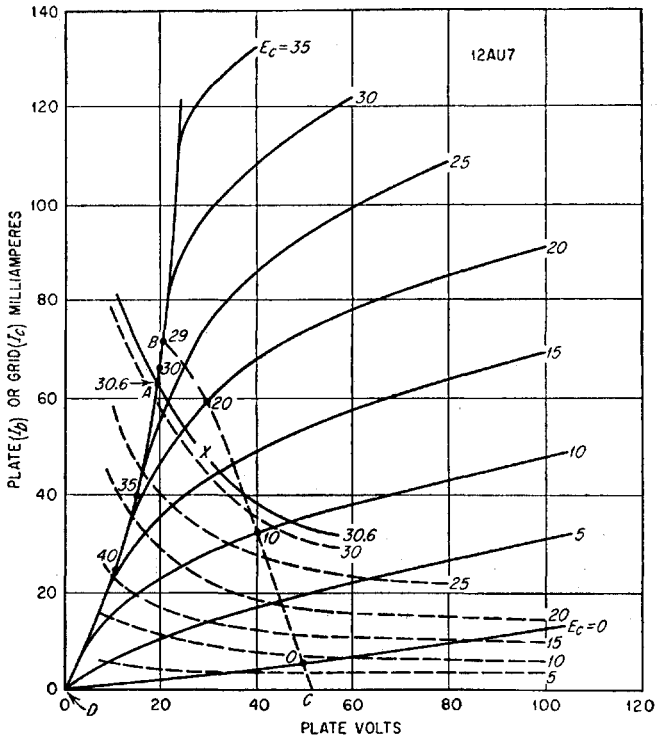
Fig. 10.13. Path of operation of a typical blocking oscillator.

decrease because of transformer action until point B is reached in Fig. 10.13.   Prior to reaching point B, the grid is in the saturated region, consequently the transformer action is not regenerative.   At point B the grid is no longer in the saturated grid region and the circuit becomes regenerative.   As a result, the plate current is cut off and the circuit is restored to the quiescent condition.   During the interval between points A and B, the plate voltage increased by approximately 2 volts; therefore, the grid voltage decreased by approximately the same amount.   The associated plate current increase is 10 ma, and from the grid current curves, the grid current decrease is approximately 12 ma.   The total magnetizing current is, therefore, 22 ma at the end of the pulse.   The time required for the operating point to shift from point A to point B establishes the pulse width τ and is approximately given by

$$\tau \simeq \frac{L_p i_{m(\text{max})}}{E} \tag{10.7}$$

where τ = pulse width, sec

    E = average voltage across pulse transformer primary in interval in which operating point moves from point A to point B

    $i_{m(max)}$ = magnetizing current at point B

At the instant the grid stops conducting in the regenerative shutoff phase, the plate current then consists only of the magnetizing current and the equivalent circuit is that shown in Fig. 10.12c. In the final regenerative shutoff phase, the magnetizing current through the tube is cut off. This induces a large positive and a large negative voltage in the plate and grid circuits, respectively. These waveforms are shown at the end of the pulse in Fig. 10.11. The plate voltage has an exponential decay with a time constant t_d equal to

$$t_d = \frac{L_p}{R} \tag{10.8}$$

where R = total shunt resistance across L_p when tube is cut off, e.g., transformer-core loss resistance and any resistive loading across primary and secondary windings all referred to primary winding

The reverse voltage swing at the end of the pulse is approximately equal to the amplitude of the plate-voltage swing during the pulse in most blocking oscillators. The shunt capacitance across the transformer resonates with the transformer inductance and produces a damped oscillation which is superimposed on the exponential decay of the reverse swing as shown in Fig. 10.11. The oscillation is not present during the pulse because of the heavy damping across the transformer by the grid and plate resistances. Neglecting the shunt capacitances and the leakage inductance, the generation of the pulse in accordance with the action just described is illustrated in idealized form in Fig. 10.14.

In conventional blocking oscillators using medium μ triodes with approximately 250 to 300 plate volts, the grid is driven much more positive during the pulse than indicated in the preceding example and the plate current during the pulse may reach 1 amp or more. Although the oxide-coated cathodes of such tubes are capable of supplying such peak currents, the average current rating of the tube must not be exceeded if normal tube life is desired. Average grid and plate power dissipations must also be kept within the ratings of the tube used.

Medium-μ triodes are, in general, well suited for use in blocking oscillators. For example, the 5687 triode operates very satisfactorily in blocking-oscillator circuits. With a plate supply voltage of +250 volts, it is capable of producing peak pulse currents of over 1 amp with a 1:1 turns ratio pulse transformer.

Peak currents in tetrode and pentode blocking oscillators are considerably less than for triodes having the same dissipation ratings and which utilize the same supply voltage. Triode-connected beam tetrodes are sometimes used in higher-power blocking oscillators because of their high plate dissipation capabilities.

In the previously discussed example, the capacitor in the grid circuit was assumed to be large enough that its potential did not change appreciably during the pulse. If the value of this capacitor is reduced sufficiently, the negative potential across the capacitor is increased by the tube grid current during the pulse, and E_c in Eq. (10.6) increases as a function of time. As a result, the plate current does not increase as far as point B in Fig. 10.13 before the regenerative cycle which ends the pulse occurs. Thus, C_c determines the reduction of the pulse length from the limit set by the transformer, and, as previously discussed, the open-circuit inductance of the transformer determines the maximum length of the pulse. The charge accumulated on C_c during the pulse discharges through R_c after the end of the pulse. This exponentially decaying voltage which has a time constant equal to R_cC_c adds to the grid reverse swing voltage to give the complete grid-cathode voltage at the end of the pulse. This can be seen in Fig. 10.11 by noting the differences in plate and grid waveforms.

If R_c is returned to ground instead of to a negative supply voltage, the voltage developed across C_c by the grid current during one pulse will prevent the tube from conducting again until the voltage across C_c becomes equal to the tube cutoff bias E_{co}. Astable blocking oscillators are biased in this manner. The time interval between pulses is determined by the voltage developed across C_c during the pulse and the discharge time constant R_cC_c. When a value of C_c has been found which gives the desired pulse length, the pulse repetition frequency is adjusted by the selection of R_c. Because the grid voltage is increasing relatively slowly at the instant cutoff bias is reached, the initial rate of increase of current through the transformer plate winding
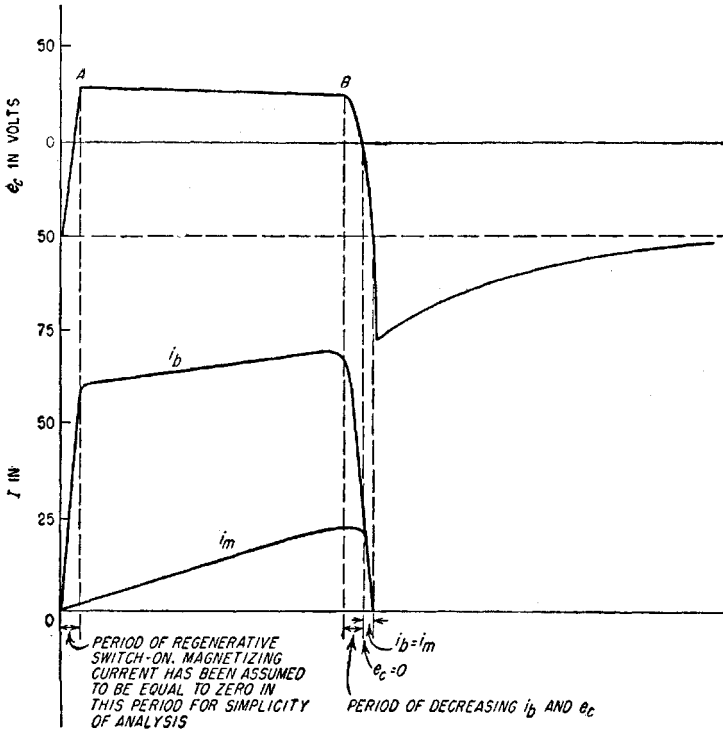
FIG. 10.14. Idealized waveform of blocking oscillator described in Figs. 10.12 and 10.13.

is low, and consequently the voltage initially developed across the transformer secondary is also low. As a result, the regenerative action driving the grid into conduction builds up relatively slowly and the rise time of the output pulse may be several tenths of a microsecond. However, if the blocking oscillator is triggered by a high-amplitude pulse having a short rise time, the grid will be driven positive very rapidly and the build-up of the regenerative action will be very fast. For these conditions, the output pulse will have a rise time very close to that of the initiating trigger.

Monostable blocking oscillators can be triggered by introducing a negative trigger across the plate winding of the transformer or by adding a positive trigger in series with the grid winding. These methods are illustrated in Fig. 10.15a and b, respectively. It is usually desirable to incorporate a separate trigger tube to minimize the interference of the blocking oscillator on the trigger source.

In the parallel triggering circuit, the plate of the trigger tube can be connected to a third winding on the transformer with the same polarity as the plate winding, or
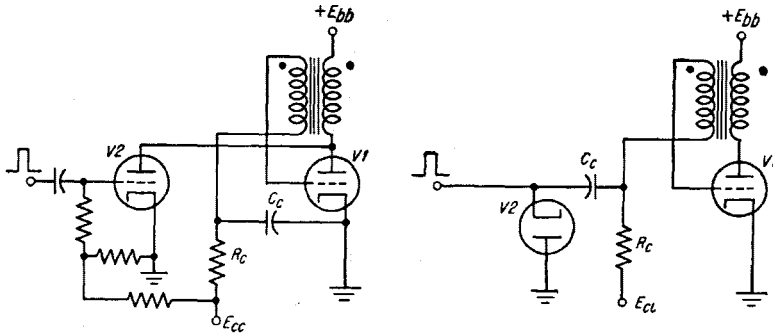
(a) PARALLEL TRIGGERING   (b) SERIES TRIGGERING

FIG. 10.15. Methods of triggering blocking oscillators.
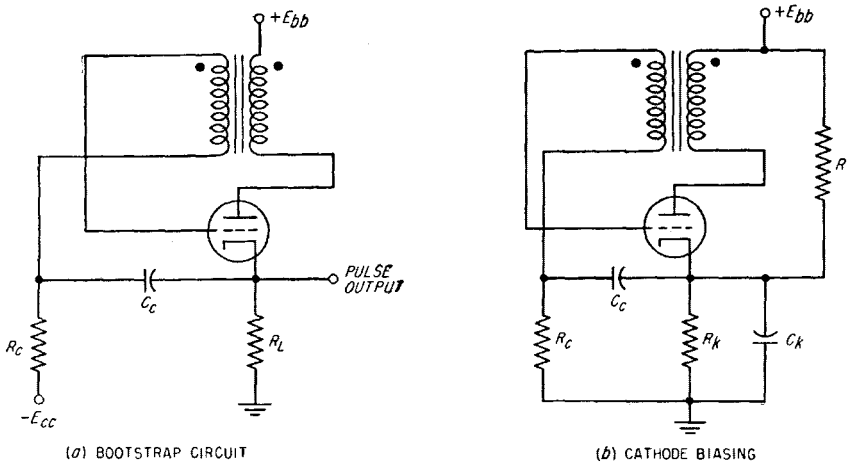
(a) BOOTSTRAP CIRCUIT   (b) CATHODE BIASING
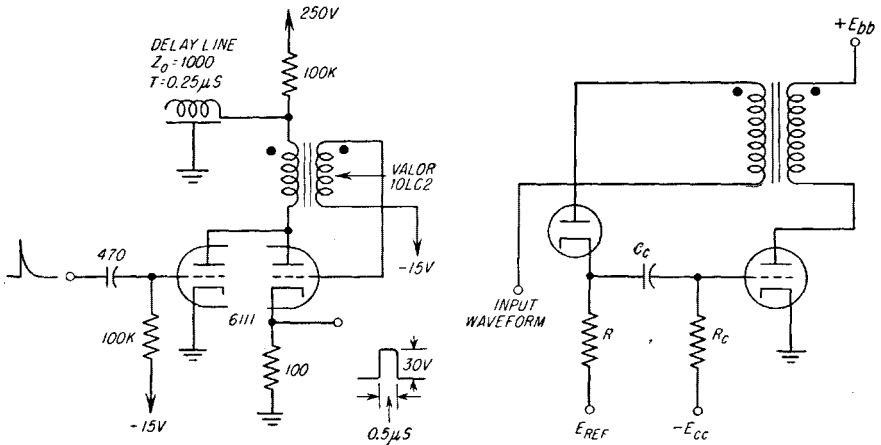
(c) TYPICAL DELAY LINE TIMING CIRCUIT FOR PRECISION PULSE WIDTH CONTROL   (d) REGENERATIVE PICK-OFF

FIG. 10.16. Several blocking-oscillator-circuit configurations.

it can be connected to the plate of the blocking oscillator. The quiescent current through the trigger tube should be zero or as small as possible to minimize the possibility of transformer core saturation.

In the series triggering circuit, the reverse swing across the grid winding of the transformer at the end of the pulse can cause circuit difficulties in the triggering stage if a cathode follower is used, and therefore a damping diode is desirable. For this reason, parallel triggering is preferable.

In addition to the waveforms available at the plate and grid of the blocking oscillator, additional outputs can be obtained from auxiliary windings on the transformer or from small resistances in series between the transformer plate winding and the supply voltage or between the cathode and ground. Typical resistance values are 50 to 250 ohms, and pulse amplitudes of 50 to 100 volts are obtainable by this method. These waveforms are derived from the plate current through the tube and, therefore, possess the advantage of not having any reverse swing. When a series cathode resistor is used, C_c should be returned to the cathode of the tube rather than to ground. This allows the blocking oscillator to operate without cathode degeneration. Returning C_c to the cathode is commonly referred to as bootstrapping. It allows the use of the minimum value of C_c for a given pulse length.

Several typical blocking-oscillator-circuit configurations are shown in Fig. 10.16.
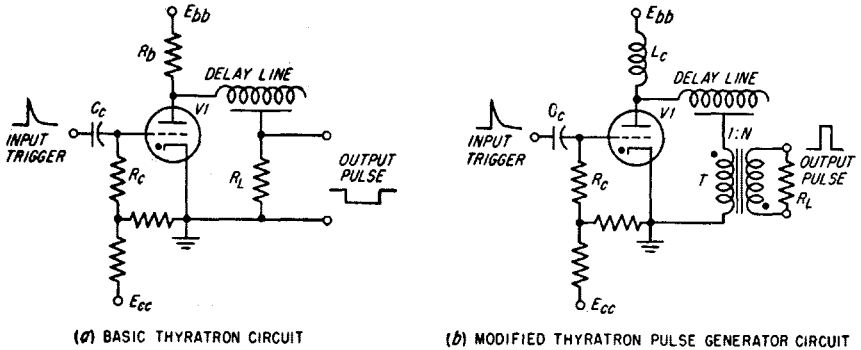
(a) BASIC THYRATRON CIRCUIT   (b) MODIFIED THYRATRON PULSE GENERATOR CIRCUIT

FIG. 10.17. Thyratron pulse-generator circuits.

10.5. Thyratron Pulse Generators. In many applications it is desirable to generate a pulse having a time duration which can be held to close tolerances without adjustment. One way in which this can be readily accomplished is by the use of a delay line (see Sec. 20.8) to establish the pulse duration.

Although delay lines can be used in many circuits to establish the duration of a pulse or a gate, e.g., in a blocking oscillator, the most common circuit is that of Fig.
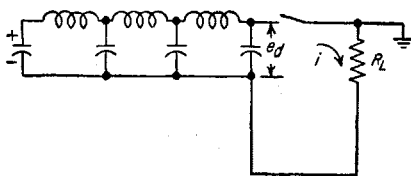
FIG. 10.18. Equivalent circuit for discharge of delay line.

10.17a. The delay line can be either a lumped-constant or a distributed transmission line of sufficient length to provide the desired time delay. The characteristic impedance Z_o of the line is made equal to the load resistance R_L. The line has an open-circuit termination. When the thyratron is cut off, the shunt capacitances of the delay line are charged to E_{bb}. When an input trigger is applied to the grid of the thyratron, the gas in the tube ionizes and the plate potential drops very rapidly to a typical value in the order of 10 volts for most thyratrons. The thyratron acts as a very low resistance switch and connects the delay line directly

to the load resistance R_L. The equivalent discharge circuit is given in Fig. 10.18, assuming that $R_b \gg R_L$ and that the resistance of the fired thyratron is zero. The voltage e_d at the terminals of the delay line is given by

$$e_d = \frac{R_L}{Z_o + R_L} E_{bb}\Gamma^n \qquad (10.9)$$

for $n\tau < t < (n+1)\tau$

where Γ = reflection coefficient when delay line is connected to load R_L

$$= \frac{R_L - Z_o}{R_L + Z_o}$$

t = time from instant thyratron is ionized, connecting delay line across load R_L

n = multiple of the two-way delay time of the delay line (0 or any integer)

NOTE: $\displaystyle\lim_{R_L \to Z_0} \left(\frac{R_L - Z_o}{R_L + Z_o}\right)^0 = 1$

τ = two-way delay time of delay line

Voltage and current waveforms for the delay line are given in Fig. 10.19 for R_L greater than Z_o, R_L equal to Z_o, and R_L less than Z_o. A single output pulse is generated when the load resistance is exactly equal to the characteristic impedance of the delay line.
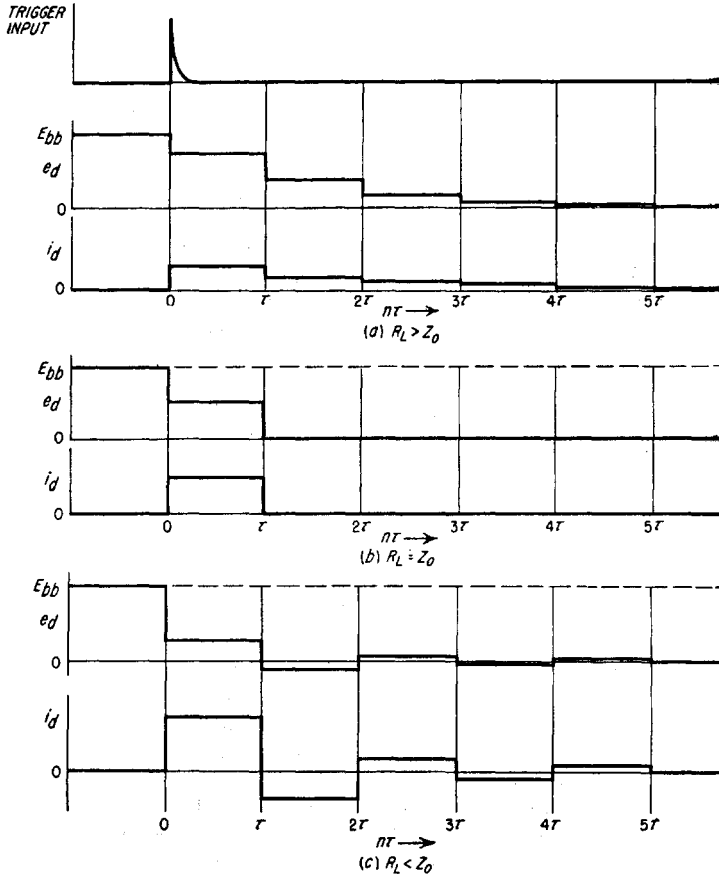
FIG. 10.19. Waveforms at delay-line terminals.

When the load resistance is lower than the line impedance, a series of alternately positive and negative pulses are generated as shown in Fig. 10.19 provided the thyratron continues to conduct after the end of the first positive pulse. However, the negative plate potential of the thyratron at the end of the first pulse initiates the deionization of the gas in the thyratron and aids ion cleanup in the tube. The amplitudes of the succeeding positive pulses then become a function of the deionization time and grid control characteristics of the thyratron and the pulse length. When R_L is greater than Z_o, a number of pulses of decaying amplitude occur after the first pulse. Since the plate potential does not become negative after the first pulse, the thyratron continues to conduct until a sufficient number of pulses have occurred to lower the plate potential to the deionization point. This type of operation is ordinarily avoided. If a positive output pulse is desired, R_L can be placed between the thyratron cathode and ground.
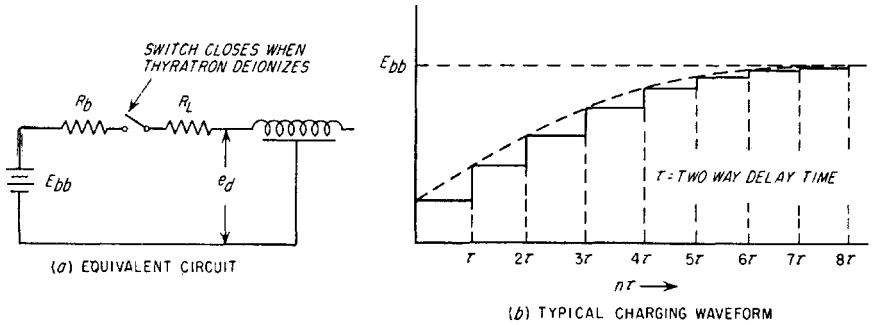
FIG. 10.20. Charging of delay line between pulses.

The minimum interval between successive pulses is determined by the time required for the delay line to be recharged to the supply voltage after the thyratron has deionized. The equivalent circuit for the charging of the delay line is given in Fig. 10.20a, and a typical charging waveform is shown in Fig. 10.20b. The equation for the charging of the delay line is given by

$$e_d = E_{bb}\left(1 - \frac{R_b + R_L}{R_b + R_L + Z_o}\,\Gamma^n\right)$$ (10.10)

for $n\tau < t < (n + 1)\tau$

where Γ = reflection coefficient

$$= \frac{R_b + R_L - Z_o}{R_b + R_L + Z_o}$$

R_b = plate-load resistance of thyratron
Z_o = characteristic impedance of delay line
τ = two-way delay time of line, sec
n = multiple of the two-way delay time of the delay line
t = time in seconds from instant thyratron is deionized

In most thyratron pulse generators of this type, the duty cycle of the pulse generator is made approximately 0.01 or less. When this condition exists, R_b is much larger than either R_L or Z_o and the term $(R_b + R_L)/(R_b + R_L + Z_o)$ in Eq. (10.10) can be assumed to be equal to unity.

The circuit of Fig. 10.17b is widely used as a pulse modulator for magnetrons. It differs from the basic circuit of Fig. 10.17a in that an inductance is used between the supply voltage and the delay line to provide a resonant charging circuit to allow the delay line to be charged to twice the supply voltage. Also a pulse transformer

is used to provide a voltage step-up to the magnetron and to isolate the modulator from the magnetron.[1]

Example 10.3

Design a thyratron pulse generator to supply a positive 0.5-μsec pulse having an amplitude of 100 volts when connected to a 100-ohm load and capable of being triggered at a maximum prf of 2,000 pulses/sec.

Solution

1. Determine the basic circuit.
Since a positive output pulse is desired, R_L is placed in the cathode circuit of the thyratron of Fig. 10.17a.
2. Determine the characteristic impedance Z_o of the delay line.
This should equal the load impedance to produce a single output pulse as indicated in Fig. 10.19. Therefore, $Z_o = R_L = 100$ ohms.
3. Determine the plate supply voltage.
The voltage to which the delay line is charged should be twice the desired output voltage plus twice the voltage drop across the thyratron when ionized. Assuming a 10-volt drop across the thyratron, the delay-line voltage should be 220 volts. If there is insufficient time between pulses for the delay line to be recharged to the full supply voltage, the supply voltage should be increased by the amount required to allow the delay-line voltage to be 220 volts at the instant each pulse occurs. Therefore, make the supply voltage +250 volts to allow for incomplete charging between pulses.
4. Determine the thyratron to be used.
The thyratron must have a peak current rating of 1.0 amp and a deionization time sufficiently less than the desired 500-μsec interpulse period to allow the delay line to be recharged to the full supply voltage in the remaining portion of the interpulse period.
A suitable tube is the 2050 tetrode. It has a peak current rating of 1.0 amp and a minimum deionization time of approximately 50 μsec. The tube voltage drop when ionized is approximately 8 volts.
5. Determine the value of the charging resistor R_b.
The delay line must be recharged from 0 to 220 volts in 500 μsec minus the tube deionization time, or approximately 450 μsec.
Using Eq. (10.10)

$$220 = 250(1 - \Gamma^n)$$
or
$$\Gamma^n = 0.12$$

where n = number of two-way delay periods of delay line in 450 μsec
$$= 900$$
Therefore
$$\Gamma = 0.99765$$

Since R_L is in the cathode circuit,

$$
\begin{aligned}
\Gamma &= \frac{R_b - Z_o}{R_b + Z_o} \\
&= \frac{R_b - 100}{R_b + 100} \\
0.00235 R_b &= 199.765 \\
R_b &= 85,000 \text{ ohms}
\end{aligned}
$$

If R_b is made smaller than 85,000 ohms, the delay line will charge more closely to +250 volts in the interpulse period.
6. Determine the value of the grid resistor R_c and the grid bias voltage E_c.
Deionization of the thyratron is speeded up by a negative grid voltage since the remaining positive ions in the gas will be attracted to the grid. The positive ions thus collected constitute a current flow in the grid circuit, resulting in a positive bias across any grid-circuit resistance. Deionization is, therefore, hastened by having as low a value of grid-circuit resistance as possible and by returning the grid resistor to a high negative voltage.

---

[1] For an analysis of this type of charging circuit, see G. N. Glasoe and J. V. Lebacqz, "Pulse Generators," vol. 5, Radiation Laboratory Series, McGraw-Hill Book Company, Inc., New York, 1949.

The maximum grid bias is determined by the available trigger pulse amplitude and the requirement that the grid be raised above the critical value for ignition of the tube when the input trigger is applied.   For a plate voltage of +220 volts and with the shield grid
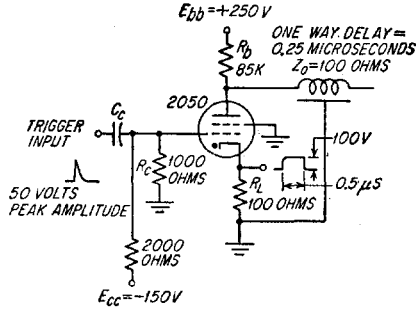
FIG. 10.21. Complete circuit of the thyratron pulse generator for Example 10.3.

grounded, the critical grid voltage is −2.3 volts.   The grid-circuit resistance should be 1,000 ohms or less if the deionization time of 50 μsec is to be achieved for the 2050 thyratron.   The complete circuit is shown in Fig. 10.21.

Sawtooth Generators

11.1. Sawtooth Generators. A sawtooth generator is desirable whenever a linearly changing voltage is required. For example, sawtooth-voltage generators are used (1) in oscilloscopes and synchroscopes to horizontally deflect the electron beam at a uniform speed and (2) to produce a sweep-frequency generator by applying a sawtooth voltage to a signal generator whose output frequency is a function of the input sawtooth voltage. Sawtooth generators can be either free-running or driven.

A free-running sawtooth generator produces a series of sawtooth waveforms. The repetition frequency is normally determined entirely by the sawtooth-generator circuit values; however, it is usually possible to increase the repetition frequency somewhat by introducing a synchronizing signal from an external source. The synchronizing signal need not occur at the sawtooth frequency but may be at a harmonic frequency.

A driven sawtooth generator produces a single sawtooth waveform whenever driven from an external source such as a trigger or gate generator. The desired result is an individual sawtooth generated for each trigger or gate input.

In the most common type of sawtooth generator, the sawtooth is created as a result of a capacitor charging or discharging through a resistor. In Fig. 11.1 is a simple RC circuit and the associated charging voltage across the capacitor. The time constant T is given by Eq. (11.1)

$$T = RC \qquad (11.1)$$

where T is in seconds, R is in megohms, C is in microfarads.

Fig. 11.1. Plot of the voltage appearing across capacitor C as a function of time.

t/T	e_o in % of E	t/T	e_o in % of E
0.01	0.995	1.4	75.34
0.02	1.98	1.6	79.81
0.03	2.96	1.8	83.47
0.04	3.92	2.0	86.47
0.05	4.88	2.2	88.92
0.06	5.82	2.4	90.93
0.07	6.76	2.6	92.57
0.08	7.69	2.8	93.92
0.09	8.61	3.0	95.02
0.10	9.52	3.2	95.92
0.20	18.13	3.4	96.66
0.30	25.92	3.6	97.27
0.40	32.97	3.8	97.76
0.50	39.35	4.0	98.17
0.60	45.12	4.2	98.50
0.70	50.34	4.4	98.77
0.80	55.07	4.6	98.99
0.90	59.34	4.8	99.18
1.0	63.21	5.0	99.33
1.2	69.88		

The voltage appearing across the capacitor is relatively linear for a length of time equivalent to about $0.1T$ or less. For this reason the time constant of most RC charging networks in sawtooth generators is made equal to about 10 times the duration of the desired sawtooth.

The instantaneous voltage across C at any time t can be determined from Eq. (11.2) if the accuracy with which Fig. 11.1 can be read is not considered adequate or if the

11-2

particular value has not been tabulated in Fig. 11.1.

$$e_o = E(1 - \epsilon^{-t/T}) \tag{11.2}$$

where e_o = instantaneous voltage across C at time t

E = source voltage

t = time in seconds from initiation of sawtooth

T = time constant of network, sec

11.1a. Free-running Sawtooth Generators. In a free-running sawtooth generator, a finite time interval is required between recurrent sawtooth waveforms for the recovery of the circuit. In most applications there is a requirement for maximizing the ratio of the sawtooth time interval to the recovery interval. In general, performance ratios of 10:1 or 20:1 are considered satisfactory. To illustrate how the deionization period of a gas tube limits the maximum usable frequency at which it might be used, consider the case where a minimum performance ratio of 10:1 is acceptable. If the gas tube is assumed to have a deionization time t_2 of 50 μsec, the sawtooth must have a length t_1 of at least 500 μsec to achieve the desired ratio of 10:1. The maximum frequency f_{max} at which this particular sawtooth generator will function acceptably can then be determined as follows:

$$\begin{aligned}
f_{max} &= \frac{1}{t_1 + t_2} \\
&= \frac{1}{500 \times 10^{-6} + 50 \times 10^{-6}} \\
&= 1{,}818 \text{ cycles}
\end{aligned}$$

Using vacuum tubes, special types of free-running multivibrators can be made to have a recovery period equal to or less than 1 μsec. In this manner, a free-running multivibrator can be designed to operate at 100 kc or higher and still retain the desired 10:1 performance ratio.

The adjustment of the sawtooth-generator circuit values plus the introduction of a synchronizing signal make it possible to synchronize the frequency of the sawtooth generator to either the synchronizing signal or a subharmonic of the synchronizing signal. In the case of an oscilloscope, the sawtooth generator must be synchronized to either the frequency or a subharmonic of the frequency being observed on the oscilloscope to allow a stable or nonjittery presentation.

1. Gas Tube as a Free-running Sawtooth Generator. Shown in Fig. 11.2 is a free-running sawtooth generator which uses a gas tube. The input circuit is used to synchronize the free-running sawtooth generator to an external signal. This particular sawtooth generator and synchronizing circuit combination is frequently employed in commercial oscilloscopes to generate the voltage for horizontally deflecting the electron beam.

FIG. 11.2. Gas tube used as a free-running sawtooth generator with associated synchronizing circuit.

The frequency of the sawtooth is dependent on E_{bb}, the time constant R_6C_3, and both the ionizing and ionized voltages of the gas tube. After the supply voltage is first turned on, the voltage appearing across C_3 and $V2$ increases toward the supply voltage because of the charging of C_3 through R_6. The voltage across C_3 continues to increase until it reaches the critical value E_F at which the gas tube ionizes. This value is determined by the type of tube and the

amount of bias. During the interval in which the gas tube conducts, C_3 discharges, and its terminal voltage is reduced to the gas-tube regulating voltage E_R.

The firing or ionizing voltage E_F of a typical biased gas tube might be 150 volts, and the ionized voltage E_R might be 15 volts. As soon as the tube has discharged C_3 to this new voltage, it deionizes, since R_6 has been made large enough to reduce the plate current flow to a value smaller than that required to maintain ionization. As soon as the tube is deionized, C_3 again starts to charge toward E_{bb}. As soon as the tube ionization voltage is reached, the tube again conducts, C_3 is again discharged to a very low voltage, etc. Grid resistor R_5 is used to prevent excessive grid current when the tube ionizes. Frequently a resistor is placed in series with C_3. The value of this resistor is chosen to limit the peak current through the gas tube.

This type of circuit can easily be synchronized to an external signal or to a subharmonic of some higher-frequency signal if the synchronizing signal is coupled into the grid circuit. The circuit shown in Fig. 11.2 illustrates a typical synchronizing circuit and the method by which either polarity of the synchronizing signal can be injected into the grid circuit of the sawtooth generator.

In Fig. 11.1, the value of E for establishing the generated sawtooth characteristics is equal to the difference between E_{bb} and the voltage to ground existing at the plate of $V2$ when ionized. The peak amplitude of the sawtooth is equal to $E_F - E_R$, and if this quantity is expressed as a percentage of E, the value of t/T can be determined from Fig. 11.1. The length t of the sawtooth can then be calculated since the time constant R_6C_3 is equal to T. The linearity of the sawtooth is not affected as the time constant R_6C_3 is varied. This time constant affects the frequency only.

The maximum frequency of the typical free-running gas-tube sawtooth generator is about 20 to 25 kc. This upper frequency is limited by the gas-tube deionization time. At the maximum operating frequencies, the major portion of each cycle is required for deionization. This results in a very low performance ratio and, consequently, limits the usefulness of the generated waveform.

2. Free-running Sawtooth Generators Using Vacuum Tubes. In general, a free-running sawtooth generator employing vacuum tubes consists of an astable plate-coupled multivibrator which is used to "gate on" some form of sawtooth generator circuit. The sawtooth generator circuits shown in Figs. 11.6 through 11.10 are in this category since they require an enabling gate. The operation of these sawtooth generators is discussed in Sec. 11.1c. The plate-coupled multivibrator shown in Fig. 11.3 and any one of the aforementioned sawtooth generators permit the generation of either free-running or externally synchronized sawtooth waveforms. The circuit shown in Fig. 11.3 is widely used in oscilloscopes and test equipment where a stable and easily synchronized circuit is required. It can easily be made to operate at frequencies as high as 60 kc with performance ratios as high as 10:1, i.e., a sawtooth duration ten times the recovery period. The duration t of the sawtooth is determined by the interval for which $V2$ conducts and can be determined as described in Sec. 8.5. The characteristics of the individual sawtooth waveforms can be established from the data in Sec. 11.1c.
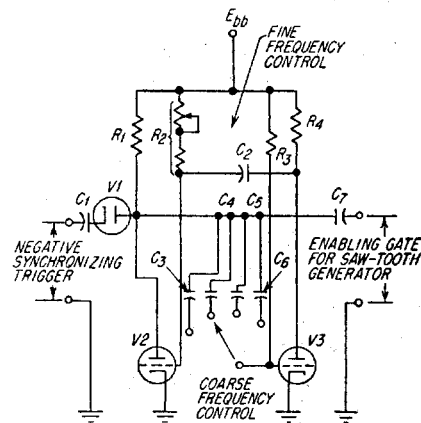
FIG. 11.3. Free-running multivibrator gate generator for enabling a sawtooth generator (see Figs. 11.6 through 11.10).

11.1b. Sawtooth Generators Requiring a Driving Trigger. The circuits shown in Figs. 11.4 and 11.5 generate a single sawtooth waveform for each input trigger. These circuits are members of a family of circuits which have one stable and one unstable position and are usually identified as either *one-shot* multivibrators or *monostable* multivibrators (see Secs. 8.3 and 8.4).

1. Trigger-driven Positive Sawtooth Generator. The basic one-shot multivibrator shown in Fig. 11.4 has been modified by the addition of C_3. A positive trigger applied to the grid of $V1$ initiates the regenerative cycle which puts the triggered tube into conduction and drives the grid of $V2$ far beyond cut off. The first tube will remain in conduction until the negatively driven grid of $V2$ returns sufficiently positive, because of the changing charge on the coupling capacitor C_2, to again cause plate current to flow in $V2$ and initiate the regenerative cycle which causes the circuit to restore itself to the quiescent condition. The duration of the gate is primarily a function of the bias on $V1$ and the time constant $R_5 C_2$ and can be determined as described in Sec. 8.4.
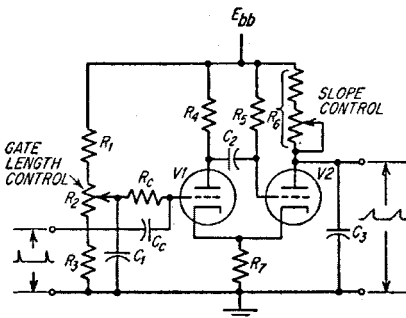
Fig. 11.4. One-shot multivibrator used as a positive sawtooth generator.
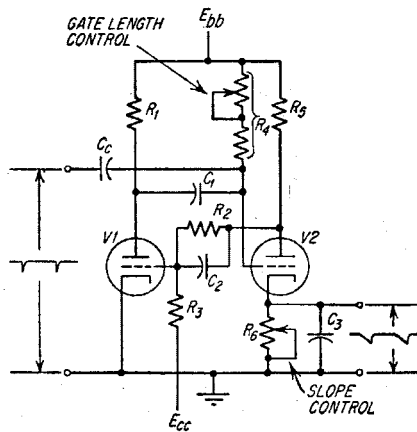
Fig. 11.5. One-shot multivibrator used as a negative sawtooth generator.

The generated sawtooth waveform follows the typical RC charging curve shown in Fig. 11.1. For determining the output waveform of this particular circuit, the charging voltage E of Fig. 11.1 is equal to E_{bb} less the quiescent voltage at the plate of $V2$. The charging time constant T is equal to $R_6 C_3$ and should be made several times larger than the duration t of the sawtooth to obtain good linearity (see Fig. 11.1).

2. Trigger-driven Negative Sawtooth Generator. The circuit shown in Fig. 11.5 is a plate-coupled one-shot multivibrator which has been modified by the addition of R_6 and C_3 in the cathode circuit of $V2$. During quiescence, $V1$ is cut off and $V2$ is conducting. The negative trigger applied to the grid of $V2$ initiates a regenerative cycle which drives $V2$ to cut off. After a period of time t, the charge on C_1 will have changed sufficiently so that the voltage at the grid of $V2$ will no longer be beyond cut off and $V2$ will once again conduct. As soon as $V2$ starts to conduct, a regenerative cycle restores the circuit to the quiescent condition. During quiescence, the product of the current through $V2$ and the value of R_6 will establish the quiescent cathode voltage E_k of $V2$. Once $V2$ is cut off, C_3 starts to discharge toward ground potential. With reference to Fig. 11.1, the value of E is equal to E_k and the time constant $R_6 C_3$ is equal to T. The duration t of the sawtooth is determined by the interval for which $V2$ is cut off. Details for calculating this interval are given in Sec. 8.3. In the circuits shown in Figs. 11.4 and 11.5, the controls can be ganged so as to

maintain the amplitude of the sawtooth approximately constant as the gate duration is varied.

11.1c. Sawtooth Generators Requiring an "Enabling Gate." Several types of driven sawtooth generators are shown in Figs. 11.6 to 11.10. Each of these sawtooth generators requires an *enabling gate* which initiates the start and determines the length of the generated sawtooth. In each instance the gate generating circuit must be triggered by a master or control trigger. Sweep circuits in a synchroscope are controlled in this manner.

The degree to which a driven sawtooth-generator circuit returns to the quiescent condition after each sawtooth and before the application of another enabling gate is a function of the period between gates. Since the slope of any given sawtooth is affected by the voltage across the charging circuit at the instant the enabling gate is applied, there exists a problem of producing identical sawtooth waveforms whenever the period before the enabling gates is varied.
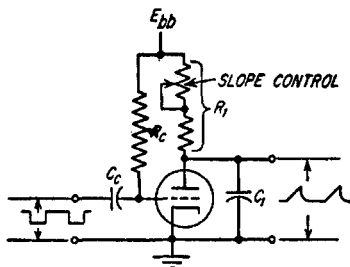
FIG. 11.6. Simple sawtooth generator.
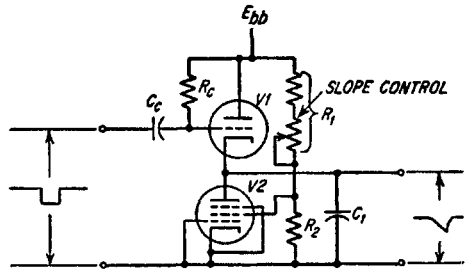
FIG. 11.7. Negative sawtooth generator utilizing a pentode as a constant current charging source.

Another consideration is the possibility of any irregularities in the enabling-gate waveform being fed through the grid-to-plate capacitance and consequently appearing on the generated sawtooth. This emphasizes the importance of minimizing any high-frequency components which might exist on the ideally flat portion of the gate. This problem becomes particularly aggravated when the plate-circuit capacitance to ground is small, as is usually the cases in those circuits which generate sawtooths having very large slopes.

1. Simple Positive Sawtooth Generator. One of the simplest positive sawtooth generators is shown in Fig. 11.6. The quiescent voltage across C_1 can be determined by establishing the plate voltage E_{bo} at which the d-c load line for the tube intersects the zero bias curve. Application of the negative enabling gate at the grid causes the plate current to be cut off. C_1 starts to charge to the supply voltage E_{bb}. The slope of the sawtooth voltage which appears across C_1 is a function of the time constant R_1C_1, the duration t of the gate, and the charging voltage E. E (see Fig. 11.1) is equal to the difference in voltage between E_{bb} and the quiescent plate voltage E_{bo}. The absolute value of the sawtooth voltage appearing across C_1 can be established by adding E_{bo} to the signal voltage determined by utilizing the charging curve in Fig. 11.1. The larger the time constant R_1C_1 with respect to the gate duration t, the more linear will be the resulting sawtooth. For most applications, it is sufficient to make this time constant equal to or greater than ten times the gate duration. The actual deviation from linearity can be approximated from Fig. 11.1 and can be established accurately from Eq. (11.2).

To ensure that the tube does not conduct during the enabling gate, the time constant R_cC_c should be several times longer than the gate or the gate must be of sufficient

amplitude that the changing charge on C_c during the gate does not permit the grid to return to the region where plate current flows. At the end of the enabling gate, the tube conducts and discharges C_1 toward its initial value. The return of the charge on C_1 to very nearly its initial value is not instantaneous. Consequently, sufficient recovery time must be allowed before another enabling gate is applied. The same considerations for recovery are applicable to the grid ciruit. In this case, very nearly all the charge acquired on C_c during the gate is discharged through the positively driven grid in a relatively short period immediately following the gate.

2. *Negative Sawtooth Generator.* $V1$ and $V2$ of Fig. 11.7 are normally conducting. The generation of a negative sawtooth is initiated by the application of a negative enabling gate to the grid of $V1$ which cuts off the plate current through $V1$. The magnitude of the gate appearing at the grid of $V1$ must be equal to or larger than the sum of the cutoff bias for $V1$ and the magnitude of the generated sawtooth waveform. When the gate cuts off $V1$, $V2$ starts to discharge C_1. As C_1 discharges, the plate voltage on $V2$ decreases. The plate current, however, remains essentially constant because of the high dynamic plate resistance of $V2$ which is characteristic of a pentode. The slope of the resulting sawtooth is approximately equal to $-i/C_1$ where i is the quiescent plate current. The same considerations for the values of R_c and C_c applicable to Fig. 11.6 apply here. The quiescent voltage E_{bo} across $V2$ is determined by assuming equal values of plate current for $V1$ and $V2$ and establishing from the tube characteristics the associated values of plate voltage. This problem is straightforward since both tubes are at zero bias. That value of plate current which causes the sum of the two plate voltages to be equal to E_{bb} is the quiescent plate current.

3. *Bootstrapped Sawtooth Generator.* The bootstrapped sawtooth generator in Fig. 11.8 is capable of generating a very linear positive sawtooth. The application of the negative enabling gate at the grid of $V1$ causes plate current cutoff of $V1$ and as a result C_1 starts to charge toward E_{bb}. As the voltage across C_1 starts to increase, essentially the same voltage increase appears at the output of the cathode follower provided that the cathode-follower gain is near unity. Since the cathode-follower output is coupled to the top of R_1 through C_2, the signals appearing at each end of R_1 are very nearly identical. The magnitudes of these signals are not exactly the same only because the cathode follower does not have a gain of unity and because C_2 slowly discharges during the gate. If these limitations did not exist, identical signals would appear at each end of R_1 and the voltage drop across R_1 would remain constant and equal to the quiescent voltage drop. A constant voltage drop across R_1 during the gate would ensure a constant current flow through R_1 into C_1. If the current flowing into C_1 is constant, the sawtooth waveform will be linear. The slope of the sawtooth waveform which appears across C_1 under these conditions can be determined from Eq. (11.3).

$$\frac{dV}{dt} = \frac{I_{bo}}{C_1} \qquad\qquad (11.3)$$

where C_1 is in farads and I_{bo} is the quiescent plate current through $V1$ in amps Although the bootstrapping is not perfect because the cathode follower gain is not unity and because C_2 discharges a certain amount during the gate, the error in using Eq. (11.3) is usually insignificant.

In the design of the bootstrapped sweep generator, the first step is to establish the quiescent plate current I_{bo} through $V1$. Different values of plate current are assumed until a value is found which makes the sum of the plate voltages $V1$ and $V3$ plus the voltage drop across R_1 equal to the supply voltage. Although bootstrapping does help in linearizing, it should not be considered as a "cure-all"; consequently, if very linear sawtooth waveforms are desired, it is recommended that the time constant R_1C_1 be made equal to at least ten times the gate length. A value for C_1 can therefore

be established, and in turn dV/dt across C_1 can be determined from Eq. (11.3). C_2 must be large compared to C_1 since the change in voltage across C_2 is equal to the peak value of the output sawtooth across C_1 multiplied by C_1/C_2. The cathode-follower stage should be designed for as high a gain as possible without making R_2 excessively large and, consequently, incompatible with the recovery limitations as described in the following paragraph.

During the gate, the cathode of the diode $V3$ is driven in a positive direction to a value higher than E_{bb}. The diode is in series with the recharging path for C_2 during the recovery period, but the recovery time is usually limited primarily by the time constant R_2C_2. The recovery time can be reduced by making R_2 as small as is practical or by returning R_2 to a higher negative voltage. R_2 should not be reduced to a value which causes the gain of the cathode follower to be materially reduced. Frequently $V3$ is replaced by a resistor, but this causes an increase in the recovery time. If a resistor is substituted for $V3$, its value must be large enough so as neither to load the cathode follower excessively nor to cause excessive discharge of C_2 during the gated period.
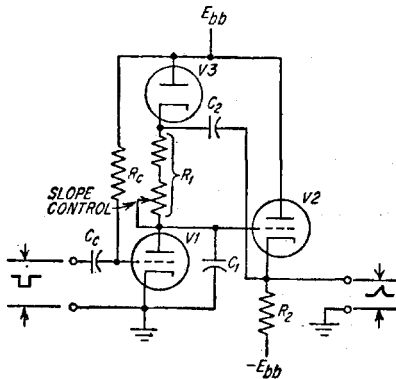
FIG. 11.8. Bootstrapped sawtooth generator.
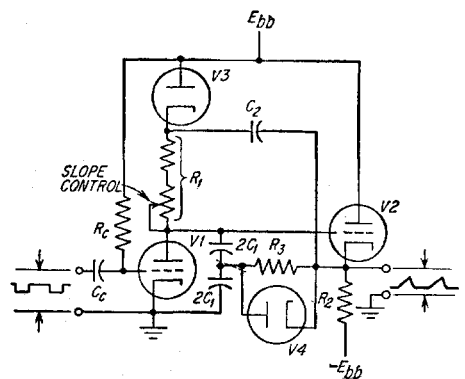
FIG. 11.9. Linearized bootstrapped sawtooth generator.

The considerations in the choice of values for R_c and C_c are the same as for the circuits shown in Figs. 11.6 and 11.7.

4. Linearized Bootstrapped Sawtooth Generator. The circuit in Fig. 11.9 is capable of generating a more linear sawtooth than is that in Fig. 11.8. With the exception of R_3, $V4$, and the replacement of C_1 with two individual capacitors, each having a value to $2C_1$, the design is the same as for the circuit in Fig. 11.8. To achieve the maximum linearity, R_3 should be established from Eq. (11.4).

$$R_3 \simeq \frac{R_1}{4(C_1/C_2 + 1 - A_{cf})} \qquad (11.4)$$

where A_{cf} = the gain of cathode follower

By selecting the proper value of R_3, the bottom capacitor can be charged at a rate which increases with time by an amount equal to the rate at which the rate of charge on the top capacitor decreases with time.

The purpose of $V4$ is to minimize the circuit recovery time.

5. Miller Integrator Circuit Used as a Sweep Generator. Basically, the Miller integrator circuit consists of a typical RC charging network with a high-gain amplifier paralleling the capacitor C (see Fig. 11.10a). The generation of the sawtooth is initiated by closing the switch S_1. Because of the negative feedback configuration,

the input voltage to the amplifier tends to remain at the quiescent level. The result is that the voltage drop across the input resistance R remains approximately constant during the interval the switch is closed, and therefore the current through R into C is very nearly constant. Under these conditions, the output sawtooth has a slope which can be determined from Eq. (11.5).

$$\frac{dV}{dt} \simeq \frac{-E}{RC} \qquad\qquad (11.5)$$

where E is in volts, R is in ohms, C is in farads, and dV/dt is the slope of the output sawtooth in volts per second.

The circuit configuration shown in Fig. 11.10a is typical of the integrators used in analog computers (see Sec. 19). In most computer applications, amplifiers which have extremely high gain and which are very stable must be employed to realize the desired integration accuracies.

In Fig. 11.10b is a simple Miller integrator. The input tube $V1$ is used as the switch tube and is enabled by a negative input gate having a duration equal to that of the sawtooth to be generated. In Eq. (11.5), the value of E is equal to $E_{bb} - E_{bo1}$
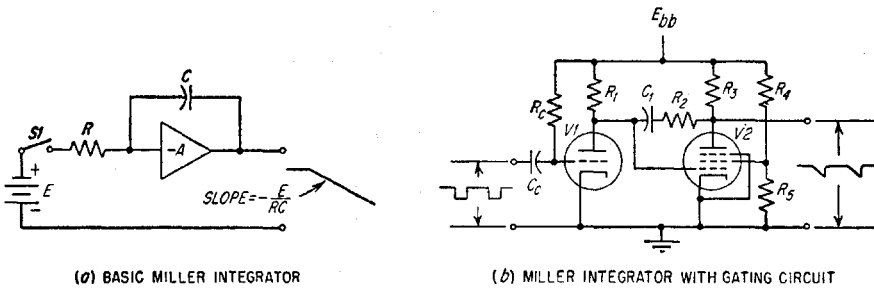
(a) BASIC MILLER INTEGRATOR   (b) MILLER INTEGRATOR WITH GATING CIRCUIT

Fig. 11.10. Driven sawtooth generator using a Miller integrator.

where E_{bo1} is the quiescent plate voltage on $V1$. The value of RC in Eq. (11.5) is equal to R_1C_1. The value of R_2 establishes whether or not there is a step associated with the output sawtooth. The polarity of the step can be positive or negative depending on the value of R_2. To eliminate the step, R_2 should be approximately equal to $1/g_{m2}$. The interval following the sawtooth required for the recovery of the circuit can be reduced by substituting a cathode follower for R_2.

6. *Automatic Shutoff of Driven Sawtooth Generators.* There are many applications in which the slope of the sawtooth must be continuously variable between wide limits. In most applications of this type there is frequently a requirement for maintaining the same peak value for the sawtooth regardless of its slope. If the gate length is held constant, the peak value of those sawtooth waveforms having large slopes is much larger than those sawtooth waveforms having much smaller slopes. However, if the gate-length control potentiometer is ganged with the slope control potentiometer so as to decrease the length of the gate as the slope is increased, it is possible to maintain an approximately constant peak value for the sawtooth regardless of the slope. The primary disadvantage of this technique is that it is frequently necessary to use precision parts and nonlinear potentiometers.

Often the problem of limiting the peak sawtooth value to some constant value is more easily solved by employing automatic shutoff circuits. The process of achieving automatic shutoff is usually accomplished by detecting the desired peak sawtooth value and feeding a control signal back to the gate generator, thereby shutting off the gate generator, which in turn causes the sawtooth to be terminated.

The control signal which determines the end of the sweep-generator enabling gate can be in the form of a trigger or a more slowly changing voltage which at some critical magnitude causes the gate generator to switch to the quiescent condition. Although many different types of gate circuits are employed, one that can be considered as being representative is the monostable multivibrator circuit, i.e., a circuit which has one stable and one unstable position (see Secs. 8.3 and 8.4). During the creation of the enabling gate, the multivibrator circuit is in the unstable position and returns to the stable position when the negatively driven grid returns in the positive direction to that value of bias which again initiates plate current flow. The length of the generated gate is dependent both on circuit time constants and voltages. It is possible to restore prematurely the multivibrator circuit to the stable position by the application of a sufficiently large positive trigger to the negatively driven grid. The required amplitude of the restoring trigger becomes larger and larger if the gate is to be made shorter and shorter with respect to its natural period. It is also possible to restore prematurely the multivibrator circuit to the stable position, thereby shutting off the gate, with a more slowly changing voltage, which gradually brings the grid of the cutoff tube into the region which causes plate current to flow. As before, the start of plate current flow in the cutoff tube initiates the regenerative cycle which restores the gate circuit to the stable position.

A diode biased at the desired sawtooth amplitude level can be used to initiate the desired automatic shutoff action. If the slope of the detected sawtooth is sufficiently large, the sawtooth output of the biased diode can be used to trigger a blocking oscillator for shutting off the gate circuit. In the case of a sawtooth having a small slope it may be necessary to amplify the detected signal before the blocking oscillator can be triggered. As previously stated, it is not always necessary to create a restoring trigger. The detected sawtooth, amplified and/or inverted as required, can be coupled back into the gate generator in any of several ways to terminate the enabling gate.

As is typical of most functions which are to be accomplished electronically, there are many ways of obtaining the same end result, e.g., the gate generator is frequently not a monostable multivibrator. Consequently, the methods described for automatic shutoff should not be considered as being complete but only as being representative.

11.2. Basic Considerations for Beam Deflection in Electromagnetic Cathode-ray Tubes. In an electromagnetic deflection system the electron beam is deflected in a direction perpendicular to the lines of magnetic flux by an amount proportional to the flux density in the magnetic field. Assuming a uniform flux field, the beam will be deflected an amount which is proportional to the current generating the flux. The linearity of the beam deflection in such a system is therefore dependent on the linearity of the sawtooth of current through the coil. If a deflection coil has neither resistance nor distributed capacitance, a rectangularly shaped voltage waveform is required to cause a linear sawtooth of current to flow through the coil. The waveshape of the impressed voltage required to cause a linear sawtooth of current to flow through a practical deflection coil is shown in Fig. 11.11. The triangular portion of the waveform represents the voltage waveform required to overcome the iR drop in the deflection coil. The spike on the leading edge of the voltage waveform is required to charge the distributed capacitance across the deflection coil. If the spike is absent, the slope of the first portion of the current sawtooth is reduced. If the total sawtooth length is reduced, the importance of a properly shaped charging spike becomes greater and greater since a larger percentage of the total sweep period might be distorted.

FIG. 11.11. Voltage waveform necessary to cause a linear sawtooth of current to flow through a physically realizable deflection coil.

The sawtooth generators discussed in Figs. 11.1 to 11.10 generate voltage sawtooth waveforms, and consequently the resulting waveforms cannot be applied directly to a deflection coil to produce a linear sweep.   There are, however, two basic methods for developing the desired voltage across the deflection coil.

One of these methods is to employ current feedback, thereby causing the current which flows through the deflection coil to tend to have the same waveshape as an input linear-voltage sawtooth.   A circuit of this type is shown in Fig. 11.12 where $V2$ is normally biased beyond cutoff to prevent current flow through the deflection coil between successive sweeps.   For this reason, it is necessary that the input sawtooth voltage have a step as shown in the figure.[1]   A small capacitor C is usually placed across the feedback circuit to attenuate the high-frequency components of the waveform in the feedback path.   In this manner, the leading edge of the trapezoidal waveform driving $V2$ is emphasized to create the desired spike. A resistor is usually placed across the deflection coil to dampen the high induced

FIG. 11.12. Current feedback for linearizing the sawtooth of current through a deflection coil.

voltage which appears across the coil at the termination of the current sawtooth through the coil.

The second method sometimes used to create the desired current sawtooth through the deflection coil is to combine a sawtooth having a step with a separately generated voltage spike and to apply the resulting waveform to the grid of the tube having the deflection coil in its plate or cathode circuit.   In this application, however, the tube used to drive the deflection coil must have linear characteristics, and the input waveform must be very carefully generated to produce the desired sweep linearity.

11.3. Basic Considerations for Beam Deflection in Electrostatic Cathode-ray Tubes. The typical electrostatic cathode-ray tube has two sets of deflection plates, one horizontal and one vertical.   Each set consists of one conductive plate on one side of the tube axis and a second conductive plate located on the opposite side of the tube axis. The relative positions of the other deflection plates are identical except that they are rotated around the tube axis by 90°.   Since the electron beam is directed very nearly along the axis of the tube, it will pass between the two plates of each set (see Fig. 2.47).

Considering one set of deflection plates, the electron-beam axis will not be deflected if the voltages on the two plates are equal.   If the voltages on the two plates are not equal, the electron beam will be deflected in the direction of the more positive plate. To minimize defocusing of the electron beam, the average of the signal voltages on the two plates, i.e., $(E_{H1} + E_{H2})/2$, should be constant and equal to some critical value which is dependent on both the average voltage of the other set of deflection plates and on the values of the voltages applied to the first and second anodes.   In practice, an attempt is usually made to make the mean voltage on one set of deflection plates very nearly equal to the mean voltage on the other set of deflection plates.   Since it is not necessary that they be exactly the same, no adjustment is normally provided for adjusting the mean voltage on either set of deflection plates.   If the average voltage on one set of plates is reasonably close to the average voltage on the other set, e.g., within 50 volts or less, it is possible to achieve proper focusing of the electron beam in

[1] With reference to the voltage sawtooth generators shown in Figs. 11.4 to 11.10, a step, or pedestal, can be added to the sawtooth by placing a small resistor in series with the charging capacitor.

both the vertical and horizontal planes by a proper adjustment of the first and second anode voltages. The resulting second-anode voltage will be reasonably close to the average voltages on the deflection plates.

Since the beam is deflected an amount which is proportional to the voltage difference between the deflection plates, a sawtooth voltage applied to only one of the deflection plates will cause the beam to be deflected an amount which increases linearly with time. Consequently, any of the sawtooth generators discussed in the
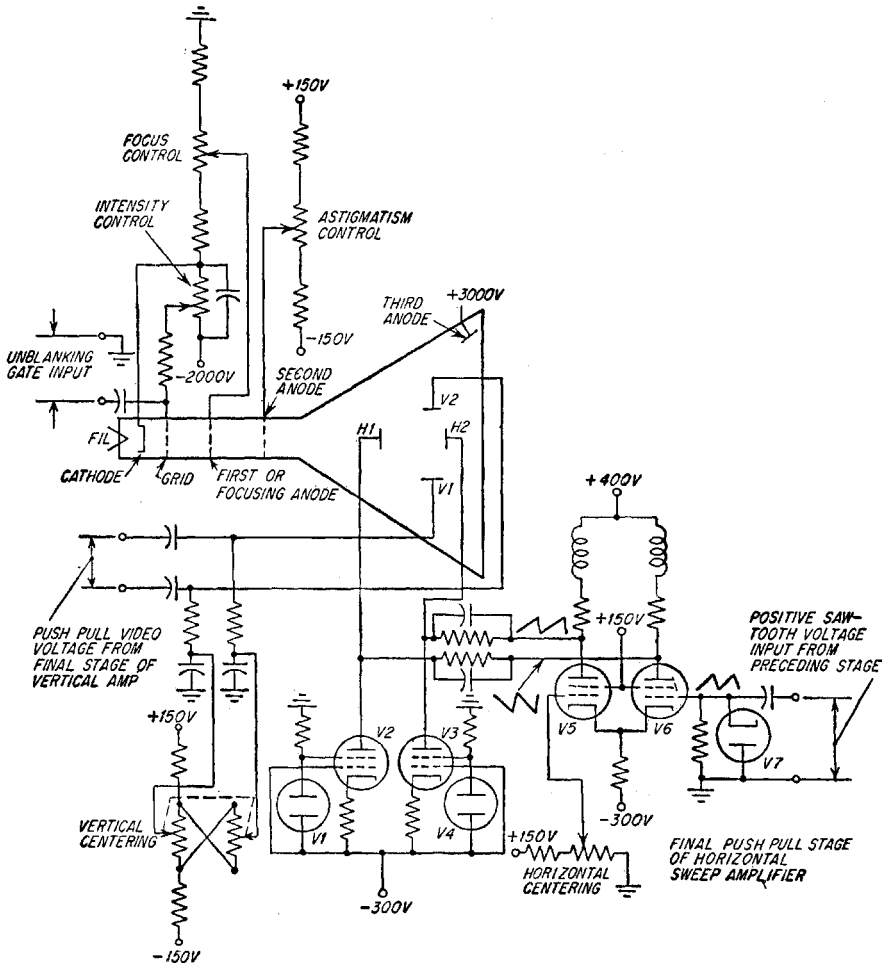
Fig. 11.13. Electrostatic cathode-ray tube with typical oscilloscope control circuits.

preceding sections, with an appropriate amplifier if required, can be used as the sweep circuit for an electrostatic cathode-ray tube. There is a distinct disadvantage, however, in applying the sweep- or beam-deflecting voltage to only one of the deflecting plates, since this tends to cause beam defocusing because the mean voltage of the two plates is not constant. This effect becomes especially severe when large deflection signals are required. For this reason, push-pull circuits for driving the deflection plates are preferred. A certain amount of defocusing exists even with push-pull

deflection because the voltage gradient which exists between the two plates causes a nonuniform acceleration of the electron beam.

The circuit shown in Fig. 11.13 is typical of the deflection systems used in cathode-ray tubes. It should be noted that there are no adjustments for establishing the mean voltage on either set of deflection plates. Instead, these voltages are established by fixed resistors and tube characteristics. Adjustment of either the vertical or horiontal centering control does not change the mean voltage on the respective deflection plates. Operation of one of these controls merely causes the voltage on one plate to increase and the voltage on the other to decrease an equal amount. The average does not change. The presence of a signal on either set of plates does not cause a change in the average voltage.

The horizontal sweep amplifier is d-c–coupled to the deflection plates to permit clamping the sweep level by $V7$ at the input grid circuit of the push-pull amplifier. Both $V2$ and $V3$ are used to establish the d-c level of the horizontal plates at a mean voltage which is very nearly equal to the mean voltage of the vertical plates. These tubes provide a high a-c impedance path to the sweep voltage and, consequently, permit a high efficiency transfer of the signal from the push-pull stage to the deflection plates.

The astigmatism and focus controls are used to adjust the voltages on the first and second anodes to the desired values for the particular mean voltages on the two sets of deflection plates. With proper adjustment it will be found that the critical voltage on the second anode for proper focusing is within 100 volts or less of the average voltages on the deflection plates.

High-speed sawtooth waveforms, i.e., having very large values of dV/dt, can be obtained by utilizing the plate voltage of a gas tube during the discharging period of the tube, the plate voltage of a blocking oscillator during the "turning on" period, or simple circuits of the type shown in Fig. 11.6 where very short time constants and gate lengths are used. As in any other sweep application, the cathode-ray tube is intensified by the application of an unblanking gate on the grid during the period of the desired deflection voltage.

Example 11.1

Design a bootstrapped sawtooth generator of the type of Fig. 11.8 to generate a sawtooth with a slope of 10^6 volts/sec and a length of 10 μsec. Assume that the available supply voltages are $+300$ and -200 volts and that the repetition frequency of the sawtooth is to be 1 kc.

Solution

1. Determine the value of R_1 and I_{bo}.

The value of I_{bo} is the quiescent current through $V1$. If R_1 is made sufficiently large, I_{bo} will very nearly be equal to E_{bb}/R_1 since $V1$ is at zero bias. Therefore, let R_1 have a value of 300,000 ohms.

$$I_{bo} \simeq \frac{300}{300,000} = 1 \times 10^{-3} \text{ amp}$$
$$= 1 \text{ ma}$$

NOTE: Since R_1 is usually a variable quantity, let R_1 consist of a 250,000-ohm fixed resistor in series with a 100,000-ohm variable resistor.

2. Determine the required value of C_1.

From Eq. (11.3)

$$C_1 = \frac{1 \times 10^{-3}}{10^6}$$
$$= 1,000 \times 10^{-12} \text{ farads}$$
$$= 1,000 \ \mu\mu f$$

NOTE: In paragraph 3 of Sec. 11.1c it was emphasized that the time constant R_1C_1 should be at least ten times as large as the length of the sawtooth waveform if particularly good

linearity is desired. In this example the ratio is 30:1, consequently the values for R_1 and C_1 are satisfactory. If the ratio had calculated to be less than approximately 10:1, a considerable improvement in linearity could have been obtained by increasing either R_1 or C_1. Associated with the improvement in linearity would be a decrease in the amplitude of the generated sawtooth waveform.

3. Determine the tube type of $V1$ and $V3$.

The average current through $V1$ and $V3$ is approximately 1 ma, and receiving-type tubes can be used. The gate tube $V1$ can be either a high μ triode or pentode. If the enabling gate has high-frequency components on the portion of the gate which is normally considered to be flat, feedthrough may be a problem and a pentode would be preferable because of its smaller grid-to-plate interelectrode capacitance.

4. Determine the desired size of R_2 and C_2.

During the generation of the sawtooth the voltage across C_2 will decrease by an amount Δe which is approximately equal to the peak value of the sawtooth E multiplied by C_1/C_2. Therefore, if a maximum discharge of 1 volt across C_2 is permitted, the size of C_2 can be calculated as follows:

$$\Delta e = E \frac{C_1}{C_2}$$

$$1 = 10 \times \frac{1,000 \times 10^{-12}}{C_2}$$

$$C_2 = 10,000 \times 10^{-12} \text{ farads}$$
$$= 0.01 \ \mu f$$

The repetition rate has been specified as 1 kc, which is equivalent to 1,000 μsec between the starts of recurrent sawtooth waveforms. Since the sawtooth duration is 10 μsec, the recovery period is 990 μsec. If the repetition frequency were variable, it would be important that the charge on C_2, which is lost during the generation of each sawtooth, be completely replaced in the shortest period which might exist between recurrent sawtooth waveforms. If the repetition frequency is constant, this is not of importance. The factors which minimize the time required for recharging C_2 are a low value for R_2, returning R_2 to a large negative voltage, and a low cathode-follower output resistance. In this particular example, R_2 can be returned to the -200-volt supply. Since the grid of $V2$ is slightly above ground potentital, the cathode will also be slightly above ground potential, and the plate-cathode voltage is therefore known to be somewhat less than 300 volts. If $V2$ is chosen to be one-half of a 12AX7, the quiescent current could be as great as 2 ma without causing excessive plate dissipation. The voltage drop across R_2 at quiescence will be about 205 to 230 volts since the cathode will be at some small voltage above ground. To achieve a plate current of 2 ma, R_2 must be in the order of 100,000 ohms. The cathode-follower gain with a 100,000-ohm cathode resistor is 0.98. The gain of the same cathode follower with an infinitely large cathode resistor would be 0.99; consequently, very little has been lost in gain by using the smallest possible cathode resistor which would not cause excessive plate dissipation.

Clippers, Limiters, and Clamps

12.1. Clippers. An ideal clipper circuit transfers only those portions of a signal which are either more positive or negative than some specific voltage. Depending on the circuit configurations, a clipper is therefore able to discriminate against signals either above or below some threshold level. A biased thermionic diode, crystal diode, or multielement vacuum tube is usually employed as the clipping element.

12.1a. Diode Clippers. Figure 12.1 illustrates four basic circuit configurations which employ a diode as either a series or a shunt clipping element. The circuits in Fig. 12.1a and c pass those signal components which are more negative than the threshold potential E_T; those in Fig. 12.1b and d pass only those components which are more positive than E_T. The potential E_T may be either positive or negative for either the series or shunt type of clipper. In a practical application, E_T may be obtained from a voltage dividing network as shown in Fig. 12.2a and b. The relative

SERIES DIODE CLIPPERS

SHUNT DIODE CLIPPERS
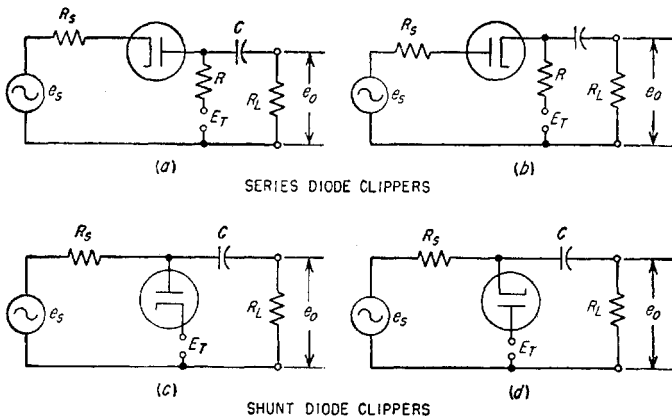
Fɪɢ. 12.1. Diode clippers.

values of R_1 and R_2 must be selected to provide the proper threshold potential, and the value of R_1 and R_2 in parallel should be (a) large compared to R_s in parallel with R_L in the series circuit and (b) small compared to R_s in parallel with R_L in the shunt circuit. Coupling capacitor C in Fig. 12.2b can be eliminated where d-c isolation is not required between the load and the source. If it is desired to use the circuit in Fig. 12.2a without coupling capacitor C, R_2 should also be eliminated and the supply voltage E_{bb} should be made high enough to permit R_1 to be large with respect to R_L. In this case, the clipping threshold is established by the divider action of R_1 and R_L.

Series Clippers. During the conduction period of the series diode clipper, the maximum output signal is realized when the sum of the source resistance R_s and the d-c diode resistance R_d is small compared to the load resistance R_L in parallel with the effective resistance in the branch which establishes the threshold potential E_T.

To analyze the performance of a series diode clipper, consider the circuit shown in Fig. 12.2a. It is necessary to know the voltage and waveform of the source voltage e_s. The instantaneous difference between the source voltage e_s and the threshold

voltage E_T is designated as e' and is the output signal voltage which would exist if the output circuit were not loaded by R_1, R_2, and R_L.

$$e' = e_s - E_T \qquad (12.1)$$

Since the signal source is loaded by R_1, R_2, and R_L during the conduction of the diode, the output voltage e_o is less than e' because of the voltage drops across R_s and the diode. The diode current i_d will vary in a nonlinear manner with respect to the magnitude of e' because of the nonlinear diode characteristics. In Fig. 12.2c is shown a typical diode E_b-I_b characteristic. To determine the diode current i_d, a
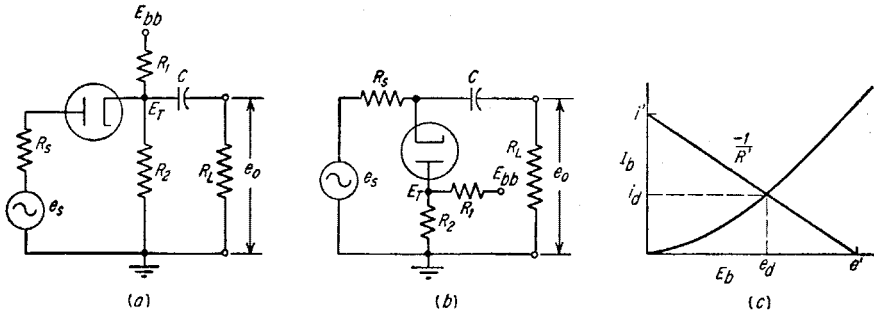
FIG. 12.2. Clipper circuits and diode E_b-I_b characteristic.

load line for the total circuit resistance R' must be plotted on E_b-I_b characteristic. R' can be calculated from Eq. (12.2).

$$R' = R_s + \frac{R_1 R_2 R_L}{R_1 R_2 + R_1 R_L + R_2 R_L} \qquad \text{(series clipper)} \qquad (12.2)$$

The load line will have a slope equal to $-1/R'$ and will intercept the E_b coordinate at e' and the I_b coordinate at i'. The value of i' is equal to e'/R'. The intersection of the load line and the diode curve determines the diode current i_d. The output voltage e_o can then be determined from Eq. (12.3).

$$e_o = i_d \frac{R_1 R_2 R_L}{R_1 R_2 + R_1 R_L + R_2 R_L} \qquad \text{(series clipper)} \qquad (12.3)$$

To determine the waveform distortion due to the nonlinearity of the diode, the values of e_o associated with a representative number of values of e_s can be established and plotted.

Shunt Clippers. With the shunt diode clipper in Fig. 12.2b, sharp amplitude discrimination will be realized if R_s in parallel with R_L is large compared to the sum of the diode resistance R_d and the parallel combination of R_1 and R_2. If the source resistance R_s is not sufficiently high to permit proper clipping, additional resistance can be added in series with the source voltage e_s. The input signal must be effectively clipped at the value of E_T when the diode conducts if ideal clipping is to be obtained. When the shunt diode is nonconducting, the output voltage e_o will be proportional to the source voltage e_s and will have a magnitude determined by the voltage divider composed of R_s and R_L. This relationship is given by Eq. (12.4).

$$e_o = e_s \frac{R_L}{R_L + R_s} \qquad \text{(shunt clipper with diode not conducting)} \qquad (12.4)$$

The presence of R_1 and R_2 and the fact that the diode has resistance cause imperfect clipping of an input signal for the range of input voltages which causes the diode to

conduct.   The output voltage e_o', which exists because of incomplete clipping, can be determined with the aid of the diode E_b-I_b characteristic.   For the shunt diode clipper, the value of the total circuit resistance R' in series with the diode is given by Eq. (12.5) (see Fig. 12.2b).

$$R' = \frac{R_L R_s}{R_L + R_s} + \frac{R_1 R_2}{R_1 + R_2} \qquad \text{(shunt clipper)} \qquad (12.5)$$

The voltage and current intercepts on the E_b and I_b coordinates shown in Fig. 12.2c are given by Eqs. (12.6) and (12.7).

$$e' = E_T - \frac{e_s R_L}{R_L + R_s} \qquad \text{(shunt clipper)} \qquad (12.6)$$

$$i' = \frac{e'}{R'} \qquad (12.7)$$

The intersection of the load line of slope equal to $-1/R'$ and the diode curve determines the voltage e_d across the diode and the diode current i_d for a specific value of e_s. The value of the output voltage e_o' during the clipping period is given by Eq. (12.8).

$$e_o' = e_d + i_d \frac{R_1 R_2}{R_1 + R_2} \qquad \text{(shunt clipper with diode conducting)} \qquad (12.8)$$

By determining values of e_o' for several values of e_s which cause the diode to conduct, a plot of e_o' versus e_s for the diode conduction interval can be made.   The values for e_o during the period in which the diode is not conducting can be calculated by Eq. (12.4).

In the analysis of both shunt and series diode clippers, the reactance of coupling capacitor C at the operating frequency is assumed to be negligible.   A further assumption is that the shunting capacitance across the load also has negligible effect.

Both shunt and series clippers have been analyzed on the basis that the diodes are either of the thermionic or silicon crystal type.   If germanium crystal diodes are used, it may be necessary to employ graphical solutions for those intervals in which the diodes are biased in the inverse direction, since the back resistance of a germanium diode may be low enough to modify materially the action of the clipper.

12.1b. Multielement Tube Clippers.   In a multielement sharp-cutoff tube, the control grid can be biased so that the tube functions as a signal clipper.   If the tube is biased beyond cutoff, only those portions of the grid-input signal sufficiently positive to cause plate current to flow are present in the output.   In Fig. 12.3a is the circuit configuration of a triode clipper which also acts as an amplifier and an inverter. With the exception of R_c, no loading of the source occurs provided that the positive peak value of the input signal does not cause the grid of the tube to conduct.

This type of circuit can be analyzed by use of the dynamic transfer characteristic (see Sec. 3.3b) illustrated in Fig. 12.3b.   An ordinate should first be drawn through the abscissa at the value of the grid bias E_c.   If the input waveform, e.g., a sine wave, is drawn to scale along this ordinate, those positive portions of the input signal which cause the tube to conduct can be transferred to the plate current scale point by point to establish the output current waveform.   It can be seen that the nonlinearity in the cutoff region of the triode causes a certain amount of distortion.   The instantaneous value of i_o multiplied by R_b in parallel with R_L will give the instantaneous output voltage e_o.

For signals which have positive components sufficiently large to cause grid current to flow, the operation is somewhat different.   If R_c has a large value and if the time constant $R_c C_c$ is much longer than the period of the input-signal waveform, the clipper will operate as a constant amplitude clipper.   This type of operation is obtained because the grid current flows when the control grid is driven positive.   The voltage across R_c, resulting from the flow of grid current through R_c, will increase the effective

bias to some new value E_c'.   Although E_c' will change as a function of input-signal amplitude, it will always have a value slightly less than that of the positive input-signal peak.   The region of tube conduction is, therefore, limited to the instantaneous values of the input signal which are between the values which cause the grid to conduct and the plate current to be cut off.   This special case of the triode clipper can also be classified as a limiter since the amplitude of the output is very nearly independent of the input amplitude.
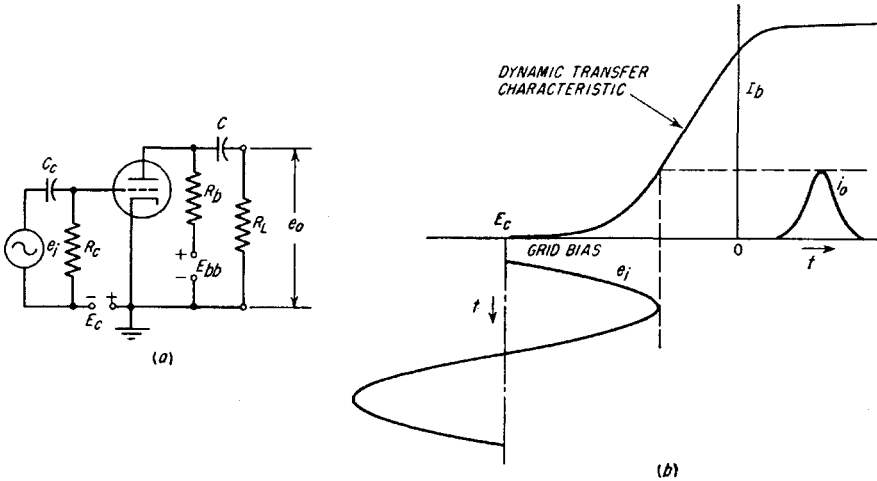
FIG. 12.3. Triode clipper circuit and graphical analysis.

12.2. Limiters.   A limiter is used to restrict the peak-to-peak or peak amplitude of a signal to a specified maximum value.   An ideal limiter has a constant ratio of output to input voltage for all values of the input signal up to the limiting threshold, and above this point the ratio of the incremental change in the output voltage to an incremental change in the input voltage abruptly assumes a value of zero.   A *factor of merit* σ for limiters can be used to compare the performance of various circuits.

$$\sigma = \frac{A}{A'} \qquad (12.9)$$

where A = ratio of incremental change in output voltage e_o to incremental increase
        in input voltage for case where $e_o < E_T$
  A' = ratio of incremental change in output voltage e_o to incremental increase in
        input voltage for case where $e_o > E_T$
  E_T = nominal limiting value of e_o

Limiting can be achieved in a number of different circuit configurations and by using the special characteristics of several different circuit elements.   Among these are nonlinear resistances, the voltage breakdown characteristics of normally non-conducting elements, diodes, and multielement vacuum tubes.   Limiter circuits generate considerable distortion in the output signal when the signal amplitude greatly exceeds the limiting threshold.   The generation of a square wave from an input sine wave is an example of the extreme waveform distortion which is often realized in a limiter.

12.2a. Limiting Circuits Employing Nonlinear Resistances.   Limiting can be accomplished in a passive network containing nonlinear resistances.   In general, the quality of the limiting achieved by this method is relatively poor since the change in

resistance as a function of either voltage or current changes is gradual rather than abrupt as desired.

Nonlinear resistances have a voltage-current relationship which can be expressed by Eq. (12.10).

$$I = kE^n \tag{12.10}$$

where k = a constant equal to initial conductance when voltage is zero (or very small)

n = an exponent which defines nonlinearity of resistance

Thyrite is an example of a nonlinear resistance which has a value of n greater than unity. The d-c resistance of this material decreases with an increase in either current or terminal voltage. A circuit for deriving a limiting action from a nonlinear resistance of this type is shown in Fig. 12.4a. The sum of R_s and R_1 should be very large compared to the maximum value of the nonlinear resistance. The performance of this type of circuit can be determined graphically if a plot of the E-I characteristic of the nonlinear resistance is constructed. A curve which is typical of this general class of material is shown in Fig. 12.4b Using the same coordinates, it is possible to plot a load line for the circuit illustrated. The load line has a slope of $-1/(R_s + R_1)$ and is plotted by assuming a value for e_s which establishes the intercept on the E coordinate. The load line intercept point on the I coordinate is equal to $e_s/(R_s + R_1)$.
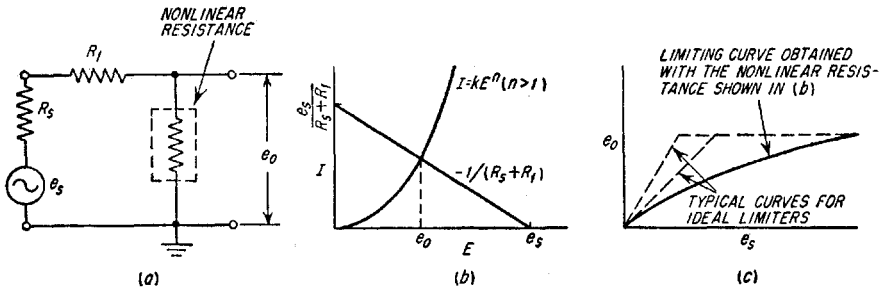
FIG. 12.4. Graphical analysis of a nonlinear resistance limiter in which $n > 1$.

The intersection of the load line and the curve determines the output voltage e_o across the nonlinear resistance for the assumed value of e_s. By assuming several values for e_s and determining the associated values of e_o, it is possible to plot e_o as a function of e_s as shown in Fig. 12.4c. When the limiting is this gradual, the factor of merit σ has little meaning except to relate the slope near the zero value of e_s to the slope at some increased value of e_s.

Nonlinear resistances which have a value of n smaller than unity increase in resistance with an increase of either current or terminal voltage. This characteristic exists in most materials. To obtain limiting with this type of element, the circuit shown in Fig. 12.5a can be used. The load line has a slope of $-1/(R_s + R_2)$ and can be drawn by the method described for the circuit shown in Fig. 12.4b. In this manner, the current i_r through the nonlinear resistance for a specific value of e_s can be established. The output voltage e_o is, therefore, equal to $i_r R_2$. In all other respects the graphical solution follows the method given for the circuit shown in Fig. 12.4. Optimum performance is obtained if the value of $R_s + R_2$ is small compared to the minimum value of the nonlinear resistance.

With the circuit of Fig. 12.6a, a pair of thermionic diodes is used as a limiter. This circuit can be analyzed by the technique employed for the circuit of Fig. 12.4a. The value of n for the diodes is approximately 1.5. A variation of this circuit, shown in Fig. 12.6b, is sometimes used as a transient (noise) limiter and does not appreciably limit the input signal. For this application, R_2 should be much greater than R_s +

$R_1 + R_d$, and the time constant R_2C_2 should be much longer than the period of the input signal e_s. R_d is the d-c resistance of the diode. When these conditions are satisfied, the voltage E_T becomes very nearly equal to the peak-to-peak value of e_s and only slight loading of the steady-state signal is observed. Any abrupt increase in the input peak amplitude, such as a noise transient, is severely limited. During the transient limiting interval, the circuit can be analyzed in the same manner as that of Fig. 12.6a except that the effective limiter input should be considered to be the difference between the instantaneous sum of the noise and signal amplitudes and the steady-state value of E_T.
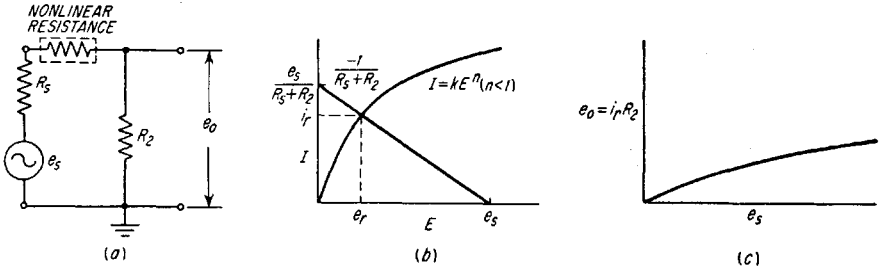
FIG. 12.5. Graphical analysis of a nonlinear resistance limiter in which $n < 1$.

Nonlinear resistance types of limiters can be cascaded to achieve improved limiting. By various combinations of linear and nonlinear elements, a great variety of transfer functions can be obtained.

12.2b. Limiting by Voltage Breakdown. One type of limiter employs an element in which conduction is initiated at some critical voltage. An example of such an element is a bidirectional gas tube.

The circuit of a gas-tube limiter is shown in Fig. 12.7a. The characteristic performance of this circuit with an increasing value of e_s is plotted in Fig. 12.7b. It will

(*a*) BACK-TO-BACK DIODES   (*b*) TRANSIENT LIMITER
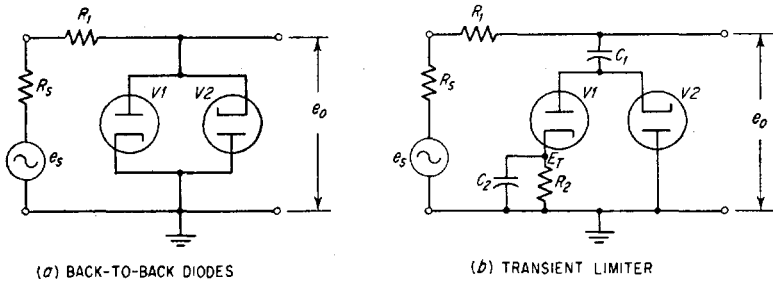
FIG. 12.6. Diode limiters.

be observed that the output voltage e_o has the same value as e_s until the gas-tube firing voltage E_F is reached. At this point, conduction is initiated in the gas tube and the output drops abruptly from the firing voltage E_F to the regulating voltage E_R. The value of e_o associated with a further increase in e_s is best determined by graphical means. A load line with a slope of $-1/R_s$ should be drawn on the gas-tube characteristics as shown in Fig. 12.7c. The output voltage e_o for an assumed value of e_s is that voltage at which the load line intersects the curve. The output waveform for any given input waveform can be determined if a representative number of values of e_s are considered.

In dynamic operation, the difference in the firing potential E_F and the conducting potential E_R results in an overshoot before limiting occurs. In Fig. 12.7d is a typical

output waveform for a sine wave input. The waveform curvature in the limited region is a function of R_s and the positive[1] dynamic resistance of the gas tube. The curvature can be minimized by making R_s large and by using a gas tube with a low dynamic plate resistance.

Since a gas tube has a finite time of ionization and deionization, its performance as a limiter will change with the frequency of the input waveform. For operation at
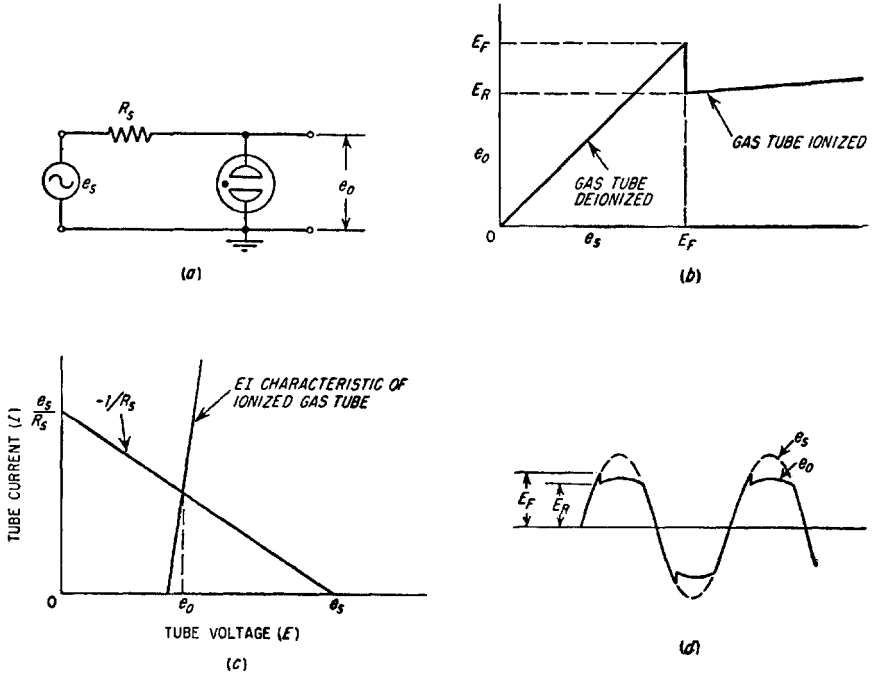
FIG. 12.7. Graphical analysis of a gas-tube limiter.

(a) SHUNT LIMITER   (b) SERIES LIMITER
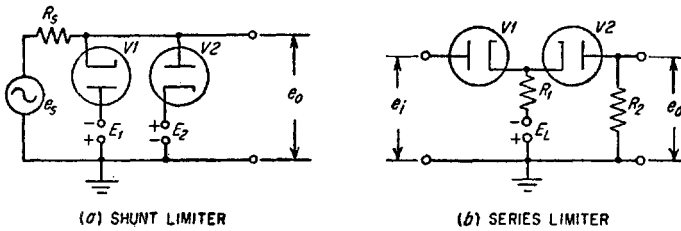
FIG. 12.8. Biased diode limiters.

frequencies up to several hundred cycles per second, it is possible to use the static E-I characteristic; for operation at higher frequencies, it is recommended that the performance of the gas-tube limiter be determined empirically.

12.2c. Biased Diode Limiters. The two basic circuit configurations for biased diode limiters are shown in Fig. 12.8. With both of these circuits, the amplitude of the input signal must exceed specific threshold values before limiting begins.

The performance of the shunt circuit in Fig. 12.8a can be evaluated by the method given for the circuit of Fig. 12.2b for those values of e_i which are more negative than

[1] Some gas tubes have a negative dynamic resistance.

E_1 and more positive than E_2. If e_i is more positive than E_1 and more negative than E_2, there is no limiting. If the diodes are identical and if $E_1 = -E_2$, a solution for only one polarity is necessary since the limiting will be symmetrical about ground potential.

In the series limiter, Fig. 12.8b, limiting is achieved when either $V1$ or $V2$ ceases to conduct, thereby causing the signal path to be open-circuited except for the tube interelectrode capacitance. For example, the limit to which the plate of $V2$ can be driven in the positive direction is ground potential since at this potential $V2$ stops conducting. The negative limit to which the plate of $V2$ can be driven occurs when the input signal is sufficiently negative to cut off $V1$. When $V1$ is cut off, the potential at the plate of $V2$ can be established by determining the current through $V2$ when conducting alone and then determining the voltage drop across R_2. The current through $V2$ when conducting alone can be determined by drawing a load line on the diode characteristic with a slope of $-1/(R_1 + R_2)$ from the voltage E_L on the abscissa. The point of intersection of the load line and the diode curve determines the current i_d' through $V2$ when $V1$ is not conducting. The limit to which the plate of $V2$ can be driven in the negative direction is, therefore, $-i_d'R_2$. In summation, the limited, output signal at the plate of $V2$ will have the limit values of 0 and $-i_d'R_2$ volts. The absolute magnitude of the limited output signal is, therefore, equal to $i_d'R_2$ volts. The minimum peak-to-peak value of the input signal e_i required to cause the output signal to reach the limit levels is equal to $e_d + |E_L| - i_d'R_1$ volts where e_d is the voltage drop across $V1$ for a tube current equal to E_L/R_1.

The polarity of E_L can be reversed if the diodes are reversed. For either polarity, however, the limiting level increases with the absolute magnitude of E_L.

This particular circuit is sometimes used as a modulator. The modulation voltage is substituted in lieu of E_L and the carrier is substituted for e_i. When used for this purpose, it is usually desirable to filter the output signal by a low-pass or a band-pass filter since the limiter introduces many harmonics due to the severe waveform distortion.

12.2d. Multielement Vacuum-tube Limiters. When it is necessary to limit small signals, multielement vacuum-tube limiters are usually employed. Limiting in grid-controlled tubes can occur (1) if plate current cutoff is reached when the grid is driven in the negative direction, (2) if the signal source is severely loaded when the grid is driven in the positive region, and (3) if the *maximum plate-current curve*[1] for the tube is intercepted. Consequently, limiting can be achieved by overdriving cascaded amplifier stages.

Positive Signal Limiting in the Positive Grid Region. Appreciable grid current starts to flow at or near a grid-to-cathode voltage of 0 volts. For most receiving-type tubes, the value of the grid current is 0.5 to 2 ma/volt for positive input signals. The result of initiating grid current is equivalent to loading the input-signal source with a resistance of 500 to 2,000 ohms whenever the positive grid condition is reached. If the signal source has a high internal impedance or if the source impedance is artificially augmented by the addition of R_1 as shown in Fig. 12.9a, the positive signal limiting due to the grid loading can be very effective.

When the signal source impedance is sufficiently low that it is not loaded appreciably by the conducting grid, limiting occurs at a higher input-signal level. In this case limiting occurs when the tube reaches its maximum degree of conduction. In all vacuum tubes there exists, for every value of plate voltage, a value of positive grid bias above which any further increase in the positive grid voltage will cause no increase in plate current. This phenomenon is observable only in the plate circuit since the

---

[1] The maximum plate-current curve for a multielement tube defines the maximum possible plate current as a function of plate voltage, i.e., the plate-current versus plate-voltage curve which is independent of any further increase in the positive grid potential.

input signal voltage is relatively unaffected. In Fig. 12.9b is shown the plate characteristics of a triode illustrating this effect. The locus of the maximum current values versus plate voltage is termed the maximum plate-current curve for the tube. For a given plate supply voltage E_{bb} and a-c plate-load resistance, the maximum possible plate current is determined by the intersection of the a-c load line and the maximum plate-current curve. In tetrodes and pentodes the intersection of the a-c load line and the maximum plate-current curve can occur at negative grid voltages.
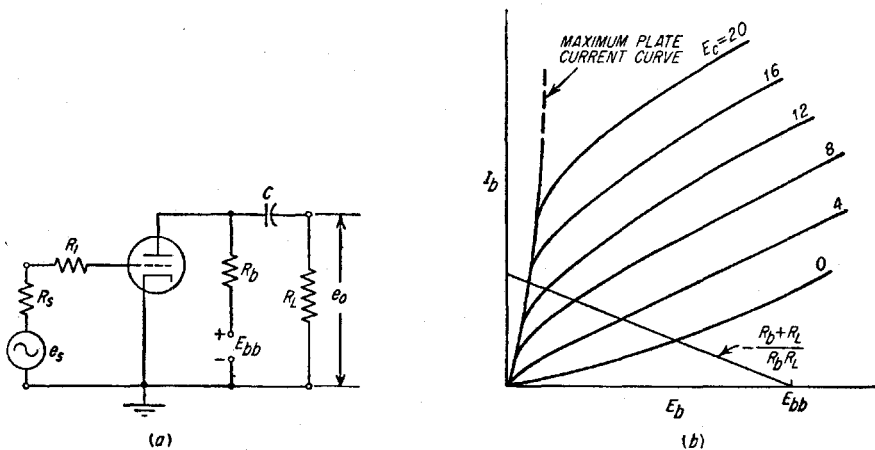
FIG. 12.9. Signal limiting in the positive grid region.

Negative Signal Limiting Due to Plate-current Cutoff. If the signal applied to the grid of a tube exceeds the negative grid bias required for plate-current cutoff, the voltage at the plate of the tube becomes equal to the supply voltage E_{bb}. In general, high-μ triodes or high g_m, sharp-cutoff pentodes are most desirable for this type of limiting. For limiting at low threshold values, a pentode having low screen and plate supply potentials should be used. These low voltages, particularly on the screen, reduce the negative control grid potential required for plate-current cutoff.

An effective limiter can be obtained by utilizing grid-current conduction whenever the grid-cathode voltage is positive and plate-current cutoff when the grid-cathode voltage is sufficiently negative. The circuit of a limiter of this type using a pentode is shown in Fig. 12.10. In this limiter, R_1 should be at least 50 to 100 times greater than the positive input grid resistance, and the time constant R_1C_1 should be large compared to $1/f$ where f is the lowest signal frequency. The value of R_1 is normally made equal to several hundred thousand ohms since the positive input grid resistance may be as high as 2,000 ohms. During the positive peaks of the input signal, grid current will flow, and consequently a charge will be developed on C_1. During the portion of the cycle that the grid is not positive with respect to the cathode, C_1 will discharge at a nearly constant rate through R_1, thereby developing "grid-leak bias" for the tube. The bias developed across R_1 will adjust itself to be slightly less than the magnitude or the positive peak value of the input signal. This is not a sym-
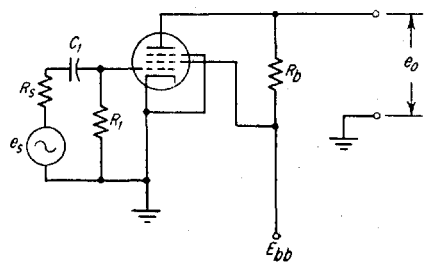
FIG. 12.10. Pentode limiter.

metrical limiter, since the most positive part of the input waveform will be amplified and the remaining portion below the grid cutoff voltage will be limited and will not be present in the output waveform." If the signal is always large compared to the value of the bias required for plate-current cutoff, the major portion of the input signal will be in the tube cutoff region. Both the limiting level and the amplification are proportional to E_{bb}. Since neither the control grid current nor the plate-current characteristics for low screen voltages is normally supplied by the tube manufacturers, the circuit parameters for optimum performance are best determined empirically.

Cathode-coupled Limiter. This excellent limiter is shown in Fig. 12.11. When the amplitude of the input signal is sufficiently negative to cut off $V1$, no further increase in the negative input signal is transferred to $V2$. When the positive excursion of the input signal

FIG. 12.11. Cathode-coupled limiter.

causes the cathode voltage to exceed the cutoff potential for $V2$ no further increase in input amplitude is conveyed to the output. The grounded grid of $V2$ minimizes capacitive feedthrough for those signals in excess of the limiting levels.

The quiescent operating voltages and the signal amplification before limiting can be established by the graphical method described in Sec. 3.16. The positive input signal $+e_i$ which causes limiting is the input signal which drives the cathode sufficiently positive to cut off $V2$. The plate voltage across $V2$ at cutoff is equal to $E_{bb} - E_{k2}$ where E_{k2} is the cathode voltage at cutoff and is equal to the absolute magnitude of the cutoff bias for $V2$. The cutoff bias E_{co2} for $V2$ can be determined by experimentally establishing a value of bias $-E_{k2}$ which will cause plate current cutoff for a plate voltage equal to $E_{bb} - E_{k2}$. After determining E_{k2}, a load line having a slope equal to $-1/R_k$ should be drawn on the E_b-I_b characteristics for $V1$. The positive input voltage $+e_i$ to $V1$ which causes $V2$ to be cut off is equal to $E_{k2} - |E_{c1}|$ where E_{c1} is the bias associated with a plate current equal to E_{k2}/R_k on the load line for $V1$. To determine the negative input signal $-e_i$ which causes limiting, it is necessary to establish the limiter cathode voltage when $V1$ is not conducting. On the E_b-I_b characteristics for $V2$, a bias line (see Sec. 3.3a) and a load line having a slope equal to $-1/(R_k + R_b)$ should be drawn. The intersection of these two lines establishes the current i_2 through R_k when $V1$ is cut off. The plate voltage across $V1$ at cutoff is equal to $E_{bb} - R_k i_2$. From the tube characteristics, the cutoff bias E_{co1} for the plate voltage can be determined. The negative input signal $-e_i$ causing limiting is equal to $-|E_{co1}| + R_k i_2$. The peak-to-peak limited output signal e_o will have a value equal to $i_2 R_b$.

Cascaded Limiters. Cascaded limiter stages produce relatively good over-all limiting. The factor of merit for cascaded limiters is the product of the factors for the individual stages. If only small signals below the limiting level are considered, the slope of the output voltage versus the input voltage represents the normal cascade gain. As the signal level is increased, limiting usually occurs first in the final stage. For this particular input level, the over-all factor of merit is proportional to the factor of merit for the last stage. If the signal level is further increased, limiting will begin in the next to last stage, the stage preceding the next to last stage, etc., until limiting occurs in all stages. For these large input levels, the ratio of the incremental increase in output voltage to an incremental increase in input voltage is very nearly equal to zero.

12.2e. Square-wave Generators. For special applications it is sometimes necessary to generate square or rectangular voltage waveforms. In general, the application of excessively large signals to the input of any amplifier tends to cause the output waveshape to become rectangular. The properties which determine the relative squareness are the characteristics of the input signal, the characteristics of the limiting device and the bandwidth of the system. Based on the assumptions: (1) the input to an amplifier is a sine wave of sufficient amplitude to cause limiting, (2) there is no shift in the operating points of the different stages, and (3) the bandwidth of the amplifier is infinite, the rise time of the rectangular output waveform can be calculated from Eq. (12.11).

$$T = \frac{1}{\pi f} \sin^{-1} \frac{e_o}{3.54 A e_i} \qquad (12.11)$$

where A = gain of amplifier

e_i = rms voltage of input sine wave

e_o = peak-to-peak voltage of output square wave

f = frequency of input sine wave, cps

T = rise time in seconds between 10 and 90 per cent points on peak-to-peak output voltage e_o

Grid current effects can be minimized by d-c coupling or by placing resistors in series with the grids to limit grid current flow. Since the quality of the output square wave is dependent on the bandwidth of the system, the values of the limiting resistors should not be excessive. Common practice is to make the frequency response of the amplifier, between half-power frequencies, approximately $0.1f$ to $10f$.

If an amplifier is used to amplify an input square wave instead of limiting an input sine wave, the rise time of the output square wave can be calculated from Eq. (12.12).

$$T = \sqrt{T_1{}^2 + T_2{}^2} \qquad (12.12)$$

where T = rise time in seconds between 10 and 90 per cent points of output square wave peak-to-peak voltage

T_1 = rise time in seconds between 10 and 90 per cent points of input square wave peak-to-peak voltage

T_2 = output square wave rise time in seconds between 10 and 90 per cent points when input is a perfect square wave

$\simeq \dfrac{0.35}{\beta}$

β = bandwidth of amplifier, cps   (NOTE: If amplifier contains compensation, value of T_2 will be somewhat greater than that given by the equation.)

12.3. Clamps. A clamp is usually employed (1) to establish a prescribed d-c level for a signal which has passed through some type of reactive coupling such as a capacitor or a transformer, (2) to prevent a shift of the operating point in an amplifier due to the signal amplitude and/or waveshape, (3) to sample periodically the instantaneous value of a varying potential, or (4) to maintain a prescribed minimum or maximum potential in a network. These objectives are accomplished by the use of a rapid-acting switch which is usually either a diode or a triode. A special form of clamping circuit can be used for the instantaneous sampling of any signal.

12.3a. D-C Restoration. The interstage coupling most frequently used in vacuum-tube circuits is of the RC type shown in Fig. 12.12. If the input to an RC interstage network is a repetitive unipolar signal, the output signal will be bipolar and the average value of the output voltage will be zero volts. The effect of the coupling capacitor is to remove the d-c component of the input waveform. This is shown in Fig. 12.12 with rectangular waveforms. When the signal duration is equal in the positive and negative periods, the output consists of equal positive and negative voltage excursions.

When the positive and negative periods are unequal, the magnitudes of the positive and negative voltages are related by the reciprocal of the time ratio. For complex repetitive waveshapes, the peak positive and negative output voltages can be determined graphically by locating the zero-voltage base line so that the enclosed waveform areas above and below the base line are equal.

The examples shown in Fig. 12.12 illustrate that in those cases where the signal has a high degree of time asymmetry, the output signal will have a peak excursion, as measured from ground, very nearly equal to the peak-to-peak amplitude of the

$$\text{OUTPUT POSITIVE AMPLITUDE} = en/(n+1)$$
$$\text{OUTPUT NEGATIVE AMPLITUDE} = e/(n+1)$$
$$\text{WHERE } n = t_2/t_1$$
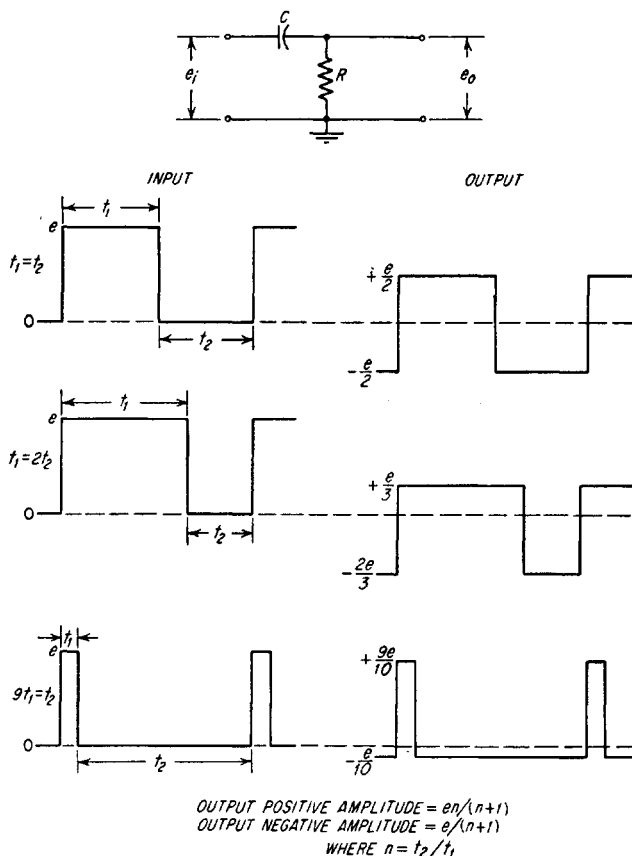
FIG. 12.12. Transmission of repetitive rectangular signals through an RC network.

input signal. The polarity of the output signal is dependent on the direction of the time asymmetry. Consequently, if the asymmetry is changed from one extreme to the other, the output-voltage values will extend over a range of almost twice the input signal peak-to-peak amplitude. In many circuits this shift in the operating region cannot be tolerated since an excessive dynamic range would be required in the following input circuit. A further undesirable feature which results from this shift is the possibility of an excessive change in gain in the following amplifier stages due to the shift in the operating region. Therefore, when the signal amplitude is great enough to occupy a major portion of the grid operating region or where the change in gain resulting from a shift in the operating point is to be avoided, a clamping circuit is desirable provided that all the signals are unipolar.

A simple clamp is shown in Fig. 12.13a where the capacitor charges through the low diode resistance and discharges through the much higher resistance R. Consequently, e_o reaches a positive voltage, with respect to ground, only great enough to replace the capacitor charge lost through R during the negative portion of the wave-form. If R is very large compared to the diode resistance, very little diode current will flow. Good clamp performance with very little waveform distortion is achieved if R is very much larger than the diode resistance and if the time constant RC is very long compared to the interval in which the waveform departs from the clamped base line. The diode in Fig. 12.13a could be reversed, thereby causing the most negative excursion of the signal to be clamped to ground potential.

Often there is a requirement for amplifying a varying amplitude unipolar signal without using a d-c coupled amplifier. This requirement can be satisfied by amplifying the signal in an RC coupled amplifier and restoring the unipolar nature of the signal in the output stage with a clamp. The base voltage for the signal is established by the clamping voltage E in Fig. 12.13b. The polarity of E can be either positive or negative.

FIG. 12.13. Simple diode clamps.

Care should be exercised in using clamps in applications where extraneous noise accompanies the signal. Noise peaks which are in the direction of diode conduction can cause the signal to be displaced out of the desired amplifier operating region.

Semiconductors such as germanium and silicon diodes are frequently used in clamp applications in lieu of thermionic diodes because of their small size and absence of Edison current effects (see Sec. 2.3e). However, in certain applications, germanium diodes may not be satisfactory due to their relatively low back resistance. Since the diode back resistance parallels the circuit resistance R during the portion of the signal waveform in which a thermionic diode would not be conducting, it is possible for the back resistance to limit the effective value of R in the circuit. With the application of an increasing inverse voltage across a crystal diode, the reverse current remains essentially constant, which indicates an increasing back resistance with an increasing inverse voltage. At some higher value of inverse voltage, however, the diode reverse current will increase significantly with a further increase in the inverse potential, which indicates a decrease in the diode back resistance. In the case of junction diodes, the decrease is regenerative within the diode, and the decrease in the back resistance is essentially a step function. This region is referred to as the *Zener region*. In point-contact diodes, significant heating ordinarily occurs within the diode before the Zener region is reached. This accounts for the more gradual decrease in back resistance, since the increase in temperature increases the energy level within the diode and, consequently, increases the conduction in the back direction. The amount of heating in point-contact diodes may actually destroy the diode before the Zener region is reached. In the various germanium crystal types, the maximum back resistance usually ranges from less than about 100,000 ohms to a few megohms and varies as an inverse exponential function of temperature. The back resistance of silicon diodes is much higher than that of germanium diodes; consequently, the problem of low back resistance ordinarily does not exist with silicon diodes. An additional limitation of crystal diodes is in regard to the time variant back-resistance phenomenon which exists when-

ever a forward voltage across a crystal diode is switched to an inverse voltage. At the instant of switching from a forward voltage to an inverse voltage, the diode has the characteristics of a bilateral element and has forward and back resistances which are very nearly equal; consequently, considerable reverse current flows at the instant the inverse potential is applied. Immediately following the application of the inverse potential, the back resistance of the diode increases exponentially as a function of time, and the large reverse current decreases with time to a very small value which is determined by the static high back resistance of the diode. In general, this characteristic presents no limitation in the use of point-contact diodes, since the back resistance recovery time (90 per cent recovery) is usually only a small fraction of a microsecond. A severe limitation exists in the use of junction diodes, however, since the back-resistance recovery time may be several microseconds. Most silicon diodes used in low-level electronic circuit applications are of the junction type, whereas most germanium diodes for similar applications are of the point-contact type; consequently, the problem of diode back-resistance recovery time is ordinarily associated with silicon diodes. For typical silicon junction diodes having low inverse voltage ratings, the back resistance recovers with an effective time constant T which is in the order of 0.5 to 1.0 μsec. (For higher inverse voltage ratings, the effective time constants are somewhat higher.) Ordinarily the time constant refers to the time required for the diode back resistance to reach a value of 50,000 ohms. If the diode external circuit impedance is small compared to 50,000 ohms, the circuit effective time constant will be significantly less than the diode back-resistance recovery time constant. In most applications, the back resistance can be considered to have attained a value approximately equal to the static value in an interval of approximately $4T$. This phenomenon must be considered when the signal contains frequency components which are in the order of $1/T$ to $10/T$ or higher. Silicon junction diodes have an additional limitation. The effective voltage at which forward conduction is initiated varies more with temperature in silicon junction diodes than in germanium point-contact diodes. This is of particular importance when the diode is used in low-level circuits.

Example 12.1

Determine the clamping effectiveness of the circuit shown in Fig. 12.13a under the following conditions:

$$e_i = 10 \text{ volts peak} \quad t_1 = 0.5 \ \mu\text{sec}$$
$$C = 1,000 \ \mu\mu\text{f} \quad t_2 = 50 \ \mu\text{sec}$$
$$R = 1 \text{ megohm} \quad \text{Source resistance} = 0 \text{ ohms}$$

Since the RC time constant is long (1,000 μsec) compared to t_2 (the interval for which the waveform departs from the clamped base line) and R is much larger than the diode resistance, e_2 can be assumed to be essentially equal to e_i.

The decay in e_2, designated Δe_2, during the interval t_2 can be determined as follows:

$$\Delta e_2 = e_2(1 - \epsilon^{-t_2/RC})$$
$$= 10(1 - \epsilon^{-0.05})$$
$$= 0.49 \text{ volt}$$

The decay e_2 corresponds to a loss of charge in capacitor C which is replaced by the diode current flow during the interval t_1.

$$\text{Average diode current} = \frac{\Delta e_2 C}{t_1}$$
$$= \frac{0.49 \times 10^{-9}}{0.5 \times 10^{-6}}$$
$$\simeq 1 \text{ ma}$$

From the diode E_b-I_b characteristic, the voltage corresponding to 1 ma can be determined. This voltage corresponds to the average value of e_1. The value of e_1 will increase if R or the clamped interval t_1 is reduced.

12.3b. One-way Clamp. A clamping circuit for establishing the initial potential in a network is frequently termed a one-way clamp. A common application of this type of clamp is in a sawtooth-waveform generator as shown in Fig. 12.14a.

The triode clamp in this application has only two operating states. In the quiescent condition, the grid is slightly positive and the plate current is limited primarily by the voltage drop across the high plate-load resistance R_b. With R_b high, the quiescent
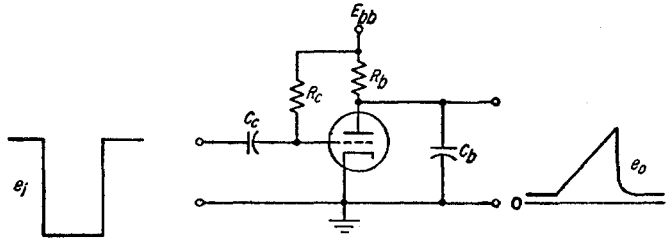
(*a*) ONE-WAY CLAMP USED IN A POSITIVE SAWTOOTH GENERATOR CIRCUIT
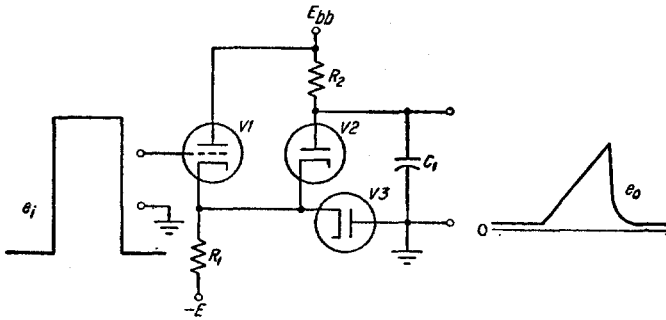
(*b*) DOUBLE DIODE ONE-WAY CLAMP
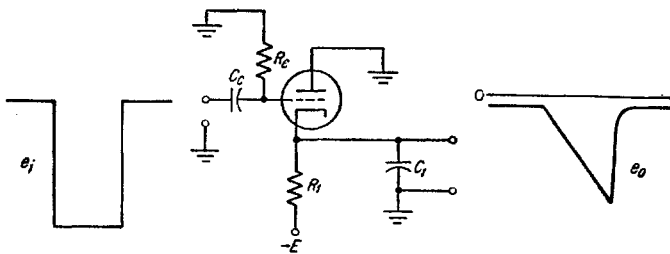
(*c*) TRIODE ONE-WAY CLAMP FOR A NEGATIVE SAWTOOTH WAVEFORM

FIG. 12.14. One-way clamp.

value of e_o will be a small percentage of E_{bb}. The other operating state of the triode is at plate current cutoff caused by a negative voltage impulse or gate e_i at the tube grid. The magnitude of the gate must be greater than the bias required for plate-current cutoff when operating with a plate voltage of E_{bb}. The time constant R_cC_c should be long compared to the time interval of the negative gate so that C_c will not discharge appreciably during the gate, thereby preventing the control grid from returning sufficiently positive to cause plate current flow. With the tube cut off, C_b starts to charge to the supply voltage E_{bb} through R_b. The voltage rise across the

capacitor will follow the typical RC exponential characteristic. If the time constant $R_b C_b$ is long compared to the gating interval, the peak value of e_o will be small compared to E_{bb} and the waveform of e_o will be a sawtooth of sufficient linearity for most applications.[1] The rate of return to the quiescent value of e_o at the end of the sawtooth is a measure of the quality of the clamp. An ideal clamp of this type would achieve a constant initial value of the sawtooth waveform independent of the repetition frequency and the gating interval.

Additional configurations of one-way clamps are shown in Fig. 12.14b and c. The one-way clamp in Fig. 12.14b uses two diodes and a cathode follower to form a positive sawtooth. Circuit design requires that the quiescent current through $V3$ be equal to or larger than that through $V2$. The magnitude of the gate appearing at the cathodes of $V2$ and $V3$ must be greater than the peak value of e_o. The cathode follower should have a low-output impedance and should be designed so that the quiescent cathode voltage is slightly negative.

A clamp for negative sawtooth generators is shown in Fig. 12.14c. The operation of this circuit is essentially identical to that of Fig. 12.14a except for the cathode output and the fact that the gate magnitude must exceed the sum of the peak value of e_o, the tube cutoff bias, and the voltage decay at the tube grid during the gate interval.

12.3c. Two-way Clamp. The clamp shown in Fig. 12.15 is capable of restoring either a positive or negative potential to a desired reference potential. When either tube is conducting, e_o is clamped to approximately the reference potential E_R. With the application of a negative unclamping gate, e_o will increase toward E_2 and will have an initial slope that is determined by the time constant RC_2. E_2 may be either positive or negative. The polarity and magnitude of E_1 determine whether the clamp is conductive or nonconductive during the ungated period. Therefore the required polarity of the enabling gate is dependent on whether or not the clamp is normally conductive or nonconductive.

FIG. 12.15. Triode two-way clamp.

If E_2 can be made to vary in magnitude through both polarities, this circuit provides a means of generating both positive and negative sawtooth waveforms. In a typical application, (1) E_1 is made equal to E_R, (2) the input gate is negative, and (3) E_2 is variable. Under these conditions, an output sawtooth waveform is generated during the interval of the negative gate. If E_2 is more positive than E_R, the polarity of the sawtooth waveform will be positive. If E_2 is more negative than E_R, the polarity of the sawtooth waveform will be negative. These waveforms have initial slopes proportional to the value of $E_2 - E_R$ at the initiation of the enabling gate.

The quality of clamping for this circuit is good only if the source resistance of the reference voltage E_R is low and if the value of R is very large.

12.3d. Keyed Clamps and Box-car Detectors. A circuit used to sample and store the instantaneous value of a varying voltage is usually termed a *sampler*, or a *box-car detector*. A simple circuit of this type can be derived from that in Fig. 12.15. If the voltage to be sampled is substituted for E_R and if E_2 and R are eliminated, a large positive pulse injected at the enabling gate input will cause the output voltage to assume a value proportional to the voltage being sampled. E_1 must be negative and greater than the sum of the most negative signal to be sampled plus the bias required

---

[1] More information on the design of sawtooth generators appears in Sec. 11.

to maintain cutoff. Obvious disadvantages of this circuit are the requirements that E_1 have such a large value and that the gating pulse be even larger.

The need for a high biasing voltage and a large gating signal are eliminated by the circuit in Fig. 12.16. The keying pulse to each grid is introduced by means of a pulse transformer winding which is returned to the cathode. The bias and the keying pulse amplitude are bootstrapped, or floated, on the cathode potential in such a manner that they are both independent of the magnitude of the cathode potential. The keying pulse will draw grid current which, by flowing through R_1 during the interpulse period, produces the bias required to keep the tubes in a normally nonconducting state. The keying pulse needs only to have a magnitude sufficient to produce a bias greater than that required to maintain cutoff with a plate voltage equal to the peak-to-peak value of the input voltage being sampled. The time constant R_1C_1 must be long compared to the interval between keying pulses.

An excellent sampling circuit is shown in Fig. 12.17. With the application of a
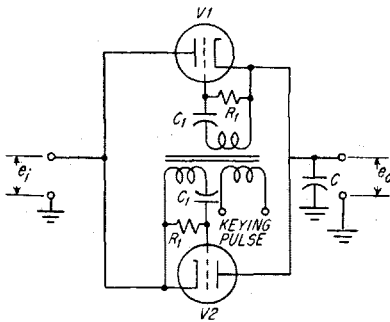
FIG. 12.16. Bootstrap keying of a keyed clamp.
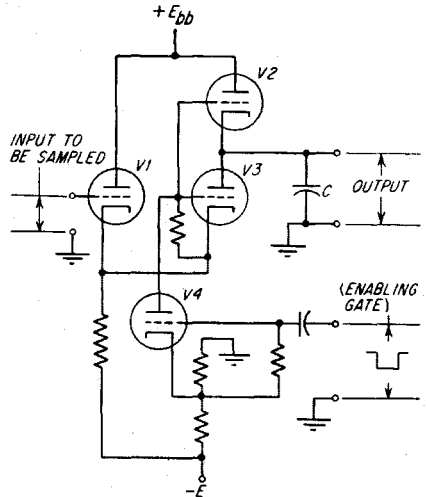
FIG. 12.17. Sampling circuit.

negative gate to the grid of $V4$, the two-clamp tubes, $V2$ and $V3$, conduct and the output voltage becomes equal to the input voltage except for a d-c offset and the distortion caused by the transmission of the signal through the cathode follower $V1$ and the clamp tube $V3$. At the end of the negative gate, $V2$ and $V3$ are cut off and the output voltage is stored across C until the next negative gate is applied. The cathode follower $V1$ and the enabling tube $V4$ have been shown for the purpose of illustrating typical methods for introducing the input signal and the enabling gate to the sampler.

When dealing with unipolar pulsed signals of short duration, it is often desirable to sample and store the last received signal until the next pulse is received. This can be accomplished in a box-car detector of the type shown in Fig. 12.18. Here, the input signal e_i consists of positive pulses which are recurrent at regular intervals but varying in amplitude. The positive pulses are reproduced by the cathode follower $V1$ and cause C_1 to charge to a value proportional to the instantaneous value of e_i. This charge is held until the occurrence of the following negative pulse e_t which restores the output voltage e_o to the voltage E_k. In any application where the occurrence of the e_i pulses can be anticipated, it is possible to generate the clearing pulses e_t with a slight time lead. Where the e_i pulses are irregular in time, it is possible to cause them to actuate a blocking oscillator (or other pulse generator) to form an e_t pulse. The e_i pulse must then be delayed in a delay line so as to be introduced into the box-car detector after clearing by the generated e_t pulse has occurred.

The voltage E_1 at the cathode of $V3$ established by R_1, R_2, and E must be more positive than the greatest value of e_o. The amplitude of the negative e_t pulses must exceed the dfference in the bias voltage E_1 and the quiescent cathode voltage E_k. C_2 should be much larger than C_1.

12.3e. Memory and Learning Time Constants. Box-car detectors are storage circuits which can be evaluated in terms of their persistence of memory and their rate of learning new data.

In most cases where these sampling and storage circuits are employed it is desirable to achieve perfect memory, i.e., have infinite time constants. Perfect memory is not feasible with practical circuit elements, but careful design will usually achieve time constants suitable for most applications. Every attempt must be made to minimize conductive paths that would discharge the storage capacitor. In general, output loading by shunt resistances across the storage capacitor is not used, the stored potential being applied to the grid of a vacuum tube for external utilization or indication. Where time constants in the order of 1 sec or longer are desired, grid conduction in this tube usually cannot be neglected. In addition, the capacitor used for the storage element should be chosen on the basis of very low leakage.
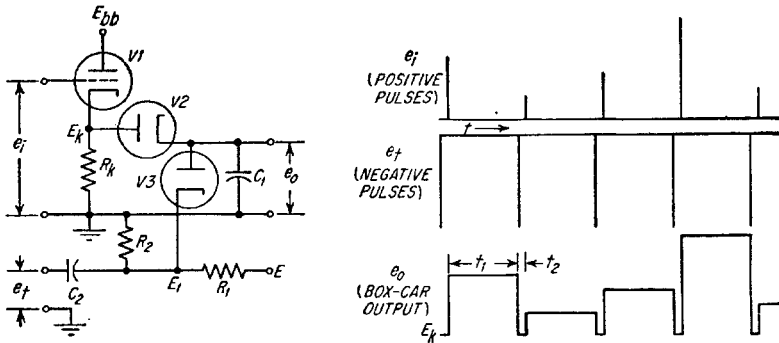
FIG. 12.18. Box-car detector which has complete restoration between samples.

An ideal sampling circuit is capable of instantaneously adopting the potential to be stored. The importance of the learning time constant for any given application is determined by the permissible sampling interval, which is usually dictated by the rate of change of the variable. In those cases where the variable is a short-duration pulse, the learning time must accordingly be very short. In some applications it is permissible, or even desirable, to integrate or delay the rate of assimilating abrupt changes in the variable, and consequently the learning time can extend over a number of sampling intervals. When the learning must be accomplished on each individual sample, the learning time constant should be much less than the sampling interval.

The learning time constant is determined by the product of the values of the storage capacitor and the total resistance of the series charging circuit. For example, in Fig. 12.18, the learning time constant is the sum of the cathode follower output impedance and the plate resistance of the diode $V2$ multiplied by the value of C_1. Both resistive terms will vary somewhat with signal amplitude. A reasonably accurate value for their sum, however, can be established by an analysis using the average signal value. If the internal impedance of the source generating e_i is as low or lower than the output impedance of the cathode follower, the cathode follower can be omitted. In Fig. 12.18, a problem similar to that of learning time exists for the discharge interval. The source impedance of the discharge pulses e_t should be low, and the pulses should be of sufficient amplitude and duration to drive the potential on C_1 below the cathode potential E_k.

Inductively Coupled Circuits

13.1. Fundamentals of Inductively Coupled Circuits. Included in this section are the basic definitions, equations, and equivalent circuits required for the detailed analysis of inductively coupled air-core and iron-core circuits.

13.1a. Coefficient of Coupling. In Fig. 13.1 a current in either the primary or secondary winding will produce magnetic flux. Depending on the orientation of the two windings, a certain portion of the total flux will link both windings. The dimensionless factor k, known as the *coefficient of coupling*, is equal to the ratio of the flux common to both windings to the total generated flux. If the coefficient of coupling k is equal to 0.5 or greater, the coils are usually said to be closely coupled, and if k is equal to 0.01 or smaller, the coils are said to be loosely coupled.
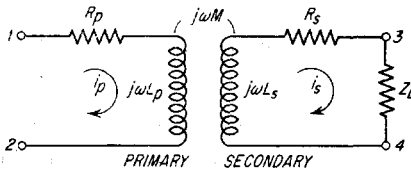
Fig. 13.1. Inductively coupled circuit.

13.1b. Mutual Inductance. Any two windings which have common flux are said to have a *mutual inductance M*. Mutual inductance in henrys can be determined from either Eq. (13.1), (13.2), or (13.3).

$$M = \frac{N_s \phi_{21} 10^{-8}}{i_p} \tag{13.1}$$

where N_s = number of secondary turns
ϕ_{21} = lines of flux common to both windings which are produced by primary current
i_p = primary current, amp

$$M = \frac{N_p \phi_{12} \times 10^{-8}}{i_s} \tag{13.2}$$

where N_p = number of primary turns
ϕ_{12} = lines of flux common to both windings which are produced by secondary current
i_s = secondary current, amp

$$M = k \sqrt{L_p L_s} \tag{13.3}$$

where L_p = primary inductance, henrys
L_s = secondary inductance, henrys

Mutual inductance can be determined experimentally for any two windings by measuring the inductances of the two coils when connected series-aiding (flux fields in the same direction) and series-opposing (flux fields opposing) and dividing the difference in the two measurements by 4.

13.1c. Leakage Inductance. Leakage inductance exists if all the flux produced by a current in one winding does not link the other winding. Primary and secondary leakage inductances in henrys can be determined from Eqs. (13.4) and (13.5).

$$L_{\text{pri. leak.}} = \frac{N_p \phi_1' \times 10^{-8}}{i_p} \tag{13.4}$$

$$L_{\text{sec.leak.}} = \frac{N_s \phi_2' \times 10^{-8}}{i_s} \tag{13.5}$$

where ϕ_1' = lines of flux produced by primary current which do not link secondary winding

ϕ_2' = lines of flux produced by secondary current which do not link primary winding

The sum of the primary and secondary leakage inductances referred to the primary and secondary terminals can be determined from Eqs. (13.6) and (13.7)

$$L_p' = 2(1 - k)L_p \tag{13.6}$$
$$L_s' = 2(1 - k)L_s \tag{13.7}$$

where L_p' = sum of primary and secondary leakage inductances referred to primary terminals

L_s' = sum of primary and secondary leakage inductances referred to secondary terminals

The measured inductance L_1 of the primary with the secondary short-circuited and L_2 of the secondary with the primary short-circuited are given by Eqs. (13.8) and (13.9).

$$L_1 = L_p(1 - k^2) \tag{13.8}$$
$$L_2 = L_s(1 - k^2) \tag{13.9}$$

Equations (13.8) and (13.9) are very nearly equal to Eqs. (13.6) and (13.7) if the value of k approaches unity. Therefore, if k is sufficiently large, the measurements L_1 and L_2 will be very nearly equal to the true values of L_p' and L_s'. The measurements are less than the true values of L_p' and L_s' by approximately $2\frac{1}{2}$ per cent if $k = 0.95$, 5 per cent if $k = 0.90$, and 11 per cent if $k = 0.80$. All short-circuit measurements should be made at frequencies low enough to minimize the effects of distributed capacitance. Where k is either not known or is not large enough to permit the determination of L_p' and L_s' by the short-circuit measurements, it is necessary to measure L_p, L_s, and M, so as to permit the determination of k from Eq. (13.3) and then to calculate L_p' and L_s' from Eqs. (13.6) and (13.7). L_p and L_s should be measured with the secondary and primary respectively open-circuited.

13.1d. Analysis of Air-core and Iron-core Coupled Circuits. Neglecting core losses, the equations for air-core and iron-core inductively coupled circuits are identical; therefore the two types of circuits can be analyzed in the same manner.

If the coupling coefficient is unity and if the copper losses, core losses, and winding capacitance are equal to zero, the voltage ratio will be equal to the turns ratio. These conditions represent the ideal iron-core power transformer. In general, all of these conditions are not satisfied, and the voltage ratio will not be equal to the turns ratio; hence it is usually necessary to make a complete circuit calculation.

For iron-core circuits which have coupling coefficients very nearly equal to unity, simplified equivalent circuits can be employed (see Sec. 14.2). The equivalent circuits discussed in this section, however, are for air-core transformers and iron-core transformers having coupling coefficients which are much less than unity.

Equivalent Circuits. To analyze an inductively coupled circuit, the secondary circuit can be referred to the primary circuit. This results in an equivalent primary circuit which permits the determination of the primary current. The secondary circuit can also be analyzed with the aid of an equivalent circuit. The solution of the secondary equivalent circuit is dependent on the solution of the primary equivalent circuit since the induced secondary voltage is equal to $-j\omega M i_p$.

Primary Equivalent Circuits.   Two equivalent primary circuits for air-core transformers are shown in Fig. 13.2.
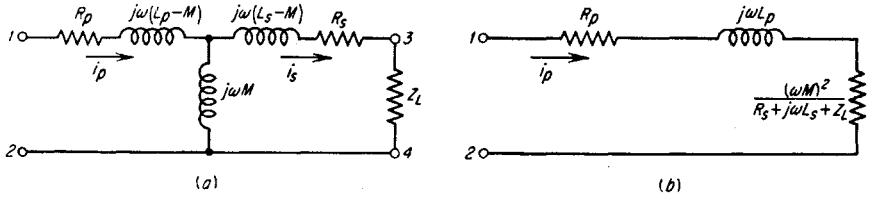
FIG. 13.2. Equivalent circuits for determining primary current.

Secondary Equivalent Circuit.   The equivalent secondary circuit for a circuit of the type shown in Fig. 13.1 is shown in Fig. 13.3.   The secondary current is that current which flows if the secondary induced voltage $-j\omega M i_p$ is placed in series with the secondary circuit.

13.1e. Primary and Secondary Circuits Each Tuned to the Same Frequency.   For
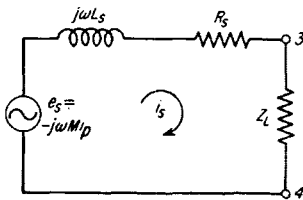
FIG. 13.3. Equivalent secondary for the circuit shown in Fig. 13.1.

inductively coupled circuits tuned to the same frequency, Fig. 13.4 illustrates how the primary current and the secondary voltage vary as a function of frequency and the coefficient of coupling.   The degrees of coupling most commonly referred to are:

1. Undercoupling
2. Critical coupling
3. Transitional coupling
4. Overcoupling

13.1f. Primary and Secondary Circuits Tuned to Different Frequencies.   If two inductively coupled circuits of equal Q are tuned to slightly different frequencies, the response curve is similar in shape but smaller in amplitude than that obtained by overcoupling the same two circuits when tuned to the same frequency.   The amplitudes of the two peaks in the secondary response curve are a function of the circuit
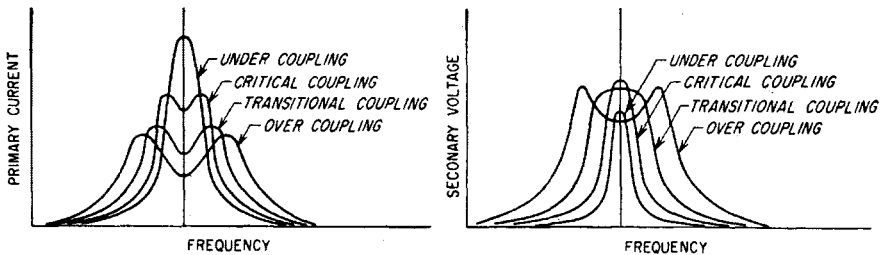
FIG. 13.4. Primary current and secondary voltage as a function of the coefficient of coupling when the primary and secondary are tuned to the same frequency.

Q's.   If Q_p does not equal Q_s the secondary response curve will be unsymmetrical. If the detuning between the primary and secondary becomes very large, the primary current will have only a single peak even though the secondary response curve has double peaks.

13.1g. T Sections as Equivalents for Inductively Coupled Circuits.   There are occasions when it is advantageous to replace inductively coupled circuits with equivalent T sections.   One particular example is the case where the inductances of the windings are to be adjusted to resonate with fixed primary and secondary capacities.   The

coupled circuit in Fig. 13.5a can be replaced mathematically and sometimes physically by the equivalent T section of Fig. 13.5b. Frequently it is found, in the calculation of the series arms of the T section, that either $L_p - M$ or $L_s - M$ is a negative inductance. This is the case when k is greater than either the quantity $\sqrt{L_p/L_s}$ or $\sqrt{L_s/L_p}$, respectively. Negative inductance presents no calculation difficulties; however, it does present a problem in a physical network. If the physical T is to replace the coupled circuit at only one frequency, the negative inductance can be represented by capacitance. If a T section is to physically replace a coupled circuit for a
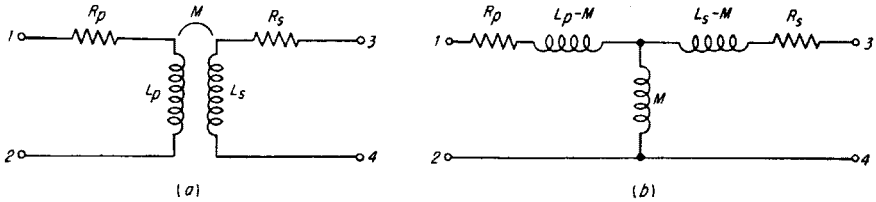
FIG. 13.5. Inductively coupled circuit and equivalent T section.

wide band of frequencies, the quantities $L_p - M$ and $L_s - M$ must have positive signs. This can be assured if the mutual inductance of the coupled circuit is made less than either the primary or secondary inductance.

In some instances there may be an advantage in converting the T network to a π network. If the T network has no negative inductances, the equivalent π network will also have no negative inductances.

It is important to note that, for an inductively coupled circuit terminated with a noncapacitive load, *it is not possible to realize a voltage gain if the mutual inductance is equal to or less than the primary inductance.* Therefore, where T sections are used to simulate inductively coupled circuits over a band of frequencies, there will be no voltage gain if the secondary load is not capacitive. For operation at *any specific frequency*, the unity-voltage gain limitation for a T section with a resistive load can be eliminated by making $L_p - M$ sufficiently negative and replacing the quantity with a capacitor.

13.1h. Gain-bandwidth Factor. A useful term in the evaluation of coupled circuits is *gain-bandwidth factor*. This is the ratio of the gain-bandwidth product of a given circuit to the gain-bandwidth product of a single-tuned circuit which has the same circuit capacitance as the sum of the primary and secondary capacitances of the coupled circuit. A single-tuned circuit is defined as the parallel combination of a single inductance, capacitance, and resistance.

Example 13.1

For the circuit shown in Fig. 13.6, determine the primary current and the voltage across terminals 3 and 4.
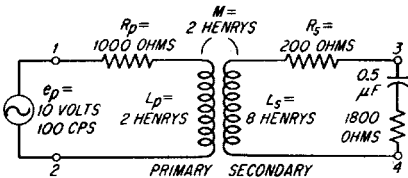
FIG. 13.6. Inductively coupled circuit.
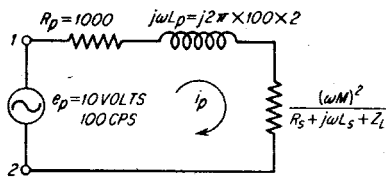
FIG. 13.7. Equivalent primary circuit for the circuit shown in Fig. 13.6.

Solution

1. Draw the equivalent primary circuit.

The equivalent primary circuit is shown in Fig. 13.7 and is of the same type as that shown in Fig. 13.2b.

2. Determine the primary current i_p.

$$i_p = \frac{\text{primary voltage}}{\text{primary input impedance}}$$

$$= \frac{e_p}{R_p + j\omega L_p + \dfrac{(\omega M)^2}{R_s + j\omega L_s + Z_L}}$$

$$= \frac{10}{1,000 + j2\pi \times 100 \times 2 + \dfrac{(2\pi \times 100 \times 2)^2}{200 + j2\pi \times 100 \times 8 + 1,800 - \dfrac{j}{2\pi \times 100 \times 0.5 \times 10^{-6}}}}$$

$$= \frac{10}{1,000 + j1,256 + 428 - j393} = \frac{10}{1,668 / +31.2°} = 0.006 / -31.2° \text{ amp}$$

or
$$= 0.0051 - j0.0031 \text{ amp}$$

3. Draw the equivalent secondary circuit as shown in Fig. 13.8.

4. Determine the secondary voltage e_L across terminals 3 and 4.

$e_s = -j\omega M i_p$
$= -j2\pi \times 100 \times 2 \times 0.006 \angle -31.2°$
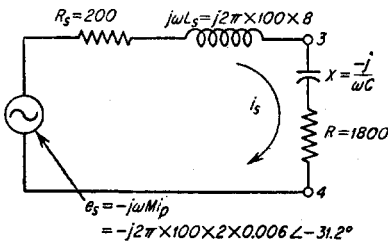
FIG. 13.8. Equivalent secondary circuit for the circuit shown in Fig. 13.6.

$$e_L = i_s Z_L$$

$$i_s = \frac{e_s}{Z_s}$$

$$\therefore e_L = \frac{e_s Z_L}{Z_s}$$

where i_s = secondary current
e_s = secondary induced voltage
Z_s = secondary circuit impedance
Z_L = impedance between terminals 3 and 4
e_L = voltage across terminals 3 and 4

$$e_L = \frac{-j\omega M i_p Z_L}{R_s + j\omega L_s + Z_L}$$

$$= \frac{(-j2\pi \times 100 \times 2)(0.0051 - j0.0031)\left(1,800 - \dfrac{j}{2\pi \times 100 \times 0.5 \times 10^{-6}}\right)}{200 + j2\pi \times 100 \times 8 + 1,800 - \dfrac{j}{2\pi \times 100 \times 0.5 \times 10^{-6}}}$$

$$= \frac{7.54 / -121.2° \times 3,657 / -60.5°}{2,716 / +42.6°}$$

$$= 10.15 / +135.7° \text{ volts}$$

13.2. Undercoupling. Two circuits which are inductively coupled and tuned to the frequency of the applied signal are said to be *undercoupled* if their orientation is such that increased coupling will cause an increase in the secondary voltage. The curves in Fig. 13.4b illustrate how the bandwidth characteristics vary as the degree of coupling is changed. The gain of an undercoupled circuit can be expressed as some value relative to the maximum possible value of gain which occurs at critical coupling (see Sec. 13.3).

The curve in Fig. 13.9 is a plot of the secondary voltage in an undercoupled circuit relative to the voltage at critical coupling. This is given as a function of the ratio between the actual coupling coefficient k to the critical coupling coefficient k_c. If the value of k is much less than critical, the shape of the secondary voltage response curve approximates the product of the response curves of two circuits having Q's which are equal to the primary and secondary circuit Q's, respectively. As the coupling coefficient is increased, the bandwidth of the secondary voltage response curve also increases.

13.3. Critical Coupling. Two circuits which are inductively coupled and tuned to the frequency of the applied signal are said to be *critically coupled* if they have been oriented so as to obtain the maximum possible secondary voltage.

The circuits in Figs. 13.10 and 13.11 illustrate the two most common forms of coupled circuit configurations. In Fig. 13.10 the circuit Q's are directly proportional to the size of the shunting resistors, provided that the Q's of the windings are much higher than the Q's of the circuits when loaded by the shunting resistors. This

$$M = k_c \sqrt{L_p L_s}$$

$$Q_p = 2\pi f_r C_p R_p$$

$$Q_s = 2\pi f_r C_s R_s$$

$$\frac{e_s}{e_p} = \sqrt{\frac{R_s}{R_p}} \quad (AT \ f_r)$$

$$f_r = \frac{1}{2\pi \sqrt{L_p C_p}} = \frac{1}{2\pi \sqrt{L_s C_s}}$$

$$Z_i = \frac{R_p}{2} \quad (AT \ FREQ \ OF \ MAX \ GAIN)$$

$\quad = INPUT \ IMPEDANCE \ AT \ TERMINALS \ a \ \& \ b \ AT \ f_r$

$$Z_{12} = \frac{\sqrt{R_p R_s}}{2}$$

$\quad = TRANSFER \ IMPEDANCE \ FROM \ TERMINALS \ a-b \ TO$
$\quad \ c-d \ AT \ FREQUENCY \ OF \ MAXIMUM \ GAIN$

$$A = \frac{e_s}{e_f} \simeq g_m Z_{12}$$

Fig. 13.9. A plot of the relative secondary output voltage as a function of the ratio between the actual coupling coefficient and the critical coupling coefficient for two circuits tuned to the same frequency and inductively coupled together.

Fig. 13.10. Shunt-loaded circuit and equations for critical coupling.

NOTE: When referring to Figs. 13.12 through 13.15, Q_p and Q_s can be identified as Q_1 and Q_2 or Q_2 and Q_1, respectively.

circuit is usually used when low-Q circuits are desired. If high-Q circuits are employed, the shunting resistors are omitted, resulting in the circuit shown in Fig. 13.11. If such is the case, the circuit Q's are determined entirely by the primary and secondary windings.

The general circuit equations for the two circuits with critical coupling are quite complex; hence, the circuit data have been plotted in Figs. 13.12 to 13.15. There is a limitation in the application of these curves since they *are not valid* for the circuit shown in Fig. 13.11 if either circuit Q is less than approximately 10. The data plotted in Figs. 13.12 to 13.15 are based on a constant current source, which is approached with most pentode-type tubes. In all the figures, Q_p and Q_s can be identified as either Q_1 and Q_2 or Q_2 and Q_L respectively.

13.3a. Gain-bandwidth Factor, Critically Coupled Circuit. This is a function of the Q ratio and the relative values of the primary and secondary circuit capacitances. The gain-bandwidth factor is greatest when $Q_p = Q_s$ and when C_p is either much larger or much smaller than C_s. Figure 13.12 is a plot of the gain-bandwidth factor for the case where the primary circuit capacitance C_p is equal to the secondary circuit capacitance C_s. Under these conditions, the Q ratio must be less than 4.5 to realize any improvement in the gain-bandwidth product over a single-tuned stage provided that the primary and secondary circuit capacitances are equal. If the capacitances are unequal, the gain-bandwidth factor established from Fig. 13.12 must be multiplied by the factor K, as determined by Eq. (13.10) to obtain the actual gain-bandwidth factor.

$$M = k_c \sqrt{L_p L_s}$$

$$Q_p = \frac{1}{2\pi f_r C_p R_p} = \frac{2\pi f_r L_p}{R_p}$$

$$Q_s = \frac{1}{2\pi f_r C_s R_s} = \frac{2\pi f_r L_s}{R_s}$$

$$\frac{e_s}{e_p} = \frac{Q_s}{Q_p}\sqrt{\frac{R_s}{R_p}} \quad (AT\ f_r)$$

$$f_r = \frac{1}{2\pi\sqrt{L_p C_p}} = \frac{1}{2\pi\sqrt{L_s C_s}}$$

$$Z_i = \frac{Q_p \pi f_r L_p}{2} = \frac{Q_p^2 R_p}{2} \quad (IF\ Q_p \geq 3)$$

$$= INPUT\ IMPEDANCE\ AT\ TERMINALS\ a\ 8\ b\ AT\ f_r$$

$$Z_{12} = \frac{Q_p Q_s \sqrt{R_p R_s}}{2} = \frac{1}{4\pi f_r}\sqrt{\frac{Q_p\ Q_s}{C_p\ C_s}} \quad (if\ Q_p \geq 3)$$

$$= TRANSFER\ IMPEDANCE\ FROM\ TERMINALS\ a\text{-}b\ TO$$
$$c\text{-}d\ AT\ f_r$$

$$A = \frac{e_s}{e_i} \approx g_m Z_{12}$$

FIG. 13.11. Series-loaded circuit and equations for critical coupling.
NOTE: When referring to Figs. 13.12 through 13.15, Q_p and Q_s can be identified as Q_1 and Q_2 or Q_2 and Q_1, respectively. Figures 13.12 through 13.15 *are not valid* for this circuit if either Q_p or Q_s is less than approximately 10.

$$K = \frac{C_p + C_s}{2\sqrt{C_p C_s}} \qquad (13.10)$$

It is therefore possible, with Q ratios larger than 4.5, to achieve gain-bandwidth products that are greater than that
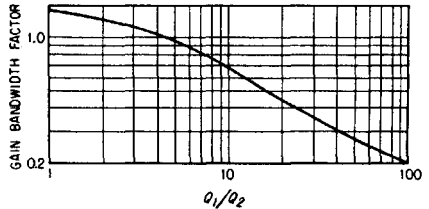
FIG. 13.12. Gain-bandwidth factor versus Q ratio for a critically coupled circuit in which $C_p = C_s$. If C_p does not equal C_s, the value as read from the curve must be corrected by the factor K given by Eq. (13.10).

achievable with a single-tuned circuit provided the improvement due to the term K is sufficiently large.

13.3b. Gain Ratio, Critically Coupled Circuit. The *gain ratio* G_r is the ratio of the voltage gain of a critically coupled stage to the voltage gain of a single-tuned stage having the same total circuit capacitance.

$$G_r = \frac{\sqrt{Q_p Q_s}}{Q_o}\frac{C_p + C_s}{2\sqrt{C_p C_s}} \qquad (13.11)$$

where Q_o = Q of single-tuned circuit
$C_p + C_s$ = total circuit capacitance of single-tuned circuit

The gain ratio is useful in comparing the voltage gain of a critically coupled stage to the voltage gain of a single-tuned stage without regard to relative bandwidths. If the critically coupled stage is adjusted to have the same bandwidth as the single-tuned stage, the gain ratio will be equal to the gain bandwidth factor. If maximum voltage gain is the prime objective and bandwidth is of no importance, there is ordinarily no advantage in using a critically coupled double-tuned circuit over a single-tuned circuit unless C_p and C_s are unequal since the maximum achievable value of $\sqrt{Q_p Q_s}$ in Eq. (13.11) will usually be no greater than the maximum value of Q_o.

13.3c. Low-Q Critically Coupled Circuits. The curves shown in Fig. 13.13 are an aid in the design of low-Q critically coupled circuits. The ratio of the primary and secondary resonant frequency f_r to the frequency f_{ac} which is the arithmetic center of the passband is plotted versus Q_2.

13.3d. Fractional Bandwidth, Critically Coupled Circuit. Figure 13.14 is a plot of the fractional bandwidth, i.e., the ratio of bandwidth β to the arithmetic-center frequency f_{ac}, versus Q_2 for different Q ratios. Bandwidth is not affected by the ratio of the primary and secondary circuit capacitances.

FIG. 13.13. $\dfrac{f_r}{f_{ac}}$ versus Q_2 for different Q ratios in critically coupled circuits.

FIG. 13.14. Fractional bandwidth versus Q_2 for different Q ratios in critically coupled circuits.
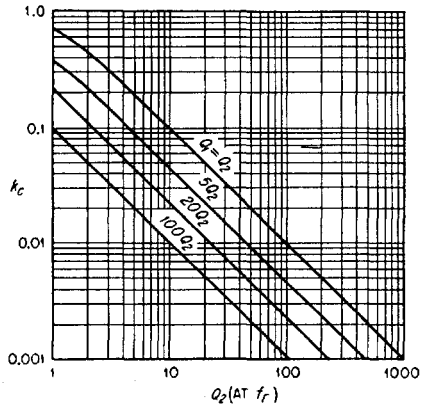
FIG. 13.15. Critical coupling coefficient versus Q_2 for different Q ratios in critically coupled circuits.

13.3e. Coupling Coefficient, Critically Coupled Circuit. Figure 13.15 is a plot of the critical coupling coefficient k_c versus Q_2 for different Q ratios. Equation (13.12) is an approximation which very nearly represents the data plotted in Fig. 13.15 provided both Q_1 and Q_2 are greater than approximately 5.

$$k_c = \frac{1}{\sqrt{Q_1 Q_2}} \qquad (13.12)$$

Example 13.2

Determine the operating characteristics of a critically coupled circuit of the type shown in Fig. 13.10 if

R_p = 2,000 ohms C_p = 2.5C_s
R_s = 1,000 ohms f_r = 1.5 Mc
Q_p = 5 $V1$ = 6AK5 tube; g_m = 5,000 μmhos, r_p = 0.3 megohm
Q_s = 1

1. Determine the Q ratio.

$$\frac{Q_p}{Q_s} = 5$$

2. Determine the gain-bandwidth factor.
From Fig. 13.12 determine the gain-bandwidth factor for $Q_1/Q_2 = 5$. The figure 0.95 must be corrected since $C_p = 2.5C_s$. Therefore the gain-bandwidth factor is equal to

$$0.95 \times \frac{2.5C_s + C_s}{2\sqrt{2.5C_s \times C_s}} = 1.05$$

3. Determine the arithmetic center f_{ac} of the secondary response curve.
From Fig. 13.13

$$\frac{f_r}{f_{ac}} = 0.95$$

$$f_{ac} = \frac{f_r}{0.95} = \frac{1.5}{0.95} = 1.58 \text{ Mc}$$

4. Determine the bandwidth β of the secondary response curve.
From Fig. 13.14

$$\frac{\beta}{f_{ac}} = 0.39$$

$$\beta = 0.39 \times 1.58 = 0.616 \text{ Mc, or 616 kc}$$

5. Determine the coupling coefficient k_c.
From Fig. 13.15

$$k_c = 0.38$$

6. Determine the gain from the grid of $V1$ to the grid of $V2$.
From Fig. 13.10

$$\frac{e_s}{e_i} \simeq g_m Z_{12} = 0.005 \times \sqrt{2,000 \times 1,000}/2$$

$$\simeq 3.54$$

Example 13.3

Design a critically coupled circuit of the type shown in Fig. 13.11 which has the secondary response arithmetically centered at 465 kc and a bandwidth of 13 kc. Assume the tube to be a 6AK5 operating at a g_m of 5,000 μmhos.
1. Determine the circuit Q's and the Q ratio.
Assume the Q ratio to be equal to 1. The fractional bandwidth is equal to 13/465, or 0.028. From Fig. 13.14 the desired value of the primary and secondary Q's is equal to 50.
2. Determine the value of the primary and secondary circuit capacitors.
To avoid any appreciable change in the total primary and secondary circuit capacitances which might be caused either by interchanging tubes or by changes in the input capacitance to $V2$, it usually is desirable to make the total circuit capacitances large compared to the expected variations. The nominal output and input capacitances of a 6AK5 tube are 2.8 and 4 $\mu\mu$f, respectively; hence it is relatively safe to assume that if both C_p and C_s were arbitrarily made equal to 50 $\mu\mu$f, the variations in tube capacitances would have little effect on the total. Therefore let

$$C_p = C_s = 50 \ \mu\mu\text{f}$$

The smaller C_p and C_s, the larger will be the gain as determined in step 3. Consequently, C_p and C_s should be made no larger than necessary to prevent excessive detuning due to variations in circuit capacitances.

3. Determine the gain from the grid or $V1$ to the grid of $V2$.
From Fig. 13.11

$$\frac{e_s}{e_i} \simeq g_m Z_{12}$$

$$\simeq \frac{g_m}{4\pi f_r} \sqrt{\frac{Q_p Q_s}{C_p C_s}}$$

$$\simeq \frac{0.005}{4\pi \times 465 \times 10^3} \sqrt{\frac{(50)^2}{(50 \times 10^{-12})^2}}$$

$$\simeq 856$$

4. Determine the critical coupling coefficient k_c.
From Fig. 13.15

$$k_c = 0.02$$

13.4. Transitional Coupling. Two inductively coupled circuits are said to be *transitionally coupled* if they are coupled to give the flattest secondary response curve possible. Transitional coupling will provide the widest passband without double peaks. Increased coupling will cause the midband portion of the selectivity curve to dip, and the circuit is then said to be overcoupled.

Transitional coupling is frequently referred to as optimum coupling. The gain-bandwidth product of a transitionally coupled circuit is always greater than that of a critically coupled circuit except in the case where the primary and secondary Q's are equal. In the case of equal-Q primary and secondary circuits, transitional coupling is identical to critical coupling. The improvement in the gain-bandwidth product over that of a critically coupled circuit is obtained at the expense of a loss in gain. The loss in gain is accompanied by an even larger increase in bandwidth, which accounts for the larger gain-bandwidth product. Unequal Q circuits which are transitionally coupled may develop pronounced asymmetry in gain with slight mistuning.

In Fig. 13.16, the circuit Q's are directly proportional to the size of the shunting resistors provided that the Q's of the windings are much higher than the Q's of the circuits when loaded by the shunting resistors. This circuit is usually used when low-Q circuits are desired.

$$M = k_t \sqrt{L_p L_s}$$

$$Q_p = 2\pi f_p C_p R_p$$

$$Q_s = 2\pi f_s C_s R_s$$

$$f_p = \frac{1}{2\pi \sqrt{L_p C_p}}$$

$$f_s = \frac{1}{2\pi \sqrt{L_s C_s}}$$

$$Z_{12} = \frac{U\sqrt{R_p R_s}}{2}$$

= *TRANSFER IMPEDANCE FROM TERMINALS a–b TO c–d*

U = *MULTIPLYING FACTOR TAKEN FROM FIG. 13. 22*

$$A = \frac{e_s}{e_i} \simeq \frac{g_m U \sqrt{R_p R_s}}{2}$$

FIG. 13.16. Shunt-loaded circuit and equations for transitional coupling.
NOTE: When referring to Figs. 13.18 through 13.22, Q_p and Q_s can be identified as Q_1 and Q_2 or Q_2 and Q_1, respectively.

In high-Q circuits, the shunting resistors are omitted, resulting in the circuit shown in Fig. 13.17. If such is the case, the circuit Q's are determined entirely by the Q's of the primary and secondary windings.

The general circuit equations for the two circuits with transitional coupling are quite complex; hence, the circuit data have been plotted in Figs. 13.18 to 13.22. There is a limitation in the application of these curves in that they *are not valid* for

the circuit shown in Fig. 13.17 if either circuit Q is less than approximately 10. The data plotted in Figs. 13.18 to 13.22 are based on a constant current source which is approached with most pentode-type tubes.

13.4a. Gain-bandwidth Factor, Transitionally Coupled Circuit. The gain-bandwidth factor is between 1.41 and 2 for the case of equal primary and secondary circuit capacitances. If the capaci-

$$M = k_f \sqrt{L_p L_s}$$
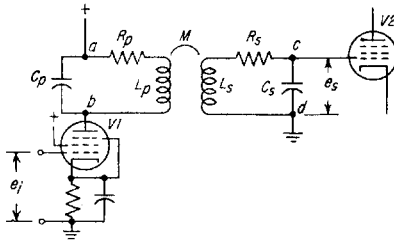
$$Q_p = \frac{1}{2\pi f_r C_p R_p} = \frac{2\pi f_r L_p}{R_p}$$

$$Q_s = \frac{1}{2\pi f_r C_s R_s} = \frac{2\pi f_r L_s}{R_s}$$

$$f_r = \frac{1}{2\pi\sqrt{L_p C_p}} = \frac{1}{2\pi\sqrt{L_s C_s}} \quad (FOR \ Q_p \ \& \ Q_s \geq 10)$$

$$Z_{12} = \frac{U Q_p Q_s \sqrt{R_p R_s}}{2} = \frac{U}{4\pi f_r} \sqrt{\frac{Q_p Q_s}{C_p C_s}}$$

= *TRANSFER IMPEDANCE FROM TERMINALS a-b TO c-d*

U = MULTIPLYING FACTOR TAKEN FROM FIG. 13.22

$$A = \frac{e_s}{e_i} \approx \frac{g_m U Q_p Q_s \sqrt{R_p R_s}}{2} = \frac{g_m U}{4\pi f_r} \sqrt{\frac{Q_p Q_s}{C_p C_s}}$$

FIG. 13.17. Series-loaded circuit and equations for transitional coupling.

NOTE: When referring to Figs. 13.18 through 13.22, Q_p and Q_s can be identified as Q_1 and Q_2 or Q_2 and Q_1, respectively. Figures 13.18 through 13.22 *are not valid* for this circuit if either Q_p or Q_s is less than approximately 10.

FIG. 13.18. Gain-bandwidth factor versus Q ratio for a transitionally coupled double-tuned circuit in which $Q_2 \geq 3$. If C_p does not equal C_s, the gain-bandwidth factor must be multiplied by $(C_p + C_s)/2 \ (\sqrt{C_p C_s}.)$

tances are unequal, the gain-bandwidth factor as determined from Fig. 13.18 should be multiplied by $(C_p + C_s)/(2 \sqrt{C_p C_s})$.

13.4b. Bandwidth of Cascaded Transitionally Coupled Stages. The bandwidth β_n of n identical transitionally coupled stages which are cascaded and synchronously tuned[1] is given by Eq. (13.13).

$$\beta_n = \beta \sqrt[4]{2^{1/n} - 1} \tag{13.13}$$

$$\beta_n \simeq \frac{\beta}{1.1 \sqrt[4]{n}} \tag{13.14}$$

where β_n = bandwidth between -3-db frequencies for n stages

n = number of stages

β = bandwidth between -3-db frequencies for one stage

13.4c. Gain Ratio, Transitionally Coupled Circuit. Without regard to relative bandwidths, the *gain ratio G_r* is the ratio of the voltage gain of a transitionally coupled stage to the voltage gain of a single-tuned stage having the same total circuit capacitance.

$$G_r = \frac{\sqrt{Q_p Q_s}}{Q_o} \frac{C_p + C_s}{2\sqrt{C_p C_s}} U \tag{13.15}$$

[1] A cascaded synchronously tuned amplifier is one in which successive stages are tuned to the same frequency.

where $Q_o = Q$ of single-tuned circuit

$C_p + C_s$ = total circuit capacitance of single-tuned circuit

$$U = \frac{\text{gain at transitional coupling}}{\text{gain at critical coupling}} \text{ (see Fig. 13.22)}$$

If the transitionally coupled stage is adjusted to have the same bandwidth as the single-tuned stage, the gain ratio will be equal to the gain bandwidth factor.
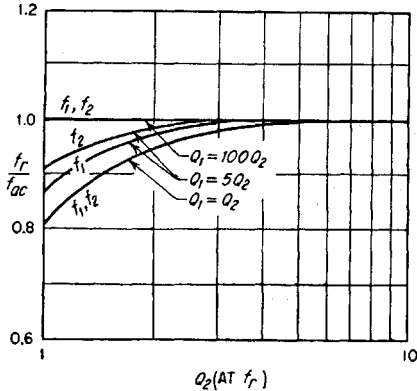
Fig. 13.19. Ratio of circuit resonant frequencies to the arithmetic center of the secondary response curve for different Q ratios versus Q_2 for transitionally coupled circuits.
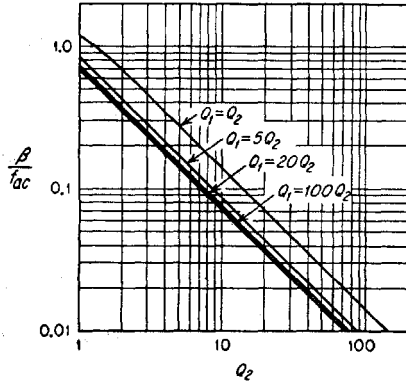
Fig. 13.20. Fractional bandwidth versus Q_2 for different Q ratios in transitionally coupled circuits.
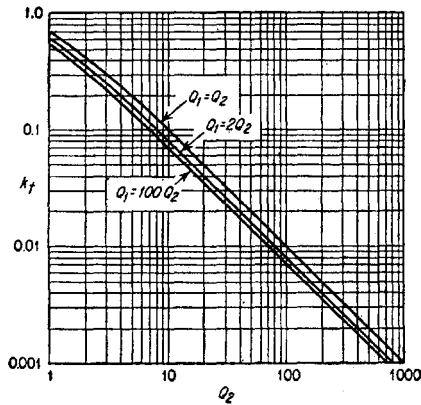
Fig. 13.21. Transitional-coupling coefficient versus Q_2 for different Q ratios in transitionally coupled circuits.
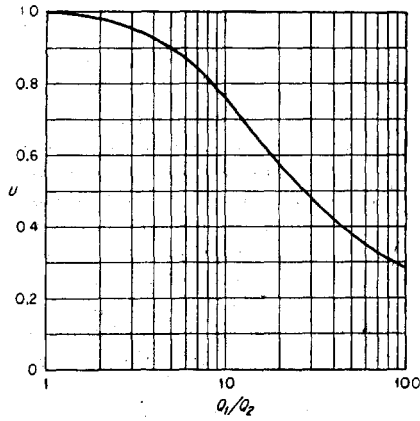
Fig. 13.22. Ratio of the gain in a transitionally coupled circuit to the gain in a critically coupled circuit versus circuit Q ratio.

13.4d. Low-Q Transitionally Coupled Circuits. In low-Q circuits, the primary and secondary resonant frequencies f_p and f_s may differ considerably from the arithmetic center f_{ac} of the passband. In most design problems f_{ac} is specified, and consequently f_p and f_s must be determined. Figure 13.19 is useful in determining f_p and f_s with respect to f_{ac}. The terms Q_2 and f_2 in the figure may refer to the Q and frequency of resonance of either the primary or secondary circuit depending on which has the lower Q.

13.4e. Fractional Bandwidth, Transitionally Coupled Circuit. Figure 13.20 shows the fractional bandwidth, i.e., the ratio of bandwidth β to the arithmetic center frequency f_{ac}, versus Q_2 for different Q ratios. Bandwidth is not affected by the ratio of the primary and secondary circuit capacitances.

13.4f. Coupling Coefficient, Transitionally Coupled Circuit. Figure 13.21 is a plot of k_t versus Q_2 for different Q ratios. Equation (13.16) very nearly represents the data plotted in Fig. 13.21 if both Q_1 and Q_2 are greater than approximately 5.

$$k_t = 0.707 \sqrt{\frac{1}{Q_p{}^2} + \frac{1}{Q_s{}^2}} \qquad (13.16)$$

Example 13.4

Determine the gain and bandwidth of a circuit of the type shown in Fig. 13.17 if

$$
\begin{aligned}
&Q_p = 30 && R_p = 66.6 \text{ ohms (a-c resistance of } L_p) \\
&Q_s = 150 && R_s = 13.3 \text{ ohms (a-c resistance of } L_s) \\
&C_p = C_s && V1 = \text{6AK5}, \ g_m = 5{,}000 \ \mu\text{mhos} \\
&f_r = 1 \text{ Mc}
\end{aligned}
$$

Solution

1. Determine the gain from the grid of $V1$ to the grid of $V2$.
From Fig. 13.17

$$\frac{e_s}{e_i} \simeq \frac{g_m U Q_p Q_s \sqrt{R_p R_s}}{2}$$

$$\simeq \frac{0.005 \times 0.895 \times 30 \times 150 \sqrt{66.6 \times 13.3}}{2}$$

$$\simeq 300$$

2. Determine the bandwidth.
From Fig. 13.20

$$\frac{\beta}{f_{ac}} = 0.029$$
$$\beta = 0.029 \times 10^6$$
$$\beta = 29 \times 10^3, \text{ or } 29 \text{ kc}$$

13.5. Overcoupled Circuits. Two circuits that are tuned to the same frequency and inductively coupled are said to be *overcoupled* if the coupling coefficient k is large enough to cause the secondary response curve to have two peaks. The difference in amplitude between the peaks is a function of the circuit Q's. With very high Q circuits the amplitudes of the peaks tend to be equal. For low-Q circuits, the low-frequency peak is larger and the high-frequency peak is smaller in amplitude than in the high-Q case. In both the low- and high-Q cases, the average of the two peaks is nearly equal to the amplitude of the response curve at transitional coupling.

13.5a. Midband Gain and Relative Gain of Peaks to Midband. For every ratio Q_p/Q_s or Q_s/Q_p there is a ratio of the actual coupling coefficient to the critical-coupling coefficient k/k_c above which increased coupling will result in double peaks in the secondary response curve. The value of k/k_c at which overcoupling begins is equal to the ratio of the transitional coupling coefficient to the critical coupling coefficient k_t/k_c. The intersection of the value of k/k_c which exists in a given circuit with the curve of Fig. 13.23 determines D where D is equal to the ratio of the midband gain (whether overcoupled or not) to the gain of the circuit with critical coupling. If the intersection of Q_1/Q_2 is at a higher point on the curve than the point of the intersection of k/k_c, the circuit is overcoupled and the ratio of the average amplitude of the peaks to the midband amplitude will be slightly less than the ratio of the two values of D read from Fig. 13.23 for Q_1/Q_2 and k/k_c. The applicable equations are

$$\text{Midband gain} = \text{gain of circuit at critical coupling} \times N \qquad (13.17)$$

$$\frac{\text{Average peak amplitude}}{\text{Midband amplitude}} \simeq \frac{RM}{N} \qquad (13.18)$$

where M = value of D at Q_1/Q_2 (see Fig. 13.23)

N = value of D at k/k_c (see Fig. 13.23)

R = a multiplying factor determined by Q ratio (see Fig. 13.24)
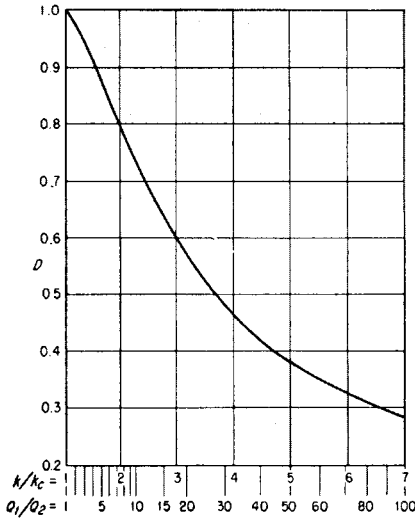
FIG. 13.23. Determination of circuit gain relative to the gain at critical coupling for ratios of k to k_c. The abscissa is calibrated in both k/k_c and Q_1/Q_2. The values of k/k_c associated with the values of Q_1/Q_2 are the maximum values at which double humps do not occur.
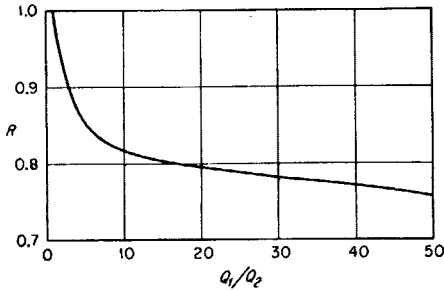
FIG. 13.24. Plot of the modifying factor R [see Eq. (13.18)] versus Q_1/Q_2.
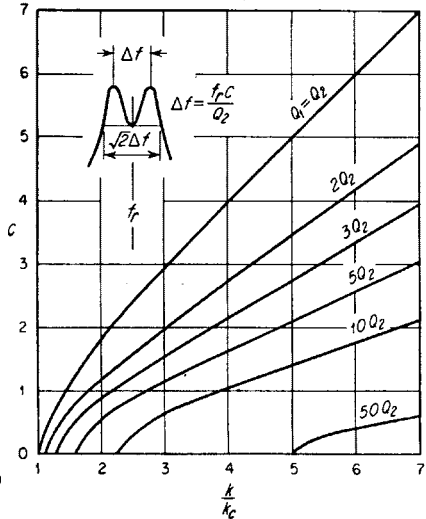
FIG. 13.25. Relationship between Δf, Q ratios, and k/k_c for overcoupled circuits which are tuned to the same frequency.

13.5b. Bandwidth of an Overcoupled Circuit. Because of the shape of the secondary response curve of an overcoupled circuit, it is not too helpful to refer to bandwidth at the −3-db points. Usually it is more desirable to define the secondary response curve as shown in Fig. 13.25.

13.5c. An Application of Overcoupled Circuits. A wide and fairly flat response curve is sometimes obtained from a two-stage amplifier in which the double-peak response of an overcoupled circuit is added to the single-peak response of a circuit that has transitional coupling or less. The overcoupled circuit is tuned so as to be symmetrical about the response curve of the other stage. To obtain this flat response, the circuit Q's, tuning and degree of coupling must be carefully chosen.

Example 13.5

In an overcoupled circuit where $k/k_c = 4$, $Q_2 = 10$, $Q_1 = 100$, and $f_r = 30$ Mc, determine Δf, ratio of the average peak amplitude to that at midband, and the gain at f_r based on the assumption that the gain at critical coupling is 30.

Solution

1. Determine Δf (Fig. 13.25)

$$\Delta f = \frac{f_r C}{Q_2} = \frac{30 \times 10^6 \times 1.05}{10} = 3.15 \text{ Mc}$$

where $C = 1.05$ (Fig. 13.25)

2. Determine the ratio of the average peak amplitude to that at midband.
From Eq. 13.18

$$\frac{\text{Average peak amplitude}}{\text{Midband amplitude}} \simeq \frac{RM}{N} = \frac{(0.82)(0.74)}{0.465} = 1.31$$

where $M = 0.74$ (from Fig. 13.23)
$\quad\quad N = 0.465$ (from Fig. 13.23)
$\quad\quad R = 0.82$ (from Fig. 13.24)

3. Determine the midband gain.
From Eq. 13.17

$$\begin{aligned} \text{Midband gain} &= \text{gain of circuit at critical coupling} \times N \\ &= 30 \times 0.465 \\ &= 13.95 \end{aligned}$$

Transformers and Chokes

14.1. Fundamentals of Magnetic Circuits. Figure 14.1 is a plot of flux density versus the magnetizing force required to alternately change the magnetization of a material between two values of flux densities B_1 and B_2. A curve of this type is known as a *hysteresis loop*. With reference to a symmetrical hysteresis loop, the negative magnetizing force required to reduce the flux density to zero is termed the *coercive force H_c*. The *residual flux density B_r* is that flux density remaining in the material when the positive magnetizing force goes to zero in a symmetrical hysteresis loop. The *retentivity* of the material is the flux density remaining in the material after it has been subjected to a magnetizing force sufficient to cause saturation flux density. A material which has a high retentivity is known to be *magnetically hard,* and a material which has a low retentivity is termed a *magnetically soft* material. Magnetically hard materials are used as permanent magnets,
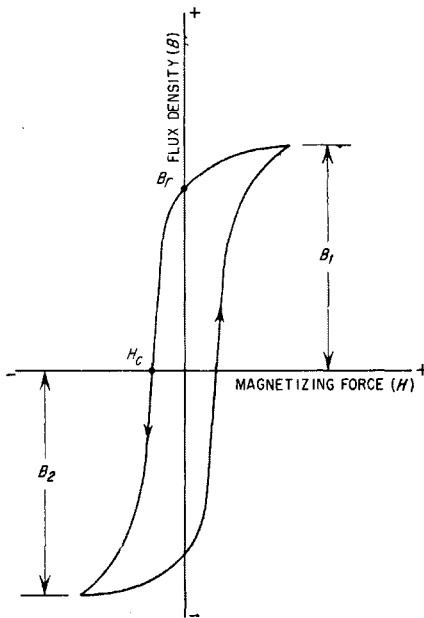
Fig. 14.1. Hysteresis loop.
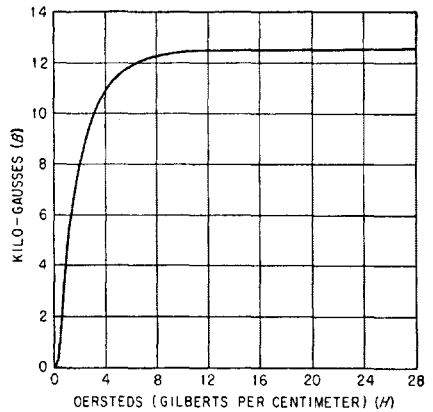
Fig. 14.2. Normal magnetization curve.

and magnetically soft materials are used as core materials for chokes and transformers. The normal *magnetization curve* for a material can be obtained if the material has no initial magnetization. The normal magnetization curve of Fig. 14.2 is a plot of flux density versus magnetizing force and is obtained by connecting the tips of a family of hysteresis loops resulting from a symmetrical hysteresis loop having been made larger and larger as shown in Fig. 14.3.

14.1a. Magnetomotive Force. In English units, the magnetomotive force F applied in a magnetic circuit is given in *ampere-turns;* in cgs units, the magnetomotive force is given in *gilberts*. A gilbert is equivalent to 0.796 ampere-turns.

14.1b. Magnetizing Force. The *magnetizing force H* is the effective magnetomotive force per unit length of core material. The unit of magnetizing force in cgs units is the *oersted* (1 gilbert/cm), which is equivalent to 2.02 ampere-turns/in.

Table 14.1 is a tabulation of the multiplication factors required to convert the English units of F, H, and B to the cgs units.

TABLE 14.1. MAGNETIC CONVERSION UNITS

English		*cgs*
F in ampere-turns	× 1.256	= F in gilberts
H in ampere-turns/in.	× 0.495	= H in oersteds
B in lines/sq in.	× 0.1550	= B in gauss

14.1c. Permeability. The ease with which a material can be magnetized is indicated by its permeability μ.
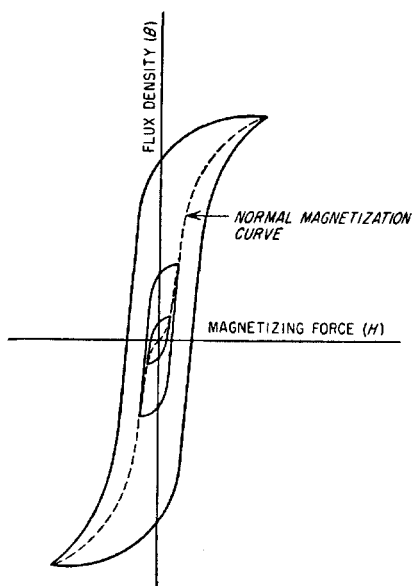
Fig. 14.3. Normal magnetization curve determined by the locus of the tips of a family of hysteresis loops.
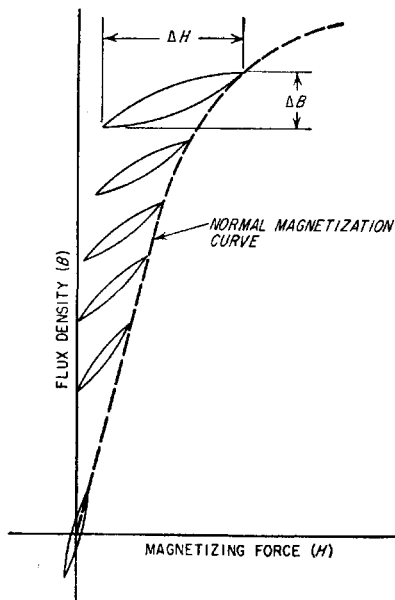
Fig. 14.4. Displaced hysteresis loops due to superimposed direct and alternating magnetizing forces.

The *static permeability* μ_s of a material at any point on the magnetization curve is defined by Eq. (14.1).

$$\mu_s = \frac{B}{H} \qquad (14.1)$$

The static permeability is therefore equal to the slope (gauss per oersted) of a line drawn from the origin through the point on the normal magnetization curve in Fig. 14.2.

The *incremental permeability* μ_{ac} of a material is the average permeability for an alternating magnetizing force. A d-c magnetizing force may or may not be present. Incremental permeability is very nearly equal to the slope (gauss per oersted) of a line connecting the tips of the hysteresis loop which may or may not have its center offset from the base line because of a d-c magnetizing force. The incremental permeability of any of the hysteresis loops shown in Fig. 14.4 can very nearly be determined from Eq. (14.2).

$$\mu_{ac} = \frac{\Delta B}{\Delta H} \tag{14.2}$$

where ΔB = incremental variation in flux density due to incremental variation ΔH in magnetizing force

14.1d. Flux Density. One line of flux is known as a *maxwell*. The unit of flux density B in cgs units is a *gauss*. One gauss is equal to one line or one maxwell per square centimeter, and in English units is equal to 6.45 lines per square inch.

In a d-c magnetic circuit, the d-c flux density can be determined from either Eq. (14.3) or (14.4).

$$B = H\mu_s \tag{14.3}$$

where B is in lines per square centimeter, H is in oersteds, and μ_s is in cgs units.

$$B = 3.192 H\mu_s \tag{14.4}$$

where B is in lines per square inch, H is in ampere-turns per inch, and μ_s is in cgs units.

A change of flux in the core of a magnetic circuit will result in an induced voltage in the winding surrounding the core. Equation (14.5) gives the peak value of the sinusoidal a-c flux density in terms of the resulting rms value of the induced voltage, number of turns, cross-sectional area of the core, and the frequency of the a-c flux.

$$B_{max} = \frac{e_{rms} \times 10^8}{4.44 f N A} \tag{14.5}$$

where B_{max} = maximum value of a-c flux density, lines/sq in.
  e_{rms} = rms value of induced voltage, volts
  f = frequency of flux, cps
  N = number of turns
  A = cross-sectional area of core, sq in.

14.1e. Inductance. A circuit has *self-inductance* if it generates a counter emf to any change in current through the circuit, i.e., a generated or induced voltage which is 180° out of phase with the voltage change causing the current variation. A coil has an inductance of 1 henry if a counter emf of 1 volt is developed when the current through the coil is changed at the rate of 1 amp/sec.

The relationship between inductance, lines of flux, and magnetizing current is given in Eq. (14.6).

$$L = \frac{0.707 N \phi_{max} \times 10^{-8}}{i_m} \tag{14.6}$$

where ϕ_{max} = maximum number of lines of flux during a cycle = $A B_{max}$
  L = inductance, henrys
  N = number of turns
  i_m = rms current in amperes producing flux

In practical applications, the value of iron-core inductances can be determined more readily from Eqs. (14.7) and (14.8).

$$L = \frac{3.192 N^2 A \mu_e \times 10^{-8}}{l_c} \tag{14.7}$$

$$\mu_e = \frac{1}{1/\mu_{ac} + l_g/l_c} \tag{14.8}$$

where N = number of turns
  A = cross-sectional area of core, sq in.
  μ_e = effective a-c permeability of iron core and air gap, cgs units

μ_{ac} = incremental permeability of iron core, cgs units

l_g = total length of air gap, in.

l_c = length of iron core, in.

L = inductance, henrys

Equations (14.7) and (14.8) are applicable to iron-core inductances in which the flux path may be through a small air gap l_g in addition to passing through the length of the iron core l_c. If the iron core is not interrupted by an air gap, $\mu_e = \mu_{ac}$.

Two inductances are said to have *mutual inductance* if they are oriented so that some portion of the flux developed by a current in one inductance will link the other inductance. Equation (14.9) expresses the relationship between mutual inductance, flux linkages, and magnetizing current. Equation (14.10) indicates that mutual inductance can also be defined as the ratio of the voltage induced in the second inductance to the rate of current change in the first inductance. For example, two inductances have a mutual inductance M of 1 henry if a current change at the rate of 1 amp/sec in the first inductance induces a voltage of 1 volt in the second inductance.

$$M = \frac{0.707 N \phi_{21} \times 10^{-8}}{i_m} \qquad (14.9)$$

$$= \frac{e_{sec}}{2\pi f i_m} \qquad (14.10)$$

where e_{sec} = induced secondary rms voltage

i_m = primary rms current in amperes producing flux

M = mutual inductance, henrys

N = secondary turns

ϕ_{21} = maximum number of primary lines of flux linking secondary during a cycle

14.1f. Core Losses. The heat losses which occur in the cores of magnetic circuits are due to (1) the dissipation of energy into heat in altering the orientation of the magnetic poles within the core material and (2) the I^2R losses within the core that occur as a result of the currents in the core which are induced by the varying flux. The first type of core loss is identified as *hysteresis loss*, and the second type is referred to as *eddy-current loss*.

Hysteresis loss in a transformer is the energy lost in alternately magnetizing the core of the transformer between two limits, such as B_1 and B_2 shown in Fig. 14.1, and is the same per cycle regardless of the frequency. Therefore, the hysteresis loss for any given maximum value of flux density is directly proportional to frequency. The addition of a d-c magnetizing force will displace the hysteresis loop and frequently change its shape as shown in Fig. 14.5. If the displacement of the hysteresis loop changes the enclosed area within the loop, the hysteresis loss will also be changed since the hysteresis loss per cycle is proportional to the area within the loop.

FIG. 14.5. Hysteresis loops with and without a superimposed d-c magnetizing force.

Eddy-current loss is the energy lost in heat because of I^2R losses caused by the flow of eddy currents within the core. These currents are induced within the core

by the varying magnetic flux and flow at right angles to the flux. The flow of eddy currents is minimized by making the core material in thin laminations and coating the laminations with either an oxide scale or some other nonconductive material. Eddy-current losses are porportional to the square of the frequency, the square of the a-c flux density, and the lamination thickness and inversely proportional to the resistivity of the core material. The addition of a d-c magnetizing force will not affect the eddy-current losses.

The sum of hysteresis and eddy-current losses can easily be measured. The usual procedure is to first excite the primary of the unloaded transformer at the rated input voltage and frequency. Under this condition, the losses within the transformer are the primary copper loss and the core loss, and their sum can be measured by a wattmeter. The primary I^2R losses can be computed if the primary current and resistance are measured. The difference between the losses read on the wattmeter and the calculated primary copper loss is the sum of the hysteresis and eddy-current losses.

Eddy-current and hysteresis losses can be separated if a voltage source is available which is variable in both amplitude and frequency. The procedure is to first determine the sum of the core losses at the transformer's normal operating frequency and flux density. The second step is to determine the core loss at several lower frequencies but at the same flux density as the initial measurement. The requirement that the same flux density be maintained requires that the amplitude of the input voltage be changed in direct proportion to the change in the frequency of the input voltage [see Eq. (14.5)]. Step three consists of plotting the core loss per cycle versus frequency and extending the curve to zero frequency. The core loss per cycle at zero frequency is the hysteresis loss per cycle for the specific value of flux density maintained during the measurements. The hysteresis loss under normal operating conditions can, therefore, be determined by multiplying the operating frequency by the hysteresis loss per cycle. The eddy-current losses can be determined by subtracting the calculated hysteresis loss from the total core loss obtained in the initial measurement.

14.1g. Exciting Currents in Iron-core Inductances. In a simple series a-c circuit containing an inductance and a generator having an output voltage e_g, the relationship between the voltage e_g at the terminals of the inductance and the induced voltage e_i is expressed by Eq. (14.11).

$$e_g = Ri_\phi + e_i \qquad (14.11)$$

where $e_i = \dfrac{N}{10^8} \times \dfrac{d\phi}{dt}$

$d\phi/dt$ = rate of change of flux, lines/sec
i_ϕ = exciting current, amp
R = resistance of winding, ohms,
e_g = generator voltage, volts
e_i = induced voltage, volts

In a well-designed iron-core choke or transformer, the product Ri_ϕ is usually a very small quantity compared to the terminal voltage e_g which means that the induced voltage e_i is very nearly equal to e_g. Therefore, if e_g is sinusoidal, e_i will be very nearly sinusoidal. It follows that the core flux ϕ must also be sinusoidal if the induced voltage is sinusoidal since e_i is equal to $N \times 10^{-8} \times d\phi/dt$. The exciting current i_ϕ, however, is not sinusoidal because of the shape of the hysteresis loop. The exciting current will have a shape similar to the waveform shown in Fig. 14.6. The exciting current is made up of a current component in phase with the induced voltage and a current component which lags the induced voltage by 90°. The first is known as the core loss current i_c, and the latter is the magnetizing current i_m.

14.1h. Effect of Direct Current through an Iron-core Inductance. Direct current through the winding of a choke or transformer displaces the hysteresis loop. In Fig. 14.5 it can be seen that the d-c flux density B_{dc} has displaced the hysteresis loop sufficiently to cause saturation of the core material at the maximum a-c flux density. It is apparent that with B_{dc} present, μ_{ac} is reduced to a much lower value. Therefore, one effect of direct current in the winding is to reduce the inductance of the winding since μ_{ac} has been reduced.

The quantity ΔH is a measure of the magnetomotive force per unit length of core material and is proportional to the exciting current in the winding. Consequently, an increase in the exciting current is required to obtain a given increase in the a-c flux density if a large value of B_{dc} exists as shown in Fig. 14.5. The exciting current i_ϕ is not sinusoidal even in the absence of B_{dc}. In a well-designed choke or transformer, however, the term Ri_ϕ is so small that the induced voltage e_i is essentially identical to e_g. If the exciting current should become very large for a portion of the hysteresis loop, i.e., when B_{dc} causes the core to become saturated at the maximum

Fig. 14.6. Typical exciting current waveform.

a-c flux density, the nonsinusoidal voltage drop Ri_ϕ in the winding becomes large enough that the induced voltage e_i also becomes appreciably distorted [see Eq. (14.11)].

The saturation effects of direct current on the core can be minimized by the insertion of an air gap in series with the iron core. This is explained in Sec. 14.4.

14.1i. Effect of an Air Gap on the Q of an Iron-core Inductor.[1] If the frequency of the voltage applied to an inductor remains constant and if the amplitude of the voltage is increased simultaneously with the insertion of an air gap in the core so as to maintain a constant a-c flux density in the core, the losses in the core of the inductor will remain constant. Insertion of an air gap in the core will cause the reluctance, i.e., the reciprocal of permeability, of the core to increase, and consequently the magnetizing current i_m must increase to maintain the same flux density. The core loss current i_c will remain constant since the a-c flux density and core losses remain constant. The approximate Q of the inductor is given by Eq. (14.12).

$$Q \simeq \frac{e_a i_m}{i_c^2 R + i_m^2 R + \text{core losses in watts}} \qquad (14.12)$$

where e_a = rms value of applied voltage, volts
    i_m = rms magnetizing current, amp
    i_c = rms core loss current, amp
    R = resistance of winding, ohms
The solution of Eq. (14.12) for maximum Q is given by Eq. (14.13).

$$i_m^2 R = i_c^2 R + \text{core losses} \qquad (14.13)$$

To obtain the highest Q's in iron-core inductors, it is usually necessary to insert an air gap in the core in order that i_m increase sufficiently to satisfy Eq. (14.13). Since the addition of an air gap reduces the inductance of a reactor, high Q's are usually obtained at the sacrifice of inductance.

---

[1] This discussion applies only in those instances where there is no direct current through the inductor.

14.1j. Magnetic and Electrostatic Shielding. A magnetic or electrostatic field can either be confined to or restricted from a given space by the use of a suitable shield.

A high-permeability shield is most effective for d-c and low-frequency magnetic fields. A high-permeability shield diverts the flux from its normal path to a path through the shield; hence, the shield should present a complete magnetic path for the flux. The inductance of a winding will be increased if enclosed by this type of shield since the permeability of the flux path will be increased. With an iron-core inductor, the increase in inductance is imperceptible, but the increase in inductance of an air-core coil can be appreciable.

At radio frequencies, shields are most effective when made of good electrical conductors such as aluminum and copper. Magnetic flux entering the shield induces eddy currents in the shield which in turn generate magnetic flux that directly opposes the flux entering the shield. Since the eddy currents can be quite large, the shield should have low-resistance joints, which means that they should be well lapped and soldered. This type of shield will reduce the inductance of a shielded inductor since the magnetic flux generated in the shield opposes the core flux and effectively reduces the permeability of the flux path. To obtain satisfactory shielding in critical low-level applications, it is sometimes necessary to employ alternate layers of high-permeability and high-conductivity shields. In many noncritical applications, a sufficiently thick, low-permeability iron casting will serve as a satisfactory magnetic shield.

Electrostatic shielding does not present as severe a problem as magnetic shielding since neither the thickness of the shield nor the degree of its conductivity is critical. An electrostatic shield can be fairly effective even though its surface is not continuous, as in the case of a wire mesh.

Frequently there is a requirement that two windings on a transformer be electrostatically shielded from each other. This can be accomplished without affecting the magnetic flux by placing one turn of metal foil which is grounded around the outside of the inner winding. Precautions must be taken to ensure that the ends of the foil are insulated from each other to prevent the foil from becoming a shorted turn. In addition, care should be taken to prevent more than one point on the foil from being grounded; e.g., the width of the foil should be equal to the width of the winding but not wide enough to touch the metal core.

14.2. Fundamentals of Iron-core Transformers. A transformer is normally used to deliver power at a voltage different from the voltage applied at its input terminals
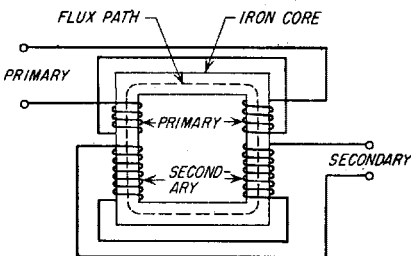
FIG. 14.7. Core-type transformer.
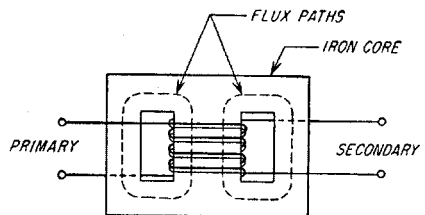
FIG. 14.8. Shell-type transformer.

and/or to provide isolation between two electrical circuits. Iron-core transformers are employed in both power and communication circuits.

14.2a. Types of Transformer Construction. The basic types of transformer construction are the core and shell types and are shown in Figs. 14.7 and 14.8, respectively.

In a core-type transformer, the primary and secondary windings are usually placed

(1) on the same leg of the core, (2) on opposite core legs, or (3) the primary and secondary windings are each equally divided and placed on opposite core legs (Fig. 14.7) to minimize the leakage inductance. The core is often made of long and short I's or of L laminations as shown in Fig. 14.9.
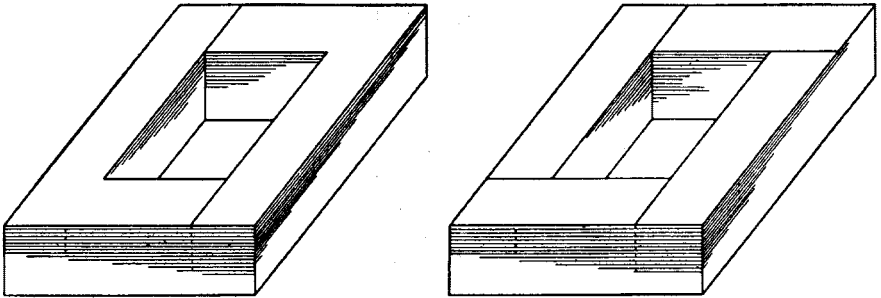
(a) CORE MADE OF L TYPE LAMINATIONS   (b) CORE MADE OF LONG AND SHORT I TYPE LAMINATIONS

FIG. 14.9. Transformer cores.

The majority of all small power and audio transformers are of the shell type. The core is usually made of E- and I-shaped laminations as shown in Fig. 14.10. These E's and I's can be stacked either interleaved or butt-jointed. Laminations ordinarily have a small burr on their edges because of the stamping process and are usually covered with a thin oxide scale.

These characteristics prevent their being stacked without a small loss in space. The term *stacking factor* is used to indicate the ratio of the effective core cross-sectional area A_e to the cross-sectional area A_c actually occupied by the laminations. The stacking factor is approximately the same for all small-gauge laminations, but varies, depending on the manner in which the core is assembled. The type of assembly denotes the arrangement of E's and I's within the core. Alternate reversal of E's and I's is denoted by 1:1, and 4:4 denotes the alternate reversal of groups of four. When there are no reversals of E's and I's, the core is said to be stacked with a butt joint. Sample stacking factors are as follows:
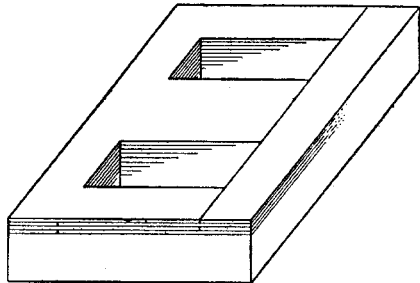
FIG. 14.10. Shell-type transformer core made of interleaved E- and I-type laminations.

Type of Assembly	Stacking Factor A_e/A_c
1:1	0.88
4:4	0.92
Butt joint	0.94

An important development in core materials is a Westinghouse product called Hipersil. Hipersil has approximately one-third higher permeability than the usual silicon steels at comparable flux densities and usually is made in thin sheets and wound on a rectangular mandrel. It is then annealed and cut into two sections which can be inserted into a completed transformer winding and banded together. Because of the high permeability of Hipersil as compared to ordinary silicon steels, Hipersil offers an advantage where space is at a premium. Hipersil cores can be used

in transformers of either the core or shell type as shown in Fig. 14.11. At present, several other manufacturers have entered this field and are manufacturing special high-grade cores in these same physical configurations.

In addition to laminated cores, solid ferrite cores are frequently employed. Since ferrite is a very poor conductor, the eddy-current losses are very small. Consequently, solid ferrite cores can be used profitably in high-frequency applications. Caution must be exercised in their use, however, since they lack the mechanical strength possessed by the more conventional cores.

14.2b. Fundamental Theory of Operation. The simplest iron-core transformer consists of a primary and a single secondary winding which share a common iron core as shown in Fig. 14.8. If a voltage is applied to the primary winding of the transformer, primary current will flow. If the secondary of the transformer is open, the current which flows in the primary is the exciting current i_ϕ. The exciting current is composed of a current in phase with the voltage induced in the primary winding and a current which is lagging the induced voltage by 90°. The first current is the core
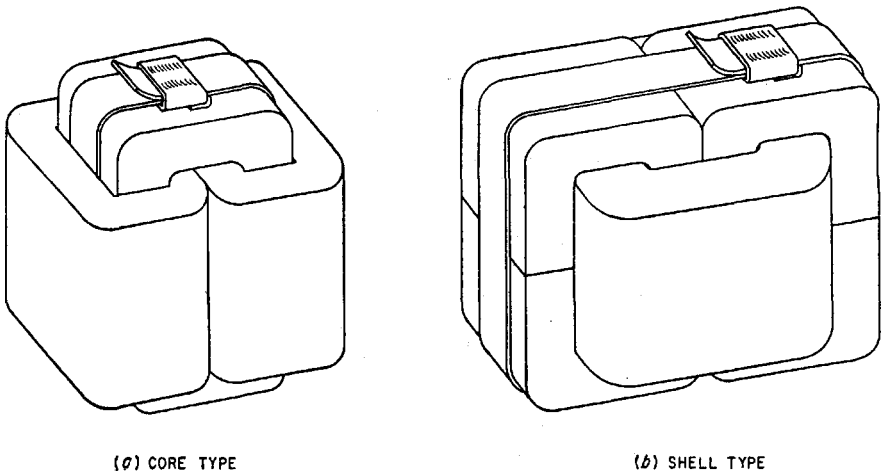
Fig. 14.11. Transformers with Hipersil C cores.

loss current i_c, and the second current is the magnetizing current i_m. The primary induced voltage is equal to the applied voltage less the vector $i_\phi R$, where R is the primary winding resistance. It is the magnetizing current which causes magnetic flux to exist in the core, and it is the magnetic flux which causes a voltage to be induced in the primary winding. The magnitude of the primary induced voltage is usually very nearly equal to the applied voltage and can be determined from Eq. (14.11). The secondary induced voltage is of the same phase as the voltage induced in the primary and is related in amplitude to the primary induced voltage by the ratio of the turns in the secondary winding to the turns in the primary winding. The application of a load on the transformer secondary causes secondary current to flow and is reflected to the primary as a decrease in the primary input impedance and, hence, causes an increase in the primary current. In an ideal transformer, loading the secondary winding will cause an increase in the primary ampere-turns which is equal and opposite to the secondary ampere-turns caused by the loading. The net result is that there is no change in the core flux because of load currents in the windings.

An extremely high exciting current often flows in the transient interval following the initial turn-on of power to a transformer. In steady-state operation, both the induced voltage and flux sinusoids are symmetrical about their zero axes. At the

initial "power turn-on," however, the flux sinusoid may be completely displaced above or below its steady-state axis. In these extreme cases, the result is a peak flux density during the first cycle of the applied voltage which is equal to twice the steady-state value which in turn causes complete core saturation and a peak exciting current which may exceed the maximum steady-state exciting current by a factor of a hundred or more (this transient phenomenon often presents a circuit protection problem when certain types of fuses and circuit breakers are employed). With reference to the sinusoidal flux existing in steady-state operation, it is only the rate of change of flux (flux waveshape) which establishes the induced voltage. Therefore, the flux sinusoid need not be symmetrically positioned on the zero flux axis to induce a sinusoidal voltage in the winding. Without a d-c current through the winding, the flux sinusoid is symmetrically located about the zero flux axis in steady-state operation; however, this ordinarily occurs after a transient period following "power turn-on." At the instant
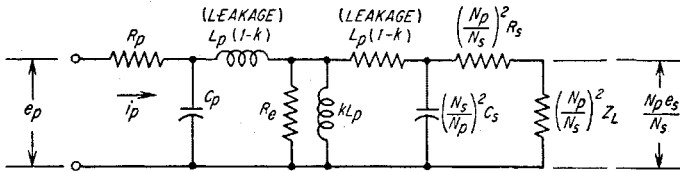
(a) EQUIVALENT CIRCUIT FOR DETERMINING THE PRIMARY CURRENT i_p. THE SECONDARY WINDING RESISTANCE AND LOAD HAVE BEEN MODIFIED BY $(N_p/N_s)^2$ SO THAT THE CIRCUIT MIGHT BE DRAWN AS A TRANSFORMER HAVING A TURNS RATIO OF ONE TO ONE
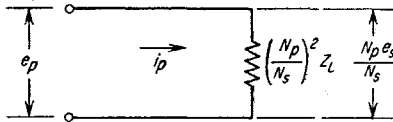
(b) EQUIVALENT CIRCUIT FOR DETERMINING THE PRIMARY CURRENT i_p IN AN IDEAL TRANSFORMER; I.E., A TRANSFORMER HAVING UNITY COUPLING, AN INFINITE INDUCTANCE PRIMARY, AND NO CORE LOSS, DISTRIBUTED CAPACITY OR WINDING RESISTANCE

FIG. 14.12. Equivalent transformer circuits; R_e represents the apparent resistance in the primary circuit due to eddy-current and hysteresis losses in the core.

of "power turn-on, the flux sinusoid is positioned in time (90° lagging the applied voltage) so that the instantaneous rate of change of flux causes the desired voltage to be induced (induced voltage is essentially equal to the applied voltage; see Sec. 14.1g). In addition, the flux sinusoid is positioned in the positive and negative directions so that the magnitude of the instantaneous initial flux is zero. The flux sinusoid will, therefore, not ordinarily be symmetrically positioned about the zero flux axis at "power turn-on." The only cases which result in the symmetrical positioning of the magnetic flux sinusoid (no transient condition) at "power turn-on" occur when power is turned on at the positive and negative crests of the applied voltage.

14.2c. Equivalent Circuits. To facilitate the analysis of a transformer, it frequently is desirable to make use of equivalent circuits. The circuits shown in Figs. 14.12 and 14.13 are equivalent primary and secondary circuits, respectively.[1]

Examination of Fig. 14.12a indicates the manner in which the magnitudes of certain circuit parameters affect transformer performance. For example, the magnitude of the shunting inductance kL_p establishes the lower-frequency limit of the transformer.

[1] Definitions of terms such as coupling coefficient k. mutual inductance M, leakage inductance, etc., are given in Sec. 13.7

The coupling coefficient k determines the magnitude of the leakage inductance, and the leakage inductance and the secondary-circuit capacitance determine the transformer upper-frequency limit.

The equivalent primary circuit of an ideal transformer is shown in Fig. 14.12b. In an ideal transformer, the core loss, resistance of windings, and distributed capacitance are all equal to zero. In addition, the coupling coefficient is equal to unity, i.e., the transformer has zero leakage inductance, and the value of the primary inductance is infinity.
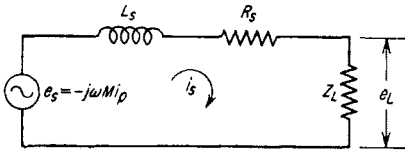
FIG. 14.13. Equivalent secondary circuit for a transformer.

14.2d. Calculation of Leakage Inductance. The magnitude of the leakage inductance[1] referred to a given winding is proportional to the square of the number of turns in the winding, mean length of winding turn, depth of windings, and separation of the winding from the other windings and is inversely proportional to the width of the winding layer. The value of the leakage inductance is almost independent of the core material and the degree of d-c saturation. Leakage inductance can be minimized by employing wide flat windings wound with a minimum of separation. Additional reduction of leakage

$$L_p' = \frac{3LN_p^2}{w}\left(d + \frac{h_1 + h_2}{3}\right)\times 10^{-5}\,MH$$
(a)

$$L_p' = \frac{3.4LN_p^2}{h}\left(d + \frac{w_1 + w_2}{3}\right)\times 10^{-5}\,MH$$
(b)

$$L_p' = \frac{0.85LN_p^2}{w}\left(d_1 + d_2 + \frac{h_1 + h_2 + h_3}{3}\right)\times 10^{-5}\,MH$$
(c)

$$L_p' = \frac{0.95LN_p^2}{h}\left(d_1 + d_2 + \frac{w_1 + w_2 + w_3}{3}\right)\times 10^{-5}\,MH$$
(d)

$$L_p' = \frac{0.35LN_p^2}{w}\left(d_1 + d_2 + d_3 + d_4 + \frac{h_1 + h_2 + h_3 + h_4 + h_5}{5}\right)\times 10^{-5}\,MH$$
(e)

$L_p' = $ SUM OF PRIMARY AND SECONDARY LEAKAGE INDUCTANCES REFERRED TO THE PRIMARY TERMINALS

$L = $ MEAN LENGTH OF TURN

$N_p = $ NUMBER OF PRIMARY TURNS

NOTE: ALL DIMENSIONS ARE IN INCHES
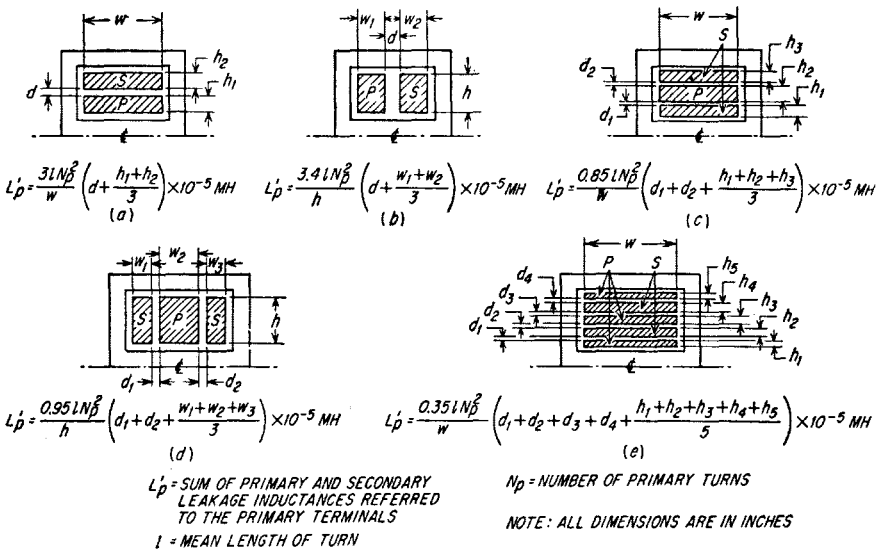
FIG. 14.14. Typical winding configurations and equations for determining leakage inductances. The value of L_s' can be determined by substituting N_s for N_p. Figures illustrate geometry of windings in transformer window.

inductance can be obtained by interleaving the primary and secondary windings. Figure 14.14 shows different winding configurations with the applicable equations for determining the total transformer leakage inductance as seen at the terminals of the primary winding. The values of the constants in the equations given in Fig. 14.14 are based on empirical data and in some instances do not agree with the calculated values for the idealized case.

[1] For measuring technique see Sec. 13.1c.

14.2e. Calculation of Distributed Capacitances. The distributed capacitance of the windings and the capacitance between the windings and the core and/or shield are important factors in the determination of the transformer upper-frequency response. Fundamental relationships for determining the distributed capacitance between the start and finish of any two adjacent layers of a transformer winding or between a single-layer winding and the core or shield can be determined by referring to the data given in Fig. 14.15.

The point at which a winding is grounded must be considered in the determination of the effective distributed capacitance across the winding. With reference to multilayer windings, Figs. 14.16 to 14.18 are applicable to (1) windings having one end grounded, (2) center-tapped windings, and (3) floating windings. The equations included in Figs. 14.16 to 14.18 are approximations. The larger the number of
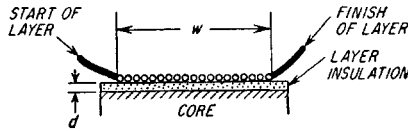
C_M is the measured capacitance between the core and winding with ends of the winding tied together.

$$C_M = \frac{0.225 l w K}{d}$$

C_E is the effective capacitance between the core and winding if one end of the winding is connected to the core.

$$C_E = \frac{C_M}{3}$$

NOTE: This equation is also valid for the effective capacitance between two layers of the same winding.

C_D is the distributed capacitance between the start and finish of a multilayer coil.

$$C_D = \frac{0.30 l w K (N_L - 1)}{d N_L{}^2}$$

where K = dielectric constant of insulation
l = mean length of turn in inches
w = width of layer in inches
d = thickness of wire insulation and paper insulation in inches
N_L = number of layers
C_M, C_E, and C_D are in $\mu\mu f$

FIG. 14.15. Fundamental equations for determining capacitance in a layer-wound coil.

layers in a given winding, the smaller will be the error. The dielectric constants of most organic insulating materials used in transformer construction are between 3 and 4.

14.2f. Physical Size of Transformers. For any specified load, the physical size of a transformer is dependent on such items as the allowable temperature rise, the type of cooling (forced air, etc.), type of core material, and the insulation requirements. The solid curve shown in Fig. 14.19 represents the average cased size versus the average of the primary and secondary volt-ampere ratings for several hundred transformers which were designed by several manufacturers for 60-cycle operation, a temperature rise of 40°C, and dependent only upon normal convection currents for cooling.

Certain types of loads, i.e., those causing current distortion in the transformer, require that the volt-ampere ratings of the primary and secondary windings differ from each other. An example is a single-phase transformer with a center-tapped secondary which operates into a full-wave rectifier with a choke-input filter (see

Table 15.1). If the power delivered to the load is equal to unity volt-amperes, the primary and secondary volt-ampere ratings should be 1.11 and 1.57, respectively, in
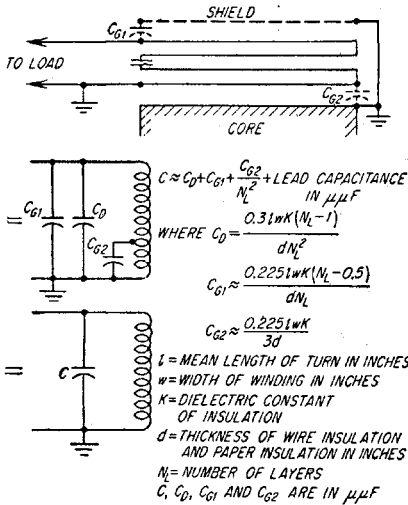
FIG. 14.16. Effective distributed capacitance across a winding having a load with one side connected to ground.
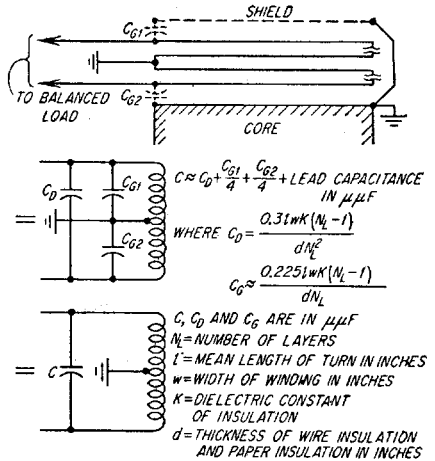
FIG. 14.17. Effective distributed capacitance across a winding which has its center tap grounded and a load balanced with respect to ground.

an ideal transformer. The average of these two values is 1.34, and it is this factor which affects the size of the transformer rather than the output power.

14.2g. Efficiency of a Transformer. A transformer will operate at its maximum efficiency when the load is adjusted so that the copper losses are equal to the core losses.

For a given maximum temperature rise, the power loss per unit of volume can be much higher in a small transformer than in a large transformer. This suggests the possibility of operating the cores of small power transformers at much higher flux densities than larger transformers, thereby permitting a reduction in the amount of core material required in small transformers. Usually this is not possible at typical power frequencies, since very high flux densities would saturate the core and cause the exciting current to be extremely large. An excessively large exciting current would cause the primary copper loss to be excessive and the secondary voltage to be appreciably distorted. Therefore, at frequencies of approximately 100 cps and

FIG. 14.18. Effective distributed capacitance across a winding which has neither load nor winding grounded.

lower, small power transformers are usually designed so that the maximum loss per pound of core material is limited by the flux density at which the core material rapidly

starts to saturate. Since normal practice is to design for a given temperature rise, a typical design for a low-frequency power transformer is one in which the core flux density is safely below saturation and the wire sizes are such that the sum of the core and the copper losses will cause the specified temperature rise. This usually
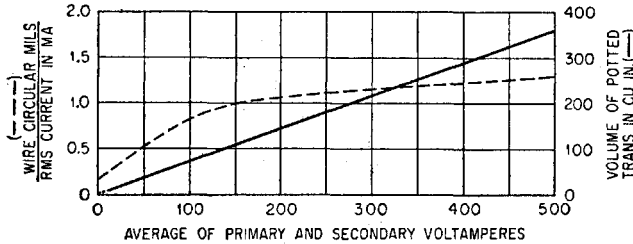
FIG. 14.19. Typical curves for 60-cycle transformers with 40°C rise.

results in the copper losses being much greater than the core losses in small power transformers having a rating of approximately 100 volt-amperes or less, operating at approximately 100 cps and lower frequencies, and having a temperature rise of 40°C.

A curve of efficiency versus output power for transformers operating into resistive loads is shown in Fig. 14.20 and is typical of small 60-cycle transformers which are designed for a temperature rise of 40°C. Higher operating efficiencies are possible if special high-grade core materials are used.

14.2h. Maximum Operating Temperatures. Insulating materials have been classified into five different groups by the American Institute of Electrical Engineers. A tabulation of the five classes with the maximum allowable temperatures is given in Table 14.2.

High operating temperatures cause organic insulating materials to dry out and become charred, which in turn causes the materials to be brittle and to lose their mechanical strength. Insulating materials weakened in this manner frequently break down if excessive stresses are developed in the windings by circuit overloads or by continued vibration.

FIG. 14.20. Curve is typical for commercial 60-cycle transformers operating into a resistive load, with a temperature rise of 40°C and dependent only on normal convection currents for cooling.

V. M. Montsinger[1] has shown that the rate of mechanical deterioration approximately doubles for each 8°C increase in temperature.

The class of insulation does not specify the allowable temperature rise of the transformer above the ambient temperature. A transformer with class A insulation (maximum of 105°C) operating in an ambient temperature of 31°C could have a temperature rise of 105°C–31°C, or 74°C. A transformer with class B insulation (maximum of 125°C) and operating in an ambient of 70°C could have but 125°C–70°C, or 55°C rise. Hence, both the temperature rise of the transformer and the maximum ambient temperature determine the class of insulation required.

The International Electrotechnical Commission refers to class 0 insulating materials as class Y. Also they have introduced classifications E and F for 120°C and 155°C, respectively.

[1] V. M. Montsinger, "Loading Transformers by Temperature," *Trans. AIEE*, vol. 49, pp. 776–790, April, 1930.

TABLE 14.2. CLASSIFICATION OF INSULATING MATERIALS

Class	Max. allowable temp.	Description of material
0	90°C	Class 0 insulation consists of cotton, silk, paper, and similar organic materials when neither impregnated nor immersed in a liquid dielectric
A	105°C	Class A insulation consists of: (1) Cotton, silk, paper, and similar organic materials when either impregnated or immersed in a liquid dielectric (2) Molded and laminated materials with cellulose filler, phenolic resins, and other resins of similar properties (3) Films and sheets of cellulose acetate and other cellulose derivatives of similar properties (4) Varnishes (enamel) as applied to conductors
B	130°C	Class B insulation consists of mica, asbestos, Fiberglas, and similar inorganic materials in built-up form with organic binding substances. A small proportion of class A materials may be used for structural purposes only
H	180°C	Class H insulating materials are the same as for class C with the exception that binding substances composed of silicone compounds may be used
C	>180°C	Class C insulation consists entirely of mica, porcelain, glass, quartz, and similar inorganic materials

In contrast with specifying the maximum allowable temperatures for insulating materials, the military classifies transformers in regard to performance, viz., maximum temperature and life expectancy. These classifications are as follows:

Class	Maximum Temperature (Ambient + Transformer Rise)
R	105°C
S	130°C
T	170°C
U	>170°C

	Life Expectancy
X	$\geq 10,000$ hours
Y	$\geq 2,500$ hours
Z	< 500 hours

As an example, a transformer with S and Y classifications would have a life expectancy of 2,500 hours or more at a transformer temperature of 130°C. This method of classification permits the transformer manufacturer to disregard the maximum allowable temperature ratings of insulating materials as specified by Table 14.2, since it is only the over-all transformer performance he is obligated to satisfy. For the preceding example, the maximum allowable ambient temperature is equal to 130°C less the transformer temperature rise.

Temperature rise in a transformer can be measured by noting the change in winding resistance. A 4 per cent increase in resistance very nearly represents a 10°C increase in temperature.

14.2i. Core Materials. Transformer core materials are usually referred to as being magnetically soft because they have low values of retentivity. Ideally, the permeability of the core material should be high at the operating flux density to minimize the exciting current. Manufacturers of core materials publish curves which show the guaranteed maximum watts loss per pound of core material as a function of frequency and maximum flux density. A sample family of curves is shown in Fig. 14.21. In addition, core manufacturers supply curves showing the incremental permeability μ_{ac} as a function of the maximum a-c flux density B_{max} and the d-c magnetizing force H_{DC}. A sample family of incremental permeability curves is shown in Fig. 14.22.

There are many grades of core materials, and for each application there are usually several different materials which will be satisfactory. The selection of a particular core material for a specific application might be on the basis of minimizing the size of the transformer, improving ease of assembly, ease of mounting the completed transformer, cost, or the fact that the desired electrical performance can be obtained only with a particular core material. The

Fig. 14.21. Family of curves showing core loss characteristics of a medium-grade transformer core material.

cores in power transformers are usually made of annealed steel laminations which have about a 1 to 4 per cent silicon content. This type of core material has relatively high permeability at high flux densities and is a fairly good compromise between cost
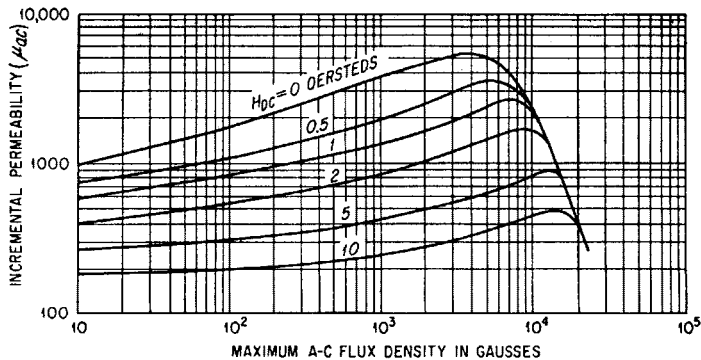
Fig. 14.22. Family of curves showing incremental permeability characteristics of a medium-grade transformer core material.

and core losses. In audio transformers, a core material having high permeability at low flux densities is often desired. Two sample core materials which have this property are mumetal and permalloy. Mumetal contains nickel, iron, copper and manganese, and permalloy contains nickel, iron, molybdenum and manganese.

TABLE 14.3. SOLID-COPPER WIRE TABLE AWG OR B&S GAGE

Gauge no.	Diameter, mils			Cross section		Ohms/1,000 ft at 20°C (68°F)	Pounds/ 1,000 ft	Feet/pound	Feet/ohm at 20°C (68°F)
	Bare	Single enamel	Double enamel	Circular mils	Square inches				
0000	460.0	......	.....	211,600.0	0.1662	0.04901	640.5	1.561	20,400.0
000	409.6	......	.....	167,800.0	0.1318	0.06180	507.9	1.968	16,180.0
00	364.8	......	.....	133,100.0	0.1045	0.07793	402.8	2.482	12,830.0
0	324.9	......	.....	105,500.0	0.08289	0.09827	319.5	3.130	10,180.0
1	289.3	......	.....	83,690.0	0.06573	0.1239	253.3	3.947	8,070.0
2	257.6	......	.....	66,370.0	0.05213	0.1563	200.9	4.977	6,400.6
3	229.4	......	.....	52,640.0	0.04134	0.1970	159.3	6.276	5,075.0
4	204.3	......	.....	41,740.0	0.03278	0.2485	126.4	7.914	4,025.0
5	181.9	......	.....	33,100.0	0.02600	0.3133	100.2	9.980	3,192.0
6	162.0	......	.....	26,250.0	0.02062	0.3951	79.46	12.58	2,531.0
7	144.3	......	.....	20,820.0	0.01635	0.4982	63.02	15.87	2,007.0
8	128.5	130.6	132.4	16,510.0	0.01297	0.6282	49.98	20.01	1,592.0
9	114.4	116.5	118.2	13,090.0	0.01028	0.7921	39.63	25.23	1,262.0
10	101.9	103.9	105.6	10,380.0	0.008155	0.9989	31.43	31.82	1,001.0
11	90.74	92.7	94.3	8,234.0	0.006467	1.260	24.92	40.12	794.0
12	80.81	82.7	84.2	6,530.0	0.005129	1.588	19.77	50.59	629.6
13	71.96	73.8	75.3	5,178.0	0.004067	2.003	15.68	63.80	499.3
14	64.08	65.9	67.3	4,107.0	0.003225	2.525	12.43	80.44	396.0
15	57.07	58.8	60.2	3,257.0	0.002558	3.184	9.858	101.4	314.0
16	50.82	52.4	53.9	2,583.0	0.002028	4.016	7.818	127.9	249.0
17	45.26	46.9	48.3	2,048.0	0.001609	5.064	6.200	161.3	197.5
18	40.30	41.8	43.2	1,624.0	0.001276	6.385	4.917	203.4	156.6
19	35.89	37.4	38.7	1,288.0	0.001012	8.051	3.899	256.5	124.2
20	31.96	33.4	34.6	1,022.0	0.0008023	10.15	3.092	323.4	98.50
21	28.46	29.9	31.0	810.1	0.0006363	12.80	2.452	407.8	78.11
22	25.35	26.6	27.7	642.4	0.0005046	16.14	1.945	514.2	61.95
23	22.57	23.8	24.9	509.5	0.0004002	20.36	1.542	648.4	49.13
24	20.10	21.3	22.4	404.0	0.0003173	25.67	1.223	817.7	38.96
25	17.90	19.0	20.1	320.4	0.0002517	32.37	0.9699	1,031.0	30.90
26	15.94	16.9	18.0	254.1	0.0001996	40.81	0.7692	1,300.0	24.50
27	14.20	15.2	16.0	201.5	0.0001583	51.47	0.6100	1,639.0	19.43
28	12.64	13.5	14.4	159.8	0.0001255	64.90	0.4837	2,067.0	15.41
29	11.26	12.2	12.9	126.7	0.00009953	81.83	0.3836	2,607.0	12.22
30	10.03	10.8	11.6	100.5	0.00007894	103.2	0.3042	3,287.0	9.691
31	8.928	9.7	10.3	79.70	0.00006260	130.1	0.2413	4,145.0	7.685
32	7.950	8.8	9.4	63.21	0.00004964	164.1	0.1913	5,227.0	6.095
33	7.080	7.8	8.3	50.13	0.00003937	206.9	0.1517	6,591.0	4.833
34	6.305	6.9	7.4	39.75	0.00003122	260.9	0.1203	8,310.0	3.833
35	5.615	6.1	6.7	31.52	0.00002476	329.0	0.09542	10,480.0	3.040
36	5.000	5.5	6.0	25.00	0.00001964	414.8	0.07568	13,210.0	2.411
37	4.453	4.9	5.4	19.83	0.00001557	523.1	0.06001	16,660.0	1.912
38	3.965	4.4	4.8	15.72	0.00001235	659.6	0.04759	21,010.0	1.516
39	3.531	3.8	4.2	12.47	0.000009793	831.8	0.03774	26,500.0	1.202
40	3.145	3.4	3.8	9.888	0.000007766	1049.0	0.02993	33,410.0	0.9534
41	2.800	3.1	3.4	7.842	0.000006159	1323.0	0.02374	42,130.0	0.7559
42	2.494	2.8	.....	6.219	0.000004884	1668.0	0.01882	53,120.0	0.5995
43	2.221	2.45	.....	4.932	0.000003873	2103.0	0.01493	66,990.0	0.4755
44	1.978	2.15	.....	3.911	0.000003072	2652.0	0.01184	84,470.0	0.3771

TABLE 14.4. WINDING DATA

B&S gage	Layer insulation	Minimum winding margins, in.	Turns per linear inch			
			Single enamel	SCE	SCC	DCC
8	0.012 in. kraft	¼	7	6	6	6
9	0.012 in. kraft	¼	8	7	7	7
10	0.010 in. kraft	¼	9	8	8	8
11	0.010 in. kraft	¼	10	9	9	9
12	0.010 in. kraft	¼	11	10	10	10
13	0.010 in. kraft	³⁄₁₆	12	11	12	11
14	0.010 in. kraft	³⁄₁₆	14	13	13	12
15	0.010 in. kraft	³⁄₁₆	16	14	14	13
16	0.010 in. kraft	³⁄₁₆	17	16	16	15
17	0.007 in. kraft	³⁄₁₆	19	17	18	17
18	0.007 in. kraft	⅛	21	19	20	18
19	0.007 in. kraft	⅛	24	21	22	20
20	0.005 in. kraft	⅛	27	23	24	22
21	0.005 in. kraft	⅛	30	26	27	24
22	0.005 in. kraft	⅛	33	29	30	27
23	0.005 in. kraft	⅛	37	32	34	29
24	0.0022 in. glassine	⅛	42	35	37	32
25	0.0022 in. glassine	⅛	46	38	41	34
26	0.0022 in. glassine	⅛	52	42	45	37
27	0.0022 in. glassine	⅛	58	46	49	40
28	0.0015 in. glassine	⅛	65	51	54	43
29	0.0015 in. glassine	⅛	72	55	58	46
30	0.0015 in. glassine	³⁄₃₂	81	60	63	49
31	0.0015 in. glassine	³⁄₃₂	90	64	68	52
32	0.0013 in. glassine	³⁄₃₂	99	69	73	55
33	0.0013 in. glassine	³⁄₃₂	111	75	79	58
34	0.0013 in. glassine	³⁄₃₂	125	81	85	62
35	0.001 in. glassine	³⁄₃₂	142	86	92	65
36	0.001 in. glassine	³⁄₃₂	155	92	98	68
37	0.001 in. glassine	³⁄₃₂	170			
38	0.001 in. glassine	¹⁄₁₆	193			
39	0.0007 in. glassine	¹⁄₁₆	215			
40	0.0007 in. glassine	¹⁄₁₆	239			
41	0.0007 in. glassine	¹⁄₁₆	265			
42	0.0005 in. ⎫	¹⁄₁₆	303			
43	0.0005 in. ⎬ condenser tissue	¹⁄₁₆	340			
44	0.0005 in. ⎭	¹⁄₁₆	369			

Transformer laminations are fabricated in many different thicknesses and grades. Probably the most common thickness used in small commercial transformers is 26 gage (0.019 in.). In special applications requiring a minimum of eddy-current losses, laminations as thin as 0.001 in. are used. Usually the very thin laminations are cemented together to form a loop of the type shown in Fig. 14.11. Otherwise the process of stacking the transformer would be relatively slow.

Solid ferrite cores (see Sec. 14.2a) are frequently used in the frequency range from a few kilocycles to several hundred kilocycles. Solid cores are formed by either pressing

or extruding mixtures of special metal oxides into the desired shapes and sintering. After the cores are sintered, they are extremely hard. Where necessary, the cores can be cut with a diamond saw and the surfaces can be smoothed by grinding. These cores have a high-volume resistivity; therefore, they have small eddy-current losses without being laminated.

14.2j. Winding Information. Tabulated characteristics of solid-copper wire are given in Table 14.3. Winding data given in Tables 14.4 and 14.5 show the recommended number of turns per inch for layer-wound coils, the layer insulation, minimum winding margins, etc. Normally, all wire smaller than No. 16 is wound on automatic winding machines, and No. 16 and larger sizes are wound by hand. In general, the windings made by automatic winders are applied before hand-wound windings.

TABLE 14.5. TYPICAL WORKING VOLTAGES FOR SEVERAL INSULATING MATERIALS

	RMS Working Voltage
Insulating Materials	*(Sine Wave)*
0.0007 in. glassine................	50
0.001 in. glassine.................	65
0.0015 in. glassine...............	100
0.0022 in. glassine...............	130
0.005 in. glassine................	250
0.002 in. kraft...................	30
0.005 in. kraft...................	60
0.007 in. kraft...................	80
0.010 in. kraft...................	100
0.002 in. varnish paper...........	370
0.005 in. varnish paper...........	750

NOTE: The recommended maximum rms working voltage gradient across the margins is 1,250 volts per ⅛ in.

In small power transformers, the number of wire circular mils per ampere will usually be between approximately 200 and 1,500. Typical values can be determined from the curve shown in Fig. 14.19. In very small transformers, the number of circular mils per ampere can be much smaller than in larger transformers. This is explained by the fact that for every unit of volume in a small transformer there is a
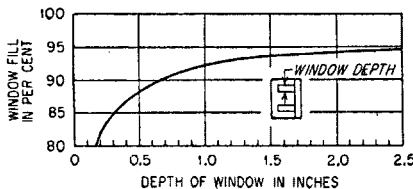
FIG. 14.23. Plot of suggested maximum window fill in per cent versus the depth of the window.

larger surface area for heat to be carried away by radiation, conduction, and convection than in a larger transformer; hence, the amount of heat dissipated per unit of volume for a given temperature rise is higher in a small transformer than in a large transformer.

In the physical design of a transformer, the thickness of the wire, insulation, and winding form must be carefully determined to ensure that the completed coil will properly fit into the lamination window. Figure 14.23 is a plot of the suggested maximum window fill in percentage versus the actual depth of the window.

Another consideration in the fabrication of the coil is the length, thickness, and internal dimensions of the form upon which the coil is wound and supported in the transformer. The length of the coil assembly is usually made between $\frac{1}{32}$ and $\frac{1}{16}$ in. less than the length of the window, and the inside dimension of the form is made approximately $\frac{1}{32}$ in. larger than the lamination tongue to facilitate stacking. If the core is solid, the internal dimensions of the form are made approximately $\frac{1}{32}$ in. larger in both directions. The thickness of the winding form varies from

about 0.025 in. for very small transformers to approximately 0.090 in. for transformers with a rating of 500 volt-amperes.

To facilitate winding, the depth of the lamination stack is usually limited to a maximum of approximately twice and a minimum of approximately one-half the width of the lamination tongue.

14.2k. Resistance of Windings. The resistance of any given transformer winding can be calculated from Eq. (14.14).

$$R = \frac{NR_w(B + C + \pi D)}{6,000} \tag{14.14}$$

where R = resistance of winding, ohms
  R_w = resistance of wire in ohms per 1,000 ft
  B = width of lamination tongue, in.
  C = depth of lamination stack, in.
  D = distance from core to center of winding, in.
  N = number of turns

14.3. Power Transformers. The factor which usually establishes the maximum power output of a transformer is its temperature rise. The exact temperature rise is somewhat difficult to calculate, but it is a function of the impregnating compound, the ratios of the wire circular mils to the rms currents through the windings, core material, type of cooling, etc. Small power transformers, i.e., approximately 0 to 500 volt-amperes, are usually designed with no special cooling provisions, and most manufacturers rely primarily on empirical data in their estimation of the expected temperature rise in any new design. The design data given in this section are for uncased transformers,[1] a medium-quality core material, and a temperature rise of
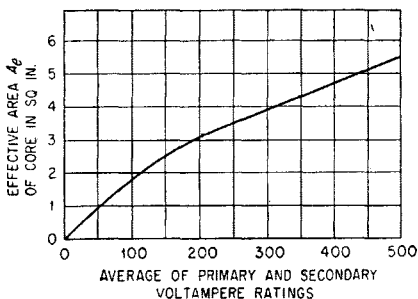
FIG. 14.24. Typical curve of recommended effective core areas versus volt-ampere ratings of 60-cycle power transformers.
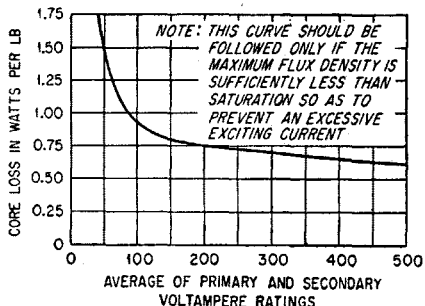
FIG. 14.25. Suggested core loss in watts per pound versus the volt-ampere rating of 60-cycle power transformers.

40°C. If 40°C is considered an excessive temperature rise, the designer should decrease the watts loss per pound of core material to some value less than that recommended in Fig. 14.25 and, in addition, should increase the ratio of wire circular mils to the rms current to a value above that recommended in Fig. 14.19.

Figures 14.24 and 14.25 indicate the recommended effective core cross-sectional area A_e and the watts loss per pound of core material for 60-cps transformers as a function of the volt-ampere rating. The total allowable loss versus the transformer surface area for a 40°C rise is given in Fig. 14.26. The design information contained in these figures plus the wire tables and the core data obtained from the core manufacturers are all that are necessary for the design of 60-cps power transformers. The actual design steps are detailed in Examples 14.1 and 14.2.

[1] Uncased transformers usually have about 10 per cent less temperature rise than transformers potted in tar.

In the early design steps of a transformer for operation at some frequency lower than 60 cps, it is recommended that the procedure detailed in the examples be followed with the exception that the recommended effective core area A_e, which is obtained from Fig. 14.24, be multiplied by the ratio of 60 to the actual operating frequency. The number of circular mils per ampere, the core loss in watts per pound, allowable
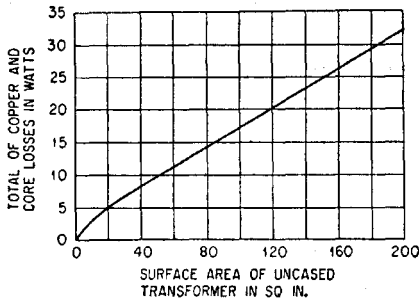
FIG. 14.26. Watts loss versus surface area for a 40°C rise in an uncased transformer.

transformer losses versus transformer surface area, etc., can be determined from Figs. 14.9, 14.25, and 14.26, as in the case of a 60-cps transformer. It may be necessary to increase the number of wire circular mils per ampere to obtain a satisfactory value of total watts loss versus the transformer surface area as plotted in Fig. 14.26. Once tentative values of core area, wire size, number of turns, etc., have been established, it may be necessary to modify slightly the values as determined from the figures to facilitate the final design. For example, it might be desirable to increase or decrease the core area, dependent on how many primary turns could be wound on the particular lamination. This type of compensation or adjustment is usually done to obtain the desired value of flux density.

Power-transformer design for frequencies higher than 60 cps can be accomplished by first designing the transformer for 60 cps and noting the core and copper losses. For operation at some higher frequency, the transformer would have essentially the same copper loss but less core loss. The core loss is reduced, since the maximum flux density is inversely proportional to frequency, assuming a constant input voltage. Therefore, it is possible at the higher frequency to reduce the size of the core until the core loss per pound is slightly larger than the initial loss per pound at 60 cps. If the core stack should become unusually thin in the process of reducing the core area, it will be necessary to use a smaller lamination size. In this process it will be necessary to increase the core area, however, since the smaller lamination will have a smaller window, and consequently the number of turns which can be put in the window will be less than in the case of the original lamination. The total losses and surface area of the completed design should be compared to the data given in Fig. 14.26 to determine if the sum of the copper and core losses needs to be either decreased or increased to obtain a 40°C rise. In high-frequency power transformers, very thin laminations or solid ferrite cores are often employed to minimize the eddy-current losses.

It is of interest to note that in most 60-cps and lower-frequency transformers, the maximum operating flux density is limited by the danger of core saturation and the associated excessive exciting current rather than by an excessive core loss. At high frequencies this is not the limitation since the core loss becomes excessive at flux densities far below saturation. For this reason, it is possible to make the core loss equal to the copper loss in small high-frequency power transformers.

If transformers designed from the included data operate at higher or lower temperatures than expected, it may be due to the use of different-grade core materials, unconventional geometries of fabrication, etc. The designer should then modify Figs. 14.24 to 14.26 to agree with the observed performance for the particular type of transformers being designed.

Example 14.1

Design a filament transformer which will operate from a 120-volt 60-cycle source and which will have an output rating of 6.3 volts at 6.5 amp and an allowable temperature rise of 40°C. Provide an electrostatic shield between the primary and secondary. Use the

29-gauge lamination shown in Fig. 14.27 and assume that its core-loss characteristics are those shown in Fig. 14.21. Assume that the magnetization curve for the core material indicates that the flux density should be kept below 80,000 lines/sq in. to avoid an excessive exciting current.

Transformers loaded with a resistive load will have a power factor less than unity because of the effect of the primary inductance across the source. In this example, assume the power factor to be equal to 0.9.

Solution

1. Determine the primary and secondary rms currents and the average of the primary and secondary volt-ampere ratings.

$$\text{Sec. } VA \simeq \text{output power } = 6.3 \times 6.5 = 41$$

NOTE: The secondary volt-ampere rating is actually equal to the product of the induced secondary voltage at full load and the secondary load current.

From Fig. 14.20, the efficiency can be expected to be about 85 per cent.

Therefore,

CORE WEIGHT IN POUNDS
$=1.55 \times (INCHES \ OF \ STACK) \times (STACKING \ FACTOR)$

FIG. 14.27. Allegheny Ludlum type EI-12 lamination.

$$\text{Pri. } VA \simeq \frac{\text{output power} \times 100}{\% \text{ efficiency} \times \text{power factor}} = \frac{41 \times 100}{85 \times 0.9} = 53.6$$

$$\text{Pri. rms current} = \frac{\text{pri. } VA}{\text{pri. voltage}} \simeq \frac{53.6}{120} = 0.447 \text{ amp}$$

Average of the primary and secondary volt-ampere ratings = $(41 + 53.6)/2 = 47.3$.

2. Determine the primary and secondary wire sizes.

From Fig. 14.19, the ratio of wire circular mils to the rms current in milliamperes should be approximately 0.50 for a 40°C rise.

Therefore,

$$\frac{\text{cir mils}}{\text{rms ma}} = 0.50$$

$$\text{Sec. cir mils} = 0.50 \times 6,500 = 3,250$$

therefore use No. 15 single enamel wire (from Table 14.3).

$$\text{Pri. cir mils} = 0.50 \times 447 = 224$$

therefore use No. 27 single enamel wire since 224 is closer to 202 than 254 (from Table 14.3).

3. Determine the tentative thickness of the lamination stack.

From Fig. 14.24 the effective area of the core material A_e should be approximately 0.90 sq in. Assuming a stacking factor of 0.88, the area occupied by the core A_c will be approximately 1.00 sq in. Therefore,

$$\text{Stack depth} = \frac{A_c}{\text{width of lamination tongue}} = \frac{1}{1} = 1 \text{ in.}$$

NOTE: If the stack depth is greater than approximately 2 in., or less than approximately 0.5 in., use a different-size lamination.

4. Determine the core flux density.

For this particular lamination, the exciting current starts to become very large at core flux densities of approximately 80,000 lines/sq in. (determined by examining the normal magnetizing curve and establishing the flux density at which very large increases in the magnetizing force are required to create small increases in the flux density). Therefore, to allow for a margin of safety, assume that the maximum flux density is to be 76,000 lines/sq in. The recommended operating core loss per pound, for a 40°C rise can be determined from Fig. 14.25. From Fig. 14.25 the permissible core loss would be approximately 1.5 watts/lb; however, the core loss of this particular lamination at a maximum flux density of 76,000 lines/sq in. and a frequency of 60 cps is only 0.8 watt/lb as determined from Fig. 14.21. Therefore, the operating core loss per pound is determined by the limitation
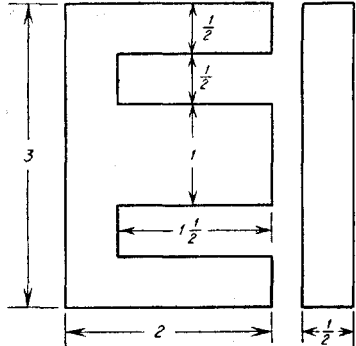
imposed by the danger of an excessive exciting current rather than by the danger of excessive heating due to a high core loss.

5. Determine the tentative number of primary turns required to operate the core at a maximum flux density of 76,000 lines/sq in.

From Eq. (14.5)

$$B_{max} = \frac{e_p \times 10^8}{4.44 f N_p A_e}$$

$$N_p = \frac{e_p \times 10^8}{4.44 f A_e B_{max}} = \frac{120 \times 10^8}{4.44 \times 60 \times 0.90 \times 76,000}$$

$$= 659$$

$$N_s \simeq N_p \frac{e_s}{e_p} = 659 \times \frac{6.3}{120}$$

$$\simeq 34.6$$

Therefore, the tentative number of secondary turns is either 34 or 35.

6. Establish whether or not the tentative number of primary and secondary turns will fit in the lamination window in accordance with the winding information given in Fig. 14.23 and Tables 14.3 and 14.4.

Primary Winding

Coil length....................................	$1\frac{7}{16}$
Margins..	$\frac{5}{32}$
Winding length................................	$1\frac{1}{8}$
Turns per layer................................	66
Number of layers..............................	10
Diameter of wire..............................	0.0152
Layer-insulation thickness.....................	0.0022

Build-up of primary wire.............................	0.152
Build-up of primary insulation	0.020
Total primary	0.172

Secondary Winding

Coil length....................................	$1\frac{7}{16}$
Margins..	$\frac{3}{16}$
Winding length................................	$1\frac{1}{16}$
Turns per layer................................	17
Number of layers*.............................	2

$$N_p \simeq N_s \frac{e_p}{e_s} = 34 \times \frac{120}{6.3}$$

$$\simeq 648 \text{ turns}$$

Diameter of wire..............................	0.059
Layer-insulation thickness.....................	0.010

Build-up of secondary wire............................	0.118
Build-up of secondary insulation.......................	0.010
Total secondary....................................	0.128

Total Fill in Window

Winding form....................................	0.075
Primary thickness (based on 648 turns)..........	0.172
Insulation over primary.........................	0.020
Electrostatic shield (0.002 copper sheet)..........	0.002
Insulation over shield...........................	0.020
Secondary thickness............................	0.128
Outside wrapper	0.020
Total thickness of coil	0.437
Width of window................................	0.500
Window fill (see Fig. 14.23)....................	87%

* Since 34 turns require only two layers and 35 turns require three layers, use only 34 turns and make the necessary adjustment in the tentative number of primary turns.

NOTE: If the build-up were too great, it would be necessary to increase the depth of the lamination stack so that the number of turns could be decreased to maintain the same flux density. If the window were insufficiently filled, the depth of the stack could be decreased with an associated increase in the number of turns.

7. Determine the resistance of primary and secondary windings based on the tentative number of turns established in step 6.

From Eq. (14.14)

$$R_p = 648 \times \frac{51.47}{6,000} [(1) + (1) + (\pi \times 0.161)]$$

$$= 13.9 \text{ ohms}$$

$$R_s = 34 \times \frac{3.184}{6,000} [(1) + (1) + (\pi \times 0.353)]$$

$$= 0.056$$

8. Determine the corrected number of primary and secondary turns based on the calculated values of resistance.

The maximum number of secondary turns allowable without going to three layers is 34; therefore let this be the number of secondary turns.

$$\text{Desired sec. induced voltage} = i_s R_s + \text{sec. terminal voltage}$$
$$= (6.5 \times 0.056) + 6.30$$
$$= 6.66 \text{ volts}$$

$$\left(\begin{array}{l}\text{Pri. induced volt. based on a pri. resist-}\\ \text{ance of 13.9 ohms}\end{array}\right) = \text{line volt.} - i_p R_p$$
$$= 120 - (0.447 \times 13.9)$$
$$= 113.8 \text{ volts}$$

$$\left(\begin{array}{l}\text{Corrected number of pri. turns based on}\\ \text{calculated values of induced voltages}\end{array}\right) = \text{sec. turns} \times \frac{\text{pri. induced volt.}}{\text{sec. induced volt.}}$$
$$= 34 \times \frac{113.8}{6.66}$$
$$= 580 \text{ turns}$$

NOTE: Since the preceding solution for primary turns is based on the resistance of 648 primary turns, it is necessary to redetermine the primary induced voltage and the corrected number of primary turns based on the value of resistance corresponding to 580 turns, i.e., $13.9 \times \frac{580}{648}$, or 12.4 ohms. Repeating the solutions to the two preceding equations establishes a value of primary turns equal to 583 and a resistance of 12.5 ohms.

9. Determine the corrected depth of the lamination stack based on the number of primary turns established in step 8 and a flux density of 76,000 lines/sq. in. Since the maximum allowable flux density has been established on the basis of limiting the exciting current, the required core area should be calculated on the basis of no secondary loading. Under this condition, the primary induced voltage (and exciting current) is the greatest and is very nearly equal to the applied voltage.

From Eq. (14.5)

$$A_e = \frac{e_p \times 10^8}{4.44 f N_p B_{\max}} = \frac{120 \times 10^8}{4.44 \times 60 \times 583 \times 76,000}$$
$$= 1.02 \text{ sq in.}$$

$$A_c = \frac{A_e}{\text{stacking factor}} = \frac{1.02}{0.88} = 1.16 \text{ sq in.}$$

$$\text{Stack depth} = \frac{A_c}{\text{width of tongue}} = \frac{1.16}{1} = 1.16 \text{ in.}$$

NOTE: The increase in stack depth will cause an increase in winding resistances. If the increase is excessive, further adjust the number of primary turns determined in step 8.

10. Determine the core loss (refer to Fig. 14.27 for core weight).

$$\text{Core loss} = \text{core weight} \times \text{watts loss per lb}$$
$$= (1.55 \times 1.16 \times 0.88) \times 0.8$$
$$= 1.26 \text{ watts}$$

11. Determine the copper loss.

$$\text{Copper loss} = i_p^2 R_p + i_s^2 R_s$$
$$= (0.447)^2 \times 12.5 + (6.5)^2 \times 0.056$$
$$= 4.87 \text{ watts}$$

12. Determine the efficiency of the transformer.

$$\text{Efficiency} = \frac{\text{output power}}{\text{output power} + \text{copper losses} + \text{core loss}} \times 100$$

$$= \frac{41}{41 + 4.87 + 1.26} \times 100$$

$$= 87\%$$

13. Determine the secondary no-load voltage.

$$\text{Sec. no-load voltage} = \frac{\text{pri. terminal voltage}}{\text{turns ratio}}$$

$$= \frac{120}{58\frac{3}{4}}$$

$$= 7.0 \text{ volts}$$

14. Estimate the temperature rise in the transformer.

Surface area of transformer = 30.5 sq in.
Watts loss = 6.13

From Fig. 14.26 the temperature rise can be estimated as being approximately equal to 36°C.

Example 14.2

Design a single-phase 117-volt 60-cycle power transformer for a full-wave rectifier and choke-input filter. The d-c output voltage and current from the filter are to be 300 volts and 200 ma, respectively. The total series resistance of each rectifier tube and the filter

$\frac{11}{16}$

$\frac{11}{16}$

$4\frac{1}{8}$ $1\frac{3}{8}$

$2\frac{1}{16}$

$2\frac{3}{4}$ $\frac{11}{16}$

CORE WEIGHT IN POUNDS
= 4.17×(INCHES OF STACK)×(STACKING FACTOR)
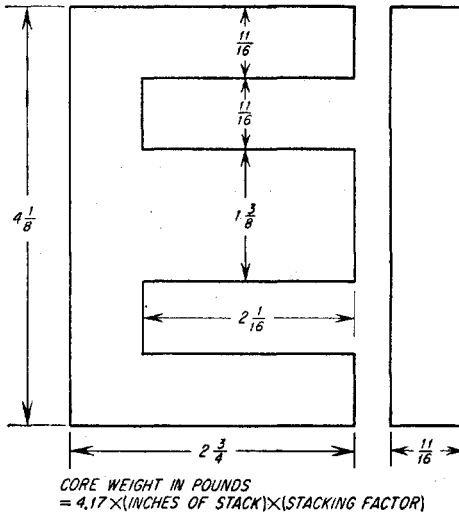
FIG. 14.28. Allegheny Ludlum type EI-138 lamination.

choke is specified as being 350 ohms. Design for a 40°C rise and use the lamination shown in Fig. 14.28. Assume the core-loss characteristics to be those of Fig. 14.21 and assume that the magnetization curve for the core material shows that the flux density should be kept below 80,000 lines/sq in. to avoid an excessive exciting current.

Solution

1. Determine the primary and secondary rms currents and the average of the primary and secondary volt-ampere ratings (refer to Fig. 14.29).

Sec. rms current[1] = 0.707 × d-c load current
= 0.707 × 200 ma = 141 ma

---

[1] See Table 15.1.

$\frac{1}{2}$ total sec. rms voltage (at terminals under load) $= 1.11 \times$ required d-c voltage

$\qquad\qquad = 1.11 \times$ d-c output voltage $+$ drop in choke and rectifier

$\qquad\qquad = 1.11 \times [300 + (0.200 \times 350)]$

$\qquad\qquad = 411$ volts

$$\text{Pri. rms current} \simeq \frac{e_s}{e_p} \times \frac{\text{peak sec. current} \times 100}{\text{efficiency in } \%}$$

$$\simeq \frac{411}{117} \times \frac{200 \times 100}{89}$$

$$\simeq 790 \text{ ma}$$

Assuming the efficiency to be 89 per cent for an output power of 370×0.2, or 74 watts

$$\text{pri. } VA \simeq 117 \times 0.790 = 92.4$$
$$\text{Total sec. } VA \simeq 2 \times 411 \times 0.141 = 116$$

NOTE: The secondary volt-ampere rating is actually equal to the product of the total induced secondary voltage and the secondary rms current.

Average of pri. and sec. $VA \simeq 104$.

PRIMARY CURRENT (ASSUMING RECTANGULAR WAVESHAPE)

$$PEAK = R\text{-}M\text{-}S \approx \frac{N_s}{N_p} \times SEC \; PEAK \; CURRENT$$

$$\approx \frac{e_s}{e_p} \times \frac{SEC \; PEAK \; CURRENT \times 100}{EFFICIENCY}$$

SECONDARY CURRENT (ASSUMING RECTANGULAR WAVESHAPE)

$$PEAK = H$$

$$AVERAGE = \frac{H}{2}$$

$$R\text{-}M\text{-}S = 0.707 H$$
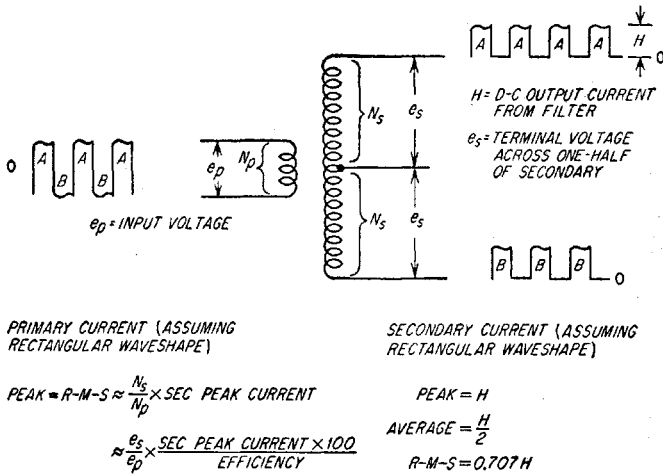
FIG. 14.29. Current waveforms in a single-phase power transformer for a full-wave rectifier and a choke-input filter.

2. Determine the primary and secondary wire sizes.

From Fig. 14.19, the ratio of wire circular mils to the rms current in milliamperes should be approximately 0.85 for a 40°C rise.

Therefore,

$$\frac{\text{cir mils}}{\text{ma}} = 0.85$$

$$\text{Pri. cir mils} = 0.85 \times 790 = 672$$

therefore use No. 21 plain enamel wire (from Table 14.3).

$$\text{Sec. cir mils} = 0.85 \times 141 = 120$$

therefore use No. 29 plain enamel wire (from Table 14.3).

3. Determine the tentative thickness of the lamination stack.

From Fig. 14.24 it can be determined that the effective area of the core material A_e will be approximately 1.9 sq in.

$$A_c = \frac{A_e}{\text{stacking factor}} = \frac{1.9}{0.88} = 2.16 \text{ sq in.}$$

$$\text{Thickness of core stack} = \frac{A_c}{\text{lamination-tongue width}} = \frac{2.16}{1.375} = 1.57 \text{ in.}$$

4. Determine the core flux density.

The maximum flux density should be below 80,000 lines/sq in. which would cause the exciting current to be excessive; therefore, assume that the maximum flux density is to be 76,000 lines/sq in. The operating core loss per pound can be the value determined from Fig. 14.25 provided the established flux density maximum of 76,000 lines/sq in. is not exceeded. It can be seen from Fig. 14.25 that the permissible core loss is about 0.9 watts/lb; however, the core loss of this particular lamination at a maximum flux density of 76,000 lines/sq in. and a frequency of 60 cps is only 0.8 watt/lb as determined from Fig. 14.21. Therefore, the maximum operating core loss per pound is determined by the limitation imposed by the danger of an excessive exciting current instead of the heating due to the core loss.

5. Determine the tentative number of primary turns required to operate the core at a maximum flux density of 76,000 lines/sq in.

From Eq. (14.5)

$$N_p = \frac{e_p \times 10^8}{4.44 f A_c B_{max}} = \frac{117 \times 10^8}{4.44 \times 60 \times 1.9 \times 76,000}$$
$$= 304$$

$$N_s \text{ (total)} \simeq N_p \frac{2e_s}{e_p} = 304 \times \frac{2 \times 411}{117}$$
$$\simeq 2,136$$

6. Establish whether or not the tentative number of primary and secondary turns will fit in the lamination window in accordance with the winding information given in Fig. 14.23 and Tables 14.3 and 14.4.

Primary Winding

Coil length..................................	2
Margins....................................	$\frac{5}{32}$
Winding length.............................	$1\frac{11}{16}$
Turns per layer.............................	51
Number of layers...........................	6
Diameter of wire...........................	0.030
Layer insulation thickness...................	0.005

Build-up of primary wire...........................	0.180
Build-up of primary insulation......................	0.025
Total primary..................................	0.205

Secondary Winding

Coil length..................................	2
Margins....................................	$\frac{1}{8}$
Winding length.............................	$1\frac{3}{4}$
Turns per layer.............................	119
Number of layers*...........................	18
Diameter of wire...........................	0.0122
Layer insulation thickness...................	0.0015

Build-up of secondary wire.........................	0.220
Build-up of secondary wire.........................	0.026
Total secondary................................	0.246

Total Fill in Window

Winding form................................	0.075
Primary thickness............................	0.205
Insulation over primary.......................	0.020
Electrostatic shield...........................	0.002
Insulation over shield.........................	0.025
Secondary thickness..........................	0.246
Outside wrapper.............................	0.025
Total thickness of coil........................	0.598
Width of window............................	0.688
Window fill (see Fig. 14.23)...................	87%

NOTE: If the build-up were too great, it would be necessary to increase the depth of the stack so that the number of turns could be decreased to maintain the same flux density. If the window were not sufficiently filled, the depth of the stack could be decreased with an associated increase in the number of turns.

* Number of layers has been made even for ease of center tapping winding, and total turns have been increased to 2,142.

7. Determine the resistances of primary and secondary windings based on the tentative number of turns established in step 5.

From Eq. (14.14)

$$R_p = 304 \times \frac{12.8}{6,000} [(1.375) + (1.57) + (\pi \times 0.178)]$$
$$= 2.27 \text{ ohms}$$

$$2R_s \text{ (total sec. resistance)} = 2,142 \times \frac{81.83}{6,000} [(1.375) + (1.57) + (\pi \times 0.450)]$$
$$= 128 \text{ ohms}$$
$$R_s \text{ (}\tfrac{1}{2} \text{ total sec. resistance)} = 64 \text{ ohms}$$

8. Determine the corrected number of primary and secondary turns based on the calculated values of resistance.

$$\left(\begin{array}{l}\text{Desired induced voltage in } \tfrac{1}{2} \text{ sec. wind-} \\ \text{ing}\end{array}\right) = \left(\begin{array}{l}\text{peak sec. current} \times R_s \times 1.11 + \text{desired} \\ \text{rms terminal voltage for } \tfrac{1}{2} \text{ sec. winding}\end{array}\right)$$
$$= (0.200 \times 64 \times 1.11) + 411$$
$$= 425.2 \text{ volts}$$

$$\left(\begin{array}{l}\text{Pri. induced voltage based on a pri.} \\ \text{resistance of 2.27 ohms}\end{array}\right) = \text{line voltage} - (i_p R_p)1.11$$
$$= 117 - (0.790 \times 2.27)1.11$$
$$= 117 - 1.99 = 115 \text{ volts}$$

$$\left(\begin{array}{l}\text{Corrected number of pri. turns based on} \\ \text{calculated values of induced voltages}\end{array}\right) = \frac{\text{total sec. turns}}{2} \times \frac{\text{pri. induced voltage}}{\text{induced voltage in } \tfrac{1}{2} \text{ sec.}}$$
$$= 1,071 \times \frac{115}{425.2}$$
$$= 290 \text{ turns}$$

NOTE: Since the above solution for primary turns is based on the resistance of 304 primary turns, it is necessary to redetermine the primary induced voltage and the corrected number of primary turns based on the value of resistance corresponding to 290 turns, i.e., $2.27 \times 290/304$, or 2.17 ohms. Repeating the solutions to the two preceding equations will again give a value of primary turns equal to 290.

9. Determine the corrected depth of the lamination stack based on the number of primary turns established in step 8 and a flux density of 76,000 lines/sq in. Since the maximum allowable flux density has been established on the basis of limiting the exciting current, the required core area should be calculated on the basis of no secondary loading. Under this condition, the primary induced voltage (and exciting current) is the greatest and is very nearly equal to the applied voltage.

From Eq. (14.5)

$$A_e = \frac{e_p \times 10^8}{4.44 f N_p B_{max}} = \frac{117 \times 10^8}{4.44 \times 60 \times 290 \times 76,000}$$
$$= 1.99 \text{ sq in.}$$

$$A_c = \frac{A_e}{\text{stacking factor}} = \frac{1.99}{0.88} = 2.26 \text{ sq in.}$$

$$\text{Stack depth} = \frac{A_c}{\text{width of lamination tongue}} = \frac{2.26}{1.375} = 1.64 \text{ in.}$$

NOTE: The increase in stack depth will cause a small increase in winding resistances. If the increase is excessive, adjust further the number of turns as shown in step 8.

10. Determine the core loss (refer to Fig. 14.28 for core weight).

$$\text{Core loss} = \text{core weight} \times \text{watts loss per lb}$$
$$= (4.17 \times 1.64 \times 0.88) \times 0.8$$
$$= 4.82 \text{ watts}$$

11. Determine the copper loss.

$$\text{Copper loss} = [(\text{pri. rms current})^2 R_p] + [(\text{sec. rms current})^2 \times 2R_s]$$
$$= (0.790)^2 \times 2.17 + (0.141)^2 \times 128$$
$$= 1.35 + 2.56$$
$$= 3.91 \text{ watts}$$

12. Determine the efficiency.

$$
\begin{aligned}
\text{Efficiency} &= \frac{\text{output power}}{\text{output power} + \text{copper loss} + \text{core loss}} \times 100 \\
&= \frac{(370 \times 0.200) \times 100}{(370 \times 0.200) + 4.82 + 3.91} \\
&= 89.5\%
\end{aligned}
$$

13. Estimate the temperature rise in the transformer.

$$
\begin{aligned}
\text{Surface area} &= 59.0 \text{ sq in.} \\
\text{Dissipation} &= 8.73 \text{ watts}
\end{aligned}
$$

From Fig. 14.26 the temperature rise is estimated as being 31°C.

14.4. Iron-core Chokes.   The inductance of a given iron-core choke is dependent on whether or not the choke carries direct current and the magnitude and frequency of the a-c voltage across the choke.   The direct current establishes the d-c flux density in the core and, hence, the value of flux density around which the a-c hysteresis loop is located.   The a-c or incremental permeability of the core with the direct current present can be determined from curves supplied by the manufacturer.   The total a-c reluctance of the core and air gap can then be established, and in turn the inductance can be calculated.

14.4a. Determination of the D-C Flux Density.   The degree of d-c flux saturation is a function of the core length, length of the air gap in the core, d-c current in the winding, number of turns, and the magnetizing characteristics of the core material.

The d-c flux density can be determined graphically.   The magnetomotive force, i.e., number of ampere-turns or gilberts, is first established, and the assumption is made that the total d-c magnetomotive force is applied across the air gap in the core. The flux density which would exist in the air gap for the assumed condition is located on the ordinate of the magnetizing curve for the core material.   For this analysis, it is necessary that the magnetizing curve be plotted with linear coordinates.   The next step is to assume that the total d-c magnetomotive force is applied only along the length of the iron core, i.e., no air gap.   The magnetomotive force per unit length of core material which would exist under this assumption, i.e., ampere-turns per inch or oersteds, is then located on the abscissa of the magnetizing curve.   If a straight line is drawn between this point and the point previously located on the ordinate, the intersection of the line and the magnetizing curve will establish the core d-c flux density and the effective d-c magnetomotive force per unit length of core material which will exist when the air gap is in series with the core.

14.4b. Determination of the A-C Flux Density.   In the absence of core saturation, the a-c flux density can be considered as not being a function of the d-c flux density; hence, the a-c flux density can be determined from Eq. (14.5).   The peak values of the hysteresis loop will be equal to the sum and difference of the d-c and a-c flux densities.   The a-c or incremental permeability for the iron can be established from a family of curves of the type shown in Fig. 14.22.

14.4c. Calculation of Inductance.   The a-c or effective permeability μ_e of the combined iron core and air gap can be determined from Eq. (14.8).   The inductance of the choke can then be established by using Eq. (14.7).

14.4d. Approximate Product of Turns and Effective Core Area.   Empirical equations for determining the approximate product of turns and effective core area as a function of the minimum desired inductance are of considerable value in the early design stages of an iron-core choke.   Equations (14.15) and (14.16) are applicable for 4 per cent silicon and C-97 Hipersil cores, respectively.

$$
\begin{array}{lll}
NA_e \geq 0.25 LI_{DC} \times 10^4 & \text{(for 4\% silicon steel)} & (14.15) \\
NA_e \geq 0.18 LI_{DC} \times 10^4 & \text{(for C-97 Hipersil)} & (14.16)
\end{array}
$$

where N = number of turns

A_e = effective area of core, sq in.

I_{DC} = d-c current, amp

NOTE: These equations are based on an appreciable d-c magnetomotive force per unit length of core material and a low a-c flux density. Similar equations can be written for other core materials provided complete design data and measured inductance values are available.

In most instances the required value of NA_e will be lower than the values obtained from the equations since the equations are based on very low values of a-c flux density. These equations are of value in determining a tentative product for NA_e and, consequently, permit the designer to make an estimate as to the physical size of the choke. The actual inductance of the choke can then be determined for the particular NA_e product in the manner outlined in Example 14.3. Depending on the calculated inductance, the product NA_e can be decreased or increased as required.

14.4e. Inductance versus Air Gap. In an inductance carrying direct current, it is usually necessary to insert an air gap in series with the core to reduce the d-c saturation of the iron.[1] Insertion of the air gap increases the incremental permeability of the iron; however, the air gap presents an additional reluctance to the a-c flux. There is an optimum length of air gap which minimizes the sum of the air gap and iron-core a-c reluctances. Figure 14.30 is a typical plot of inductance versus the length of the air gap in the core. It should be pointed out that there is an effective air gap in series with the core even when the laminations are interleaved. The following table is typical for 29-gauge (0.014-in.) laminations provided the stacking has been carefully done.

Type of Stacking	Equiv. Gap, Total, In.
1:1	0.0005
4:4	0.001
8:8	0.002
16:16	0.004
Butt joint	0.005

14.4f. Copper and Core Losses in a Choke. In a choke, the ratio of wire circular mils to the rms current through the winding cannot be specified as readily as in a power transformer. In most chokes, the core loss per pound due to the a-c flux density is much less than in a power transformer. For this reason it is usually possible to have a smaller ratio of wire circular mils to the rms current in a choke than in a power transformer of the same physical size. If the losses are primarily copper losses, the ratio of wire circular mils to the rms current in milliamperes can be reduced to about seven-tenths the value recommended for a power transformer of the same physical size.

14.4g. Swinging Chokes. It has been stated that the inductance of any given choke is dependent on the degree of the d-c flux saturation and the magnitude and frequency of the applied a-c voltage. It follows, therefore, that the inductance of an iron-core choke will vary if the d-c current through the choke is changed. This phenomenon results in the terminology "swinging choke." All iron-core chokes will swing in inductance a certain amount, although those with relatively small air gaps in the core will usually exhibit the largest change in inductance as a function of a change in the d-c current.

[1] In the design of a high-Q audio choke which may or may not have direct current through the winding, it usually is necessary to insert an air gap in the core to maximize the Q (see Sec. 14.1i).

Example 14.3

Design a choke of at least 2.5 henrys when carrying 0.40 amp d-c and with an a-c voltage of 165 at 120 cps across it. Use 4 per cent silicon iron and assume that the magnetic properties of the core material are given in Figs. 14.2, 14.21, and 14.22. Plot the inductance as a function of the air gap.

Solution

1. Solve for the tentative value of NA_e.
From Eq. (14.15)

$$NA_e \geq 0.25LI_{DC} \times 10^4$$
$$\geq 0.25 \times 2.5 \times 0.40 \times 10^4$$
$$\geq 2,500$$

2. Determine a possible combination of N and A_e.
A. Assume an effective core area A_e.
Arbitrarily assume that the lamination to be used is Allegheny Ludlum type EI-112 (tongue width = 1.125 in., window = 0.5625 by 1.6875 in., length of flux path through core = 6.75 in.) and that the effective core area A_e is 1.9 sq in.
B. Determine the ratio of wire circular mils to rms milliampere of current for a power transformer with a core size equal to that assumed in step A. Refer to Figs. 14.24 and 14.19.
The ratio should be 0.85.
C. With the tentative assumption that the heating will be entirely in the coil, reduce the ratio of wire circular mils to rms current to approximately seven-tenths the value determined in step B.

$$\text{Corrected ratio} = 0.7 \times 0.85 = 0.60$$

D. Determine the wire size and establish the number of turns which will fit into the lamination window.

$$\text{Wire cir mils} = 0.60 \times \text{rms current, ma}$$
$$= 0.60 \times 400$$
$$= 240$$

therefore use No. 26 plain enamel wire.
In the window it is possible to put 1,400 turns of No. 26 plain enamel wire. This is based on a coil length of 1.625 in., a winding length of 1.375 in. and 20 layers. The winding form has been assumed to be 0.090 in. thick, the insulation between layers to be 0.0022 in., and the winding to have a 0.030-in. wrapper. This configuration gives an 88 per cent window fill which is satisfactory, as can be seen from Fig. 14.23.
E. Compare the product of NA_e as determined in steps A and D to the product established in step 1. If the product is too large or small, it is necessary to repeat steps A to D.
In this example, the product of NA_e from steps A and D is equal to $1,400 \times 1.9$ or 2,660. This is close enough to the value determined in step 1 so that a new estimate need not be made.
3. Determine the total d-c magnetomotive force in gilberts.
From Table 14.1

$$\text{mmf} = 1.256NI_{DC}$$
$$= 1.256 \times 1,400 \times 0.40$$
$$= 703 \text{ gilberts}$$

4. Assuming that the entire magnetomotive force is applied along the iron core, determine the number of oersteds.

$$H = \frac{\text{mmf}}{\text{length of flux path through core, cm}} = \frac{703}{17.15} = 41 \text{ oersteds}$$

5. Assuming that the entire magnetomotive force is applied across the air gap, determine the d-c flux density in the gap.
NOTE: In an iron-core choke with an air gap, the laminations are stacked butt joint, i.e., all the E's are stacked together with the same orientation and, consequently, are not interleaved with the I's. The gap spacer is the insulating material which determines the physical spacing between the E's and I's. Since there is an optimum size of gap spacer, steps 5 to 9 should be repeated for several different assumed thicknesses of the gap spacer to determine the optimum. This particular series of calculations has been completed for a 0.030-in. air gap. A 0.030-in. gap is equivalent to a 0.015-in. gap spacer since the flux must pass through the gap twice.

$$B = \frac{mmf}{\text{length of gap, cm}}$$

$$= \frac{703}{0.030 \text{ in.} \times 2.54}$$

$$= 9{,}230 \text{ gausses}$$

6. Determine the effective d-c magnetomotive force per unit length of core material.

Draw a straight line on Fig. 14.2 which would connect 9.23 kilogauss on the ordinate to 41 oersteds on the abscissa. The intersection of this line with the magnetizing curve is at a flux density of 8.7 kilogauss and a magnetomotive force per unit length of core H_{DC} equal to 2.2 oersteds.

7. Determine the a-c flux density and incremental permeability.

From Eq. (14.5)

$$B_{max} = \frac{e \times 10^8}{4.44 \, fNA_e}$$

$$= \frac{165 \times 10^8}{4.44 \times 120 \times 1{,}400 \times 1.9}$$

$$= 11{,}600 \text{ lines/sq in., or } 1{,}800 \text{ gauss}$$

From Fig. 14.22 the value of μ_{ac} is approximately 900.

8. Determine the effective a-c permeability of the combination of the core and air gap.

From Eq. (14.8)

$$\mu_e = \frac{1}{\dfrac{1}{900} + \dfrac{0.030}{6.75}}$$

$$= \frac{1}{0.00111 + 0.00444}$$

$$= \frac{1}{0.00555}$$

$$= 180$$

9. Determine the inductance of the choke.

From Eq. (14.7)

$$L = \frac{3.19 N^2 A_e \mu_e}{l_c \times 10^8}$$

$$= \frac{3.19 \times (1{,}400)^2 \times 1.9 \times 180}{6.75 \times 10^8}$$

$$= 3.17 \text{ henrys}$$
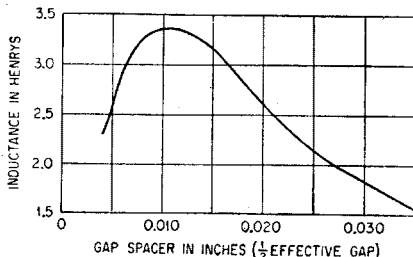
FIG. 14.30. Typical curve illustrating how the inductance of a choke carrying direct current varies as a function of the air gap.

NOTE: The inductance of this choke for different air gaps has been plotted in Fig. 14.30.

10. Determine the resistance of the winding.

$$A_c = \frac{A_e}{\text{stacking factor}} = \frac{1.9}{0.94} = 2.02 \text{ sq in.}$$

$$\text{Stack depth} = \frac{A_c}{\text{tongue width}} = \frac{2.02}{1.125} = 1.8 \text{ in.}$$

From Eq. (14.14)

$$R = \frac{1{,}400 \times 40.81}{6{,}000} [(1.125) + (1.8) + (\pi \times 0.28)]$$

$$= 36.2 \text{ ohms}$$

11. Determine the copper and core losses.

$$I^2 R = (0.4)^2 \times 36.2 = 5.8 \text{ watts}$$

The operating a-c flux density is only 11,600 lines/sq in. Therefore, the core loss will be approximately 0.08 watts/lb[1] (from Fig. 14.21). The total core weight of a 1.8-in. stack of EI-112 laminations is 3.8 lb; therefore, the total core loss is approximately 0.3 watt.

---

[1] Core loss per pound with a d-c magnetization force present is not necessarily the same as the values obtained from the normal core loss curves. Manufacturers usually do not supply this information.

The total loss in the choke is 6.1 watts, and the surface area of the choke is 49 sq in.   The temperature rise is, therefore, approximately 26°C based on Fig. 14.26.

14.5. Audio Transformers.   Audio transformers are distinguished from power transformers in that they are required to pass a range of frequencies.   The problem of designing a broad-band transformer becomes increasingly difficult with both higher impedance levels and higher power ratings and if the transformer has direct current

(*a*) LOW FREQUENCY EQUIVALENT CIRCUIT   (*b*) MIDBAND EQUIVALENT CIRCUIT

(*c*) HIGH FREQUENCY EQUIVALENT CIRCUIT

$L'_p = 2L_p(1-k)$
= LEAKAGE INDUCTANCE REFERRED TO PRIMARY TERMINALS
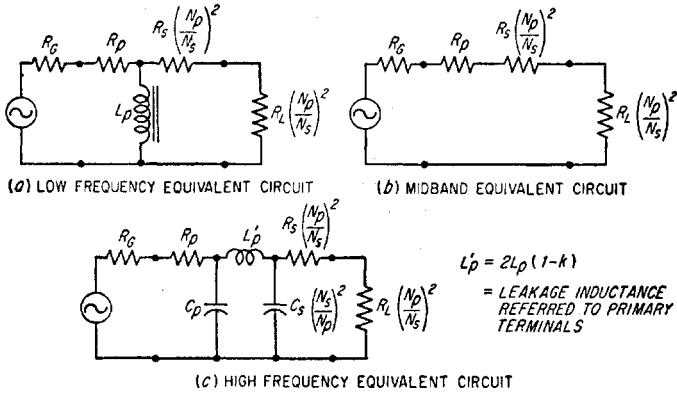
FIG. 14.31. Equivalent circuits for a transformer and driving source.   The modified secondary circuit elements, e.g., $R_s(N_p/N_s)^2$, are the equivalent secondary circuit values in a 1-to-1 turn ratio transformer having the same frequency response.
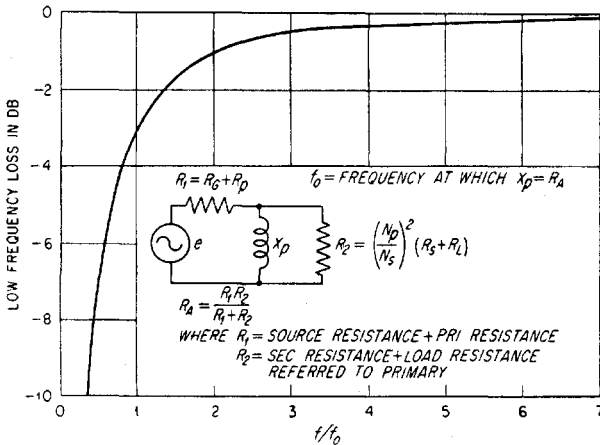
FIG. 14.32. Low-frequency response of a transformer.

through one of the windings.   For a given type of audio transformer, the relative difficulty of design depends on the number of octaves encompassed rather than the actual range in cycles.

14.5a. Input Transformers.   Input transformers couple input signals to the grid of the first amplifier tube.   They usually operate at very low power levels and, consequently, should be enclosed in a magnetic shield.   In most applications, the function of the input transformer is to provide the maximum possible voltage gain for the input signal without reducing the bandwidth below that required.

14.5b. *Interstage Transformers.* An interstage transformer is used to couple the plate circuit of one amplifier stage to the grid of another. It should provide the maximum possible gain for the bandwidth desired. Interstage transformers are frequently operated from a single-ended amplifier stage. Under these conditions the low-frequency response of the transformer will be impaired if direct current flows through the winding. This effect can be eliminated by supplying the plate voltage
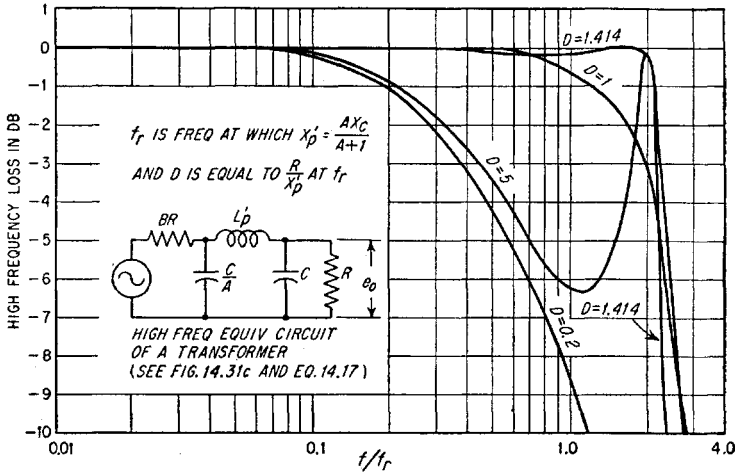
Fig. 14.33. High-frequency response of a transformer where $A = B = 1$.
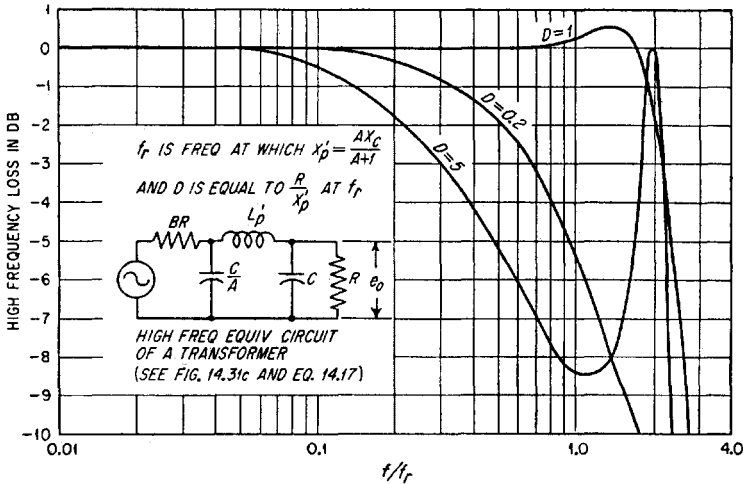
Fig. 14.34. High-frequency response of a transformer where $A = 1$ and $B = 2$.

to the tube through a resistor. The transformer primary should then be coupled to the plate of the tube through a capacitor. In certain applications, the capacitor is made to resonate with the primary inductance at some low frequency to accentuate low frequencies. For best results, the Q of the primary inductance should be high.

14.5c. *Output Transformers.* Output transformers are used to change the imped-ance level of the output signal to the impedance level of the load and/or to provide d-c isolation from the amplifier.

14.5d. Driver Transformers.   A driver transformer is a transformer which couples the plate or plates of an amplifier stage to the grids of a class AB2 or a class B2 stage. In either a class AB2 or class B2 stage, the grids are driven into the positive region for a portion of each cycle; hence, the transformer must deliver power.   The load resistance on the transformer is very high when the grids are in the negative region and relatively very low when the grids are in the positive region.   The changing load on the
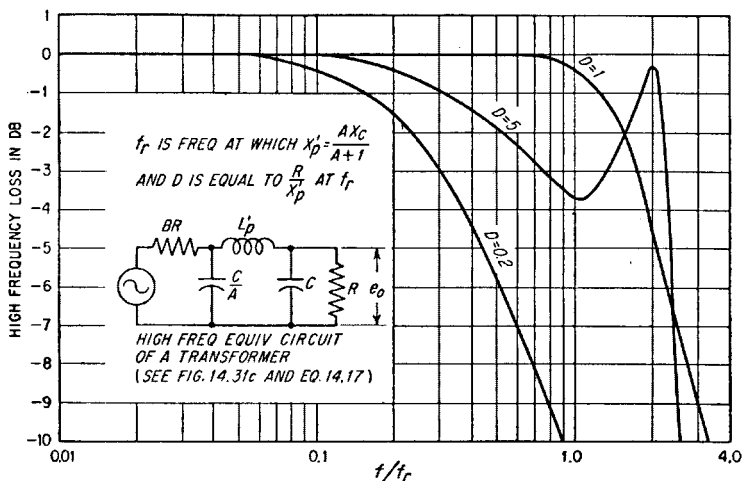
FIG. 14.35. High-frequency response of a transformer where $A = 1$ and $B = 0.5$.
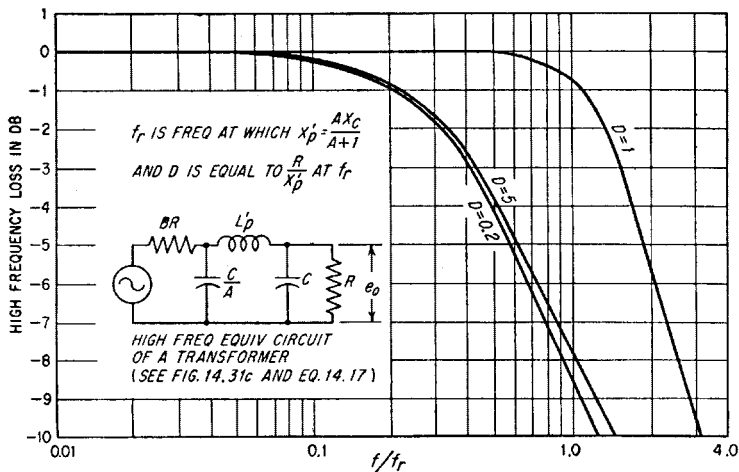
FIG. 14.36. High-frequency response of a transformer where $A = 0.1$ and $B = 1$.

transformer secondary is reflected to the driver tube or tubes and tends to cause distortion.   To reduce the effect on the driver tube, the transformer is made with a step-down ratio.   Associated with too large a step-down ratio is a decrease in the power which can be delivered to the grids of the output tubes, viz., when the positive grid input resistance is greater than the transformer output impedance; hence, there must be a compromise between distortion and power.

The winding resistances and leakage inductance are effectively in series with the secondary load as shown in Fig. 14.31c. Therefore, to minimize the poor regulation (a source of distortion) in the transformer output voltage caused by the large grid currents which flow for only a portion of each cycle, the transformer winding resistances and leakage reactance should be small. In addition, the current surges can cause transient ringing in the leakage inductance and distributed capacitances.
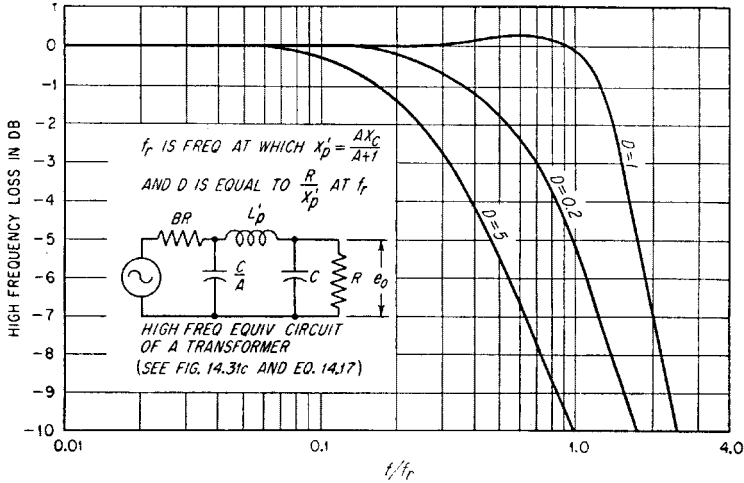
FIG. 14.37. High-frequency response of a transformer where $A = 0.1$ and $B = 2$.
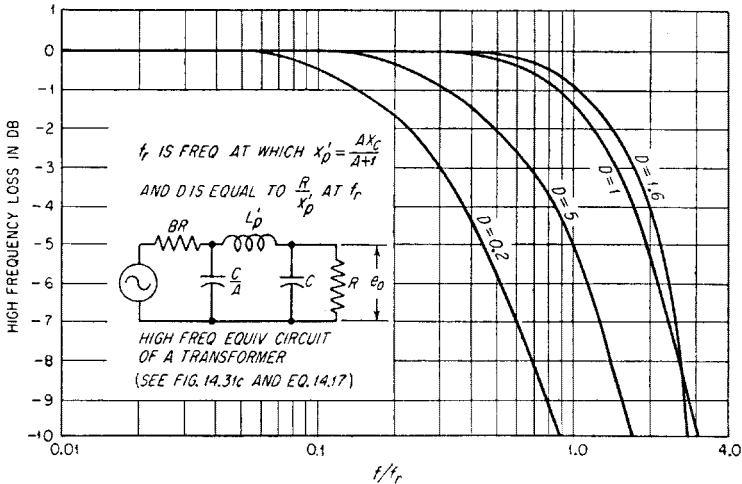
FIG. 14.38. High-frequency response of a transformer where $A = 0.1$ and $B = 0.5$.

14.5e. Equivalent Circuits. The frequency response of an audio transformer is best analyzed by examining the equivalent circuits shown in Fig. 14.31.

If a transformer should have two or more secondary windings, each winding is referred to the primary winding in the same manner as shown for a transformer having a single secondary winding. The circuit elements in any given secondary circuit are, therefore, referred to the primary, utilizing the turns ratio applicable to that

particular winding.   The equivalent secondary circuits so drawn are connected in parallel in the over-all equivalent circuit.

Low-frequency Response.   The circuit elements affecting the low-frequency response are shown in Fig. 14.31*a*, and the normalized response in decibels is plotted in **Fig. 14.32.**   The inductance of the primary can be calculated in the manner described in Sec. 14.1*e* if the winding carries no direct current.   If the winding carries direct
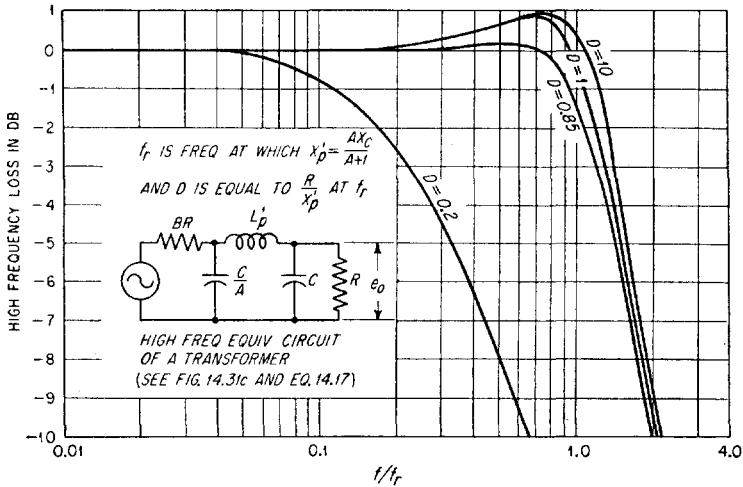
Fig. 14.39. High-frequency response of a transformer where $A = 10$ and $B = 0.1$.
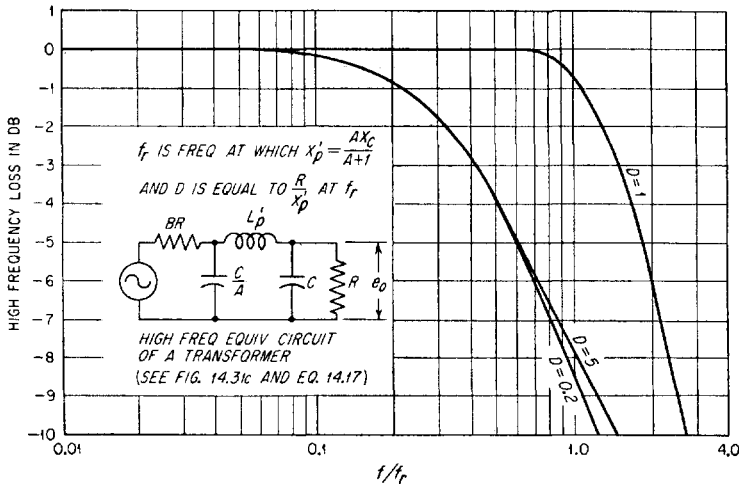
Fig. 14.40. High-frequency response of a transformer where $A = 10$ and $B = 1$.

current, refer to Sec. 14.4 and determine the primary inductance in the same manner used to calculate the inductance of a choke.

Midband Response.   The midband performance of the transformer is dependent only on the ratio of the reflected secondary winding and load resistances as compared to the source and primary resistances as shown in Fig. 14.31*b*.   From this equivalent circuit, the power dissipated in the secondary load can be determined.   The midband secondary terminal voltage can be determined by correcting the voltage across the secondary load of the equivalent circuit by the turns ratio, N_s/N_p.
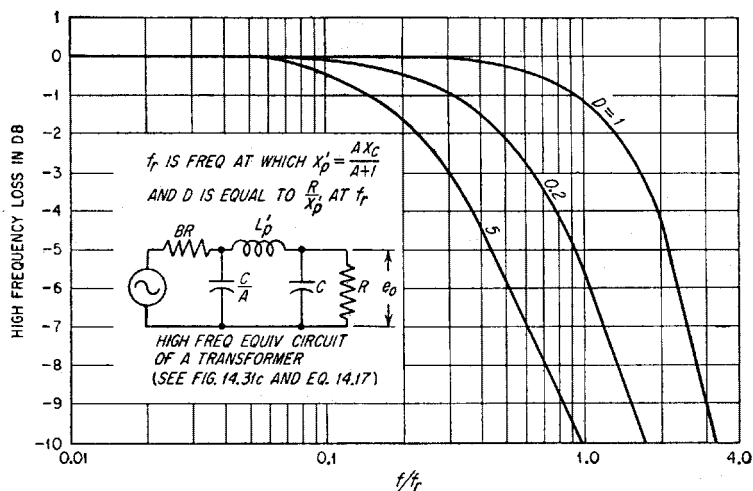
FIG. 14.41. High-frequency response of a transformer where $A = 10$ and $B = 2$.

High-frequency Response. The equivalent circuit to be used in determining the high-frequency response is shown in Fig. 14.31c. Equation (14.17) can be used to calculate the high-frequency response of the transformer; however, Figs. 14.33 to 14.41 suffice for most applications.

$$\frac{e_o}{e_{\text{midband}}} = \frac{X_C{}^2(B+1)}{\sqrt{(U+S)^2 + (V-T)^2}} \qquad (14.17)$$

where X_C = reactance of capacitance C

C = sec. circuit shunt capacitance reflected to pri. circuit

$\quad = C_s(N_s/N_p)^2$

X'_p = reactance of leakage inductance L'_p

L'_p = total leakage inductance referred to pri. circuit

$A = C/C_p$

$B = (R_G + R_p)N_s{}^2/(R_s + R_L)N_p{}^2$

C_p, C_s = total pri. and sec. circuit shunt capacitances, respectively

N_p, N_s = numbers of pri. and sec. turns, respectively

$R = (R_s + R_L)(N_p/N_s)^2$

R_G = resistance of source driving transformer

R_L = resistance of transformer load

R_p, R_s = pri. and sec. winding resistances, respectively

$S = X_C(X'_p - X_C)$

$T = \dfrac{X'_p X_C{}^2}{R}$

$U = BX_C \left(\dfrac{X'_p}{A} - X_C \right)$

$V = BR \left(\dfrac{X'_p - X_C}{A} - X_C \right)$

C_p, C_s and L'_p can be determined by utilizing the data given in Secs. 14.2d and 14.2e.

Impedance Ratio. For a given transformer, the impedance ratio is given by Eq. (14.18).

$$\frac{Z_p}{Z_s} = \frac{N_p{}^2}{N_s{}^2} \qquad (14.18)$$

14.5f. Balanced Windings. Balanced windings are windings in which the resistances and distributed capacitances are identical for each half of the winding. To obtain a balanced winding, it is necessary to make two identical coils. One of the coils is reversed to the other as it is assembled on the core. Connecting the starts of the two coils together will put the two coils in series. Ordinarily the midpoint of a balanced winding is operated at zero signal potential, e.g., at ground potential.

14.5g. Physical Size of Audio Transformers. The size of low-power audio transformers is usually independent of the power requirements. If such is the case, the required low-frequency response determines the primary inductance, and the impedance ratings of the windings determine the maximum allowable winding resistances. Common practice is to make the resistance of a given winding considerably less than one-tenth the impedance rating of the winding.

The size of an audio transformer which delivers more than a few watts of power is usually determined by the number of turns and the core size required to prevent core heating or unacceptable values of distortion due to excessive flux densities at the lower operating frequencies.

Example 14.4

Neglecting the winding resistances, determine the frequency response of the output transformer whose circuit and winding details are given in Fig. 14.42 and whose incremental permeability curves are shown in Fig. 14.22.
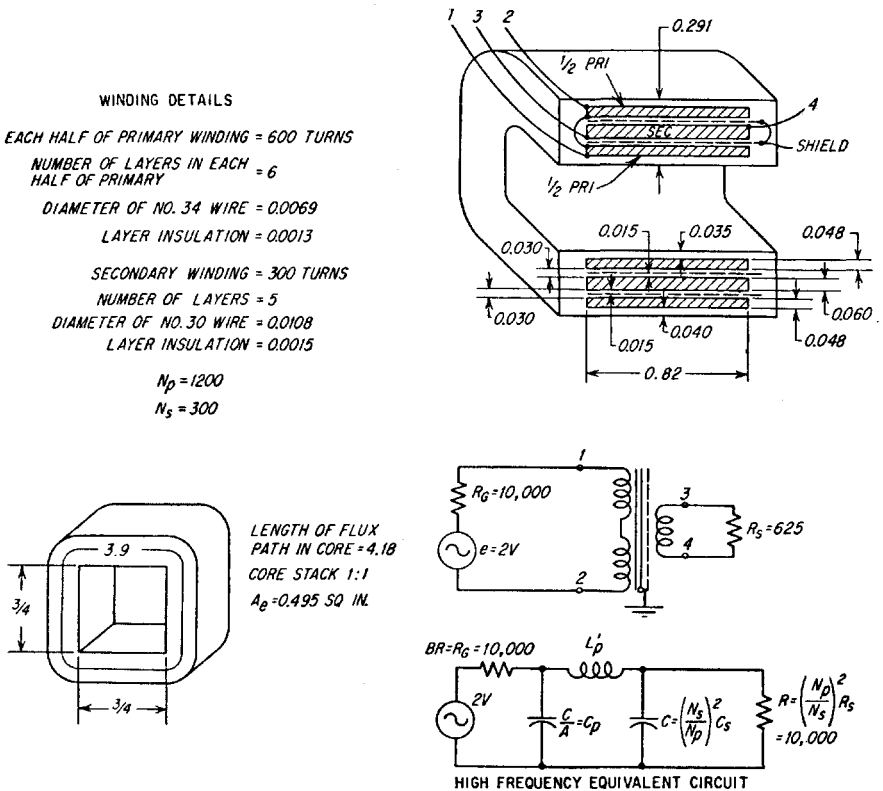
FIG. 14.42. Winding information and circuit for the output transformer analyzed in Example 14.4.

Solution

1. Determine the low-frequency response.

A. Determine the maximum flux density and the incremental permeability of the core at some arbitrarily chosen low frequency, e.g., 100 cps.
From Eq. (14.5)

$$B_{max} = \frac{1 \times 10^8}{4.44 \times 100 \times 1,200 \times 0.495}$$

$$= 379 \text{ lines/sq in., or } 58.8 \text{ gauss}$$

NOTE: The voltage appearing across the transformer primary has been assumed to be 1 volt. This is based on the assumption that the reactance of the primary inductance at 100 cycles is sufficiently high that the terminated transformer very nearly matches the 10,000-ohm generator.
From Fig. 14.22

$$\mu_{ac} \simeq 1,500 \qquad \text{since } H_{DC} = 0$$

B. Determine the effective a-c permeability of the combination of the core and air gap.
From Eq. (14.8)

$$\mu_e = \frac{1}{\dfrac{1}{1,500} + \dfrac{0.0005}{4.18}}$$

$$= 1,270$$

NOTE: For interleaving laminations 1:1, $l_g \simeq 0.0005$ in.
C. Determine the primary inductance.
From Eq. (14.7)

$$L_P = \frac{3.19 \times (1,200)^2 \times 0.495 \times 1,270 \times 10^{-8}}{4.18}$$

$$= 6.9 \text{ henrys}$$

D. Determine frequency at which the low-frequency response is down 3 db from the midband response (neglecting winding resistances).
From Fig. 14.32

$$R_A = 5,000 \text{ ohms} = 2\pi f_0 L_p$$
and
$$f_0 = 115 \text{ cps}$$

Therefore, −3 db occurs at approximately 115 cps.

NOTE: Based on a −3-db frequency of 115 cps, the voltage across the primary at 115 cps would be 0.707 volts. Repeating steps A through D utilizing a primary voltage of 0.707 volts at a frequency of 115 cps, the revised −3-db frequency is approximately 120 cps.

2. Determine the high-frequency response.

A. Determine the leakage inductance as referred to the primary terminals.
The configuration shown in Fig. 14.14c is similar to that in this example except that the primary and secondary windings are interchanged. The associated equation is applicable in determining the leakage inductance referred to either winding provided that the number of turns for that particular winding are those used in the equation.

$$L_{p}' = \frac{0.85 \times 3.9 \times (1,200)^2}{0.82} \left(0.030 + 0.030 + \frac{0.048 + 0.060 + 0.048}{3} \right) \times 10^{-5} \text{ mh}$$

$$= 6.52 \text{ mh}$$

B. Determine the capacitance C_p across the primary winding.
The primary winding consists of two separate windings connected in series to the external circuit. The external circuit, as shown in Fig. 14.42, is not grounded; therefore, the effective capacitance across the primary winding can be determined in a manner similar to that shown in Fig. 14.18. The particular configuration differs from that shown in Fig. 14.18 only in that the winding appears in two halves. The effect of the split primary winding is to reduce the layer-to-layer capacitance between the two layers which have necessarily been separated to insert the secondary winding. The reduced layer-to-layer capacitance between the two referenced layers has small effect on the net primary capacitance; therefore, it is recommended that the primary capacitance C_p be determined as shown in Fig. 14.18.
Refer to Fig. 14.18.

$$C_p = C_D + \frac{C_{G1}C_{G2}}{C_{G1} + C_{G2}} + \text{lead capacitance}$$

$$= 135 + 32 + 20 = 187 \ \mu\mu\text{f}$$

$$C_D = \text{winding capacitance}$$

$$= \frac{0.30 \times 3.9 \times 0.82 \times 3.5(12 - 1)}{(0.0013 + 0.0006)(12)^2} = 135 \ \mu\mu\text{f}$$

$\dfrac{C_{G1}C_{G2}}{C_{G1} + C_{G2}}$ = capacitance across winding due to capacitances between inside and outside of winding to core

$$= \frac{60 \times 68}{60 + 68} = 32 \ \mu\mu\text{f}$$

where $C_{G1} = \dfrac{0.225 \times 3.9 \times 0.82 \times 3.5(12 - 0.5)}{(0.0400 + 0.0003)12} = 60 \ \mu\mu\text{f}$

$C_{G2} = \dfrac{0.225 \times 3.9 \times 0.82 \times 3.5(12 - 0.5)}{(0.0350 + 0.0003)12} = 68 \ \mu\mu\text{f}$

Lead capacitance is estimated as being equal to 20 $\mu\mu$f.

  C. Determine the capacitance C_s across the secondary winding and refer this value to the primary winding (see Fig. 14.18).

$$C_s = C_D + \frac{C_{G1}C_{G2}}{C_{G1} + C_{G2}} + \text{lead capacitance}$$

$$= 237 + 74 + 20 = 331 \ \mu\mu\text{f}$$

$$C_D = \text{winding capacitance}$$

$$= \frac{0.30 \times 3.9 \times 0.82 \times 3.5(5 - 1)}{(0.0015 + 0.00077)(5)^2} = 237 \ \mu\mu\text{f}$$

$\dfrac{C_{G1}C_{G2}}{C_{G1} + C_{G2}}$ = capacitance across winding due to capacitances between inside and outside of winding to shield

$$= \frac{147 \times 147}{147 + 147} = 74 \ \mu\mu\text{f}$$

where $C_{G1} = \dfrac{0.225 \times 3.9 \times 0.82 \times 3.5(5 - 0.5)}{(0.015 + 0.00039)5}$

$$= 147 \ \mu\mu\text{f}$$

$C_{G2} = C_{G1}$

Lead capacitance is estimated as being 20 $\mu\mu$f.

  The secondary capacitance C_s referred to primary is

$$C_s(N_s/N_p)^2 = 331(^{30}\!/_{1200})^2$$
$$= 20.7 \ \mu\mu\text{f}$$

Therefore C in Figs. 14.33 to 14.41 is equal to 20.7 $\mu\mu$f.

  D. Determine the value of A [see Eq. (14.17)].

$$A = \frac{C}{C_p} = \frac{20.7}{187} = 0.11$$

  E. Determine the frequency f_r at which $X_p' = AX_C/(A + 1)$.

$$2\pi f_r L_p' = A \frac{1}{2\pi f_r C} \frac{1}{A + 1}$$

$$f_r = \frac{1}{2\pi} \sqrt{\frac{A}{(A + 1)L_p'C}}$$

$$= \frac{1}{6.28} \sqrt{\frac{0.11}{(0.11 + 1) \times 6.52 \times 10^{-3} \times 20.7 \times 10^{-12}}}$$

$$= 136{,}000 \text{ cycles, or } 136 \text{ kc}$$

F. From Figs. 14.33 to 14.41, determine the value of *D.*
At f_r,

$$X_p' = 2\pi f_r L_p'$$
$$= 6.28 \times 136 \times 10^3 \times 6.52 \times 10^{-3}$$
$$= 5{,}570 \text{ ohms}$$

therefore,

$$D = \frac{10{,}000}{5570} = 1.80$$

NOTE: Refer to equivalent circuit shown in Fig. 14.42 for value of *R.*
G. Determine the frequency at which the output is attenuated 3 db.
The source resistance is equal to the terminating resistance in the equivalent transformer; therefore, the value of *B* is 1. It is next necessary to examine Figs. 14.33 to 14.41 to determine which curve approximates this example. Since $A = 0.11$, $B = 1$, and $D = 1.80$, the curves most nearly approximating this example are given in Fig. 14.36.
Interpolating,

$$-3\text{-db frequency} \simeq 1.1 f_r$$
$$\simeq 1.1 \times 136 = 150 \text{ kc}$$

14.6. Special Magnetic-circuit Components. Several magnetic circuits, most of which employ some form of a magnetic amplifier, are treated in the following paragraphs. Their design requires an intimate knowledge of the design of iron-core chokes carrying direct current (see Sec. 14.4).
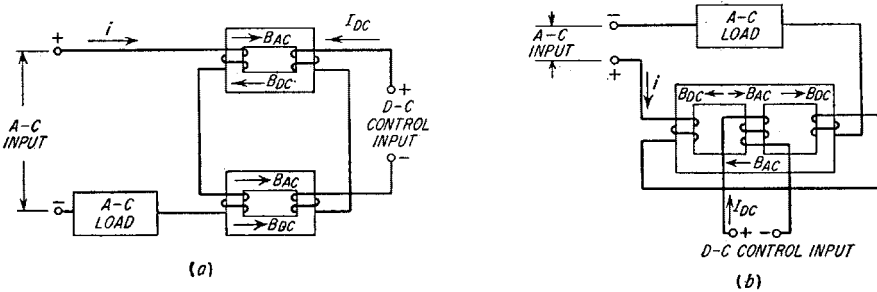
FIG. 14.43. Saturable reactors shown with instantaneous a-c currents in the windings in order to illustrate the directions of flux.

14.6a. Saturable Reactors. In a saturable reactor the degree of core saturation is controlled by a d-c winding so as to change the inductance of the reactor. Typical winding configurations are shown in Fig. 14.43. The d-c control windings are polarized so that the voltages induced from the a-c windings will oppose and cancel one another to minimize the transfer of power into the control windings. In the circuits in Fig. 14.43 the a-c flux, at any instant, will be aiding the d-c flux due to one control winding and opposing the d-c flux due to the other control winding. Consequently, at any given instant, the instantaneous impedances of the two halves of the reactor will actually differ from one another slightly, and as a result, the voltages induced in the d-c control windings will not have identical amplitudes and their cancellation will not be complete. The result is a small alternating current which flows in the d-c control windings. In most applications, however, this current is not considered objectionable.

Saturable reactors have application wherever there is a need for an adjustable reactor. For example, the power supplied to a load may be varied by changing the inductance of a saturable reactor in series with an a-c power source and the load (see Fig. 14.43). Saturable reactors introduce considerable distortion into the signal applied to the load, so their use must be limited to those cases where waveform distortion is not an objection.

The rate at which the control current can be varied is a function of the resistance and inductance in the control circuit. It is possible to increase the rate of control-current variation if resistance is placed in series with the control windings since the circuit time constant is equal to L/R. The addition of resistance requires additional power from the control circuit; hence, there is a loss in "gain" associated with the reduction of the time constant. The response time of a saturable reactor to an instantaneous control-current change is limited to approximately one cycle of the carrier frequency. However, the maximum frequency at which the control voltage can be varied and still maintain a fairly uniform response of the control current is
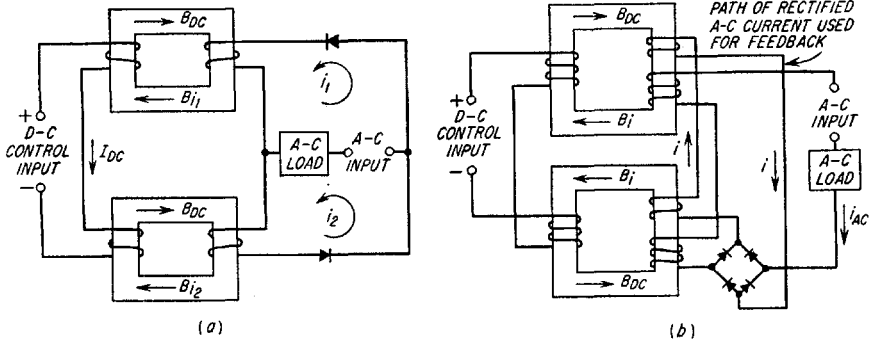
Fig. 14.44. Regenerative magnetic amplifiers which have a-c loads.
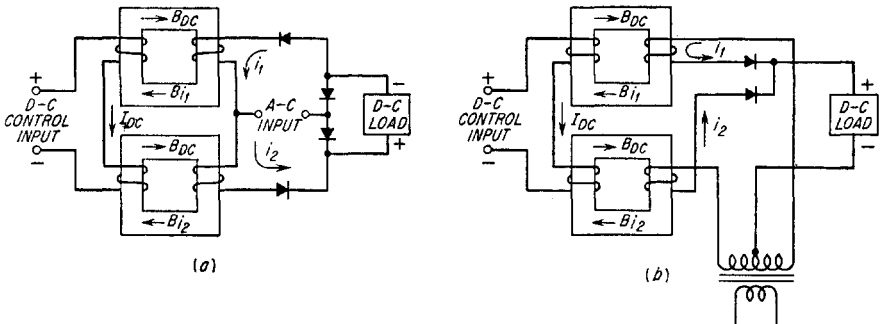
Fig. 14.45. Regenerative magnetic amplifiers which have d-c loads.

usually limited to approximately 10 or 20 per cent of the carrier frequency because of the time constant of the control circuit.

14.6b. Self-saturating Saturable Reactors. A self-saturating reactor has an additional control winding through which either a positive or negative d-c feedback current is applied. The d-c feedback current is obtained by rectification of the load current. Typical circuits which employ regeneration are shown in Figs. 14.44 and 14.45. With regeneration, the rectified load current flows through the feedback winding so as to aid the d-c control current, thereby providing more gain than is possible in the simpler form of the saturable reactor. Associated with the increase in gain is an increase in response time and a decrease in linearity. Power gains of one thousand to several hundred thousand can be obtained in regenerative magnetic amplifiers of this type. In a degenerative magnetic amplifier, there is an increase in linearity and a decrease in both gain and response time.

14.6c. Frequency Doublers. A typical magnetic circuit employed primarily for frequency doubling or d-c amplification is shown in Fig. 14.46. With no d-c control

current through the secondary windings, the instantaneous voltages induced in the individual secondary windings are equal and opposing. Thus, there is no net output voltage. With the presence of a d-c control current through the secondary windings, the two transformer cores reach a certain degree of saturation. Since the d-c control current flows in opposite directions through the two transformer secondary windings,
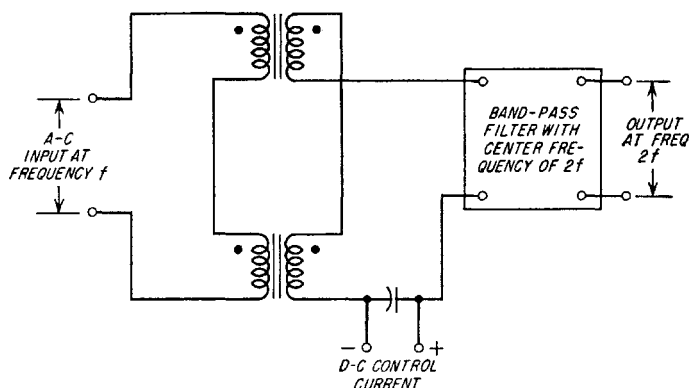
FIG. 14.46. Two transformers used to produce an a-c output proportional to a d-c control current.

it produced flux saturation on the negative voltage crest in one transformer and on the positive voltage crest in the other transformer. The instantaneous voltages induced in the two secondaries under this condition are not exactly equal and the output has a signal component at twice the input signal frequency which can be isolated by a bandpass filter. Over a fairly large range, the amplitude of the a-c output signal can be made very nearly proportional to the magnitude of the d-c control current. If the a-c output signal from the bandpass filter is rectified to produce a direct current, the over-all circuit is functioning as a d-c amplifier since the input can be considered as the d-c control current.

Another circuit for frequency doubling is shown in Fig. 14.47. In this particular circuit, the bridge is unbalanced whenever a d-c control current is applied. An unbalance in the bridge causes an output voltage to exist at twice the excitation frequency. The phase and amplitude of the output voltage are dependent on the polarity and magnitude of the control current.

FIG. 14.47. Frequency doubler using a balanced bridge. Output is proportional to magnitude of d-c control voltage. Polarity of control voltage determines phase of output.

14.6d. Current Discriminator. With the circuit of Fig. 14.47, the phase of the output voltage will be reversed if the direct current through the control winding is reversed. It follows, therefore, that in Fig. 14.48 the phase of the output voltage is dependent on whether or not I_{DC1} is larger than I_{DC2} since the two currents flow in opposite directions through the same number of turns. If the output is peak-phase-detected using the input voltage as a reference, the polarity of the detector output will indicate which d-c current is the greater, and the amplitude of the detector output will be proportional to the difference in amplitude between I_{DC1} and I_{DC2}. If better

accuracy is desired, the output of the phase detector can be amplified in a high-gain d-c amplifier and applied to an additional control winding so as to oppose the d-c core saturation resulting from the difference in the amplitudes of I_{DC1} and I_{DC2}. The feedback current will very nearly be equal to the difference between I_{DC1} and I_{DC2} if the feedback winding has the same number of turns as the other d-c windings.
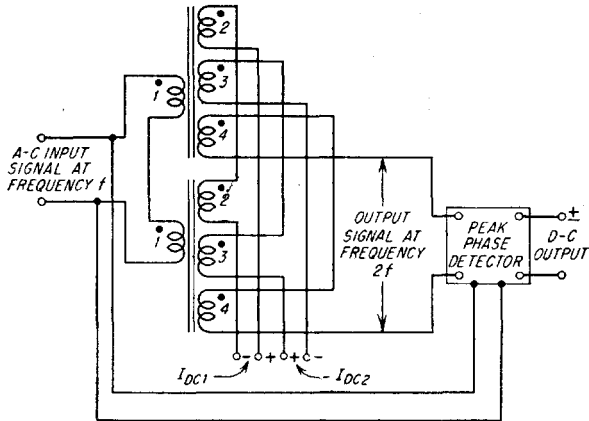
FIG. 14.48. Current discriminator.

If the nominal values of I_{DC1} and I_{DC2} are not identical, the number of turns on the two control windings should be adjusted so that the nominal values produce equal and opposite magnetizing effects.

14.6e. Saturable Transformers. Transformers connected as shown in Fig. 14.49 provide a convenient means of controlling the speed and direction of rotation of a
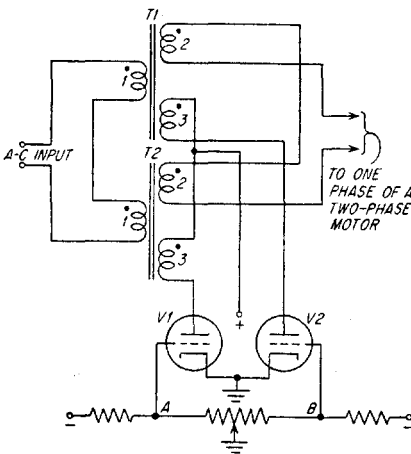
FIG. 14.49. Saturable transformers used for motor control.
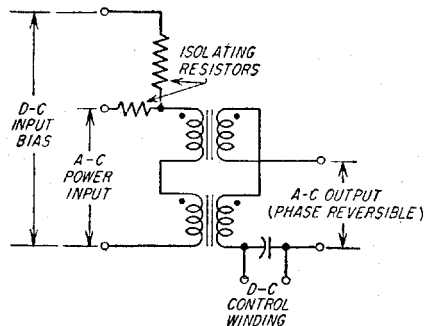
FIG. 14.50. Saturable transformers used for control of a low-power two-phase motor.

two-phase motor.   When the control potentiometer is at its mid-point, $V1$ and $V2$ will conduct equally and the output will consist of only a small 2nd harmonic component which will not drive the motor.   If the potentiometer is turned to the limit labeled A, $V1$ will conduct more heavily and $V2$ will be driven beyond cutoff.   Under this condition, the voltage induced in $T1$ is much greater than that induced in $T2$, and

consequently the transformer outputs will no longer cancel. The result is an excitation voltage applied to the motor. If the control potentiometer is turned to the other limit, the output applied to the motor will be of the opposite phase since the output of $T2$ will be larger than the output of $T1$.

Another circuit frequently used for controlling a two-phase motor is shown in Fig. 14.50. Here the d-c bias source magnetizes the cores of both transformers by equal amounts. A current in the control windings, however, aids the magnetization due to the bias current in one transformer and opposes the magnetization due to the bias current in the other transformer. It follows that a control current will unbalance the magnetization in the two transformers, causing the a-c output voltages to be unequal, which results in an a-c output voltage. The polarity and magnitude of the control current determine the phase and amplitude of the a-c output voltage.

14.7. Pulse Transformers. A rectangular pulse can be considered to consist of an infinite number of frequency components (see Sec. 22) having unique amplitude and phase relationships. Hence, to reproduce faithfully a pulse through a transformer, the transformer must introduce very nearly the same time delay and provide the same attenuation or gain to the principal frequency components in the pulse. On the assumption that the source and load resistances are constant during and after the pulse, the frequency response required to reproduce a rectangular pulse can be approximated by Eqs. (14.19) and (14.20).

$$f_1 \simeq \frac{A}{600t} \tag{14.19}$$

$$f_2 \simeq \frac{0.7}{t_r} \tag{14.20}$$

where A = allowable percentage decay of pulse from a flat top
$\quad f_1$ = lower −3-db frequency, cps
$\quad f_2$ = upper −3-db frequency, cps
$\quad t$ = pulse width, sec
$\quad t_r$ = desired rise time in seconds of pulse leading edge between 10 and 90 per cent amplitude points

In many applications, the impedance of either the pulse source or the load is low during the pulse and approaches an open circuit at the end of the pulse. For this reason, satisfying Eqs. (14.19) and (14.20) does not ensure a satisfactory trailing edge unless the source and load impedance are the same during and after the pulse. If the source or load impedance approaches an open circuit following the pulse, the trailing edge of the pulse will have a larger overshoot.

In addition to the frequency-response considerations, the leading edge of the output pulse waveform will have an overshoot of approximately 5 per cent if Eqs. (14.21) and (14.22) are satisfied (see Fig. 14.31c for high-frequency equivalent circuit). An overshoot of approximately this amount is usually desirable since it provides a significantly shorter rise time than would be obtained with critical damping.

$$R_G = \sqrt{\frac{L_p'}{C}} \tag{14.21}$$

$$R_L = \frac{N_s^2}{N_p^2} \sqrt{\frac{L_p'}{C}} \tag{14.22}$$

where $C = C_p + C_s N_s^2/N_p^2$

If the value of R_G is less than the value specified by Eq. (14.21) or if the value of R_L is larger than that specified by Eq. (14.22), the leading edge of the waveform will have a larger overshoot. In the extreme case where R_G is equal to zero and R_L is given by Eq. (14.22), the output waveform will have an overshoot of approximately

17 per cent. For this condition, the value of R_L required to obtain a 5 per cent over-shoot can be determined from Eq. (14.23).

$$R_L = \frac{0.707 N_s^2}{N_p^2} \sqrt{\frac{L_p'}{C_s}} \qquad (14.23)$$

The maximum flux density in the core caused by the application of a rectangular pulse is

$$B_{max} = \frac{et \times 10^8}{N_p A_e} \qquad (14.24)$$

where B_{max} = maximum flux density, lines/sq in.
   A_e = effective core cross-sectional area, sq in.
   e = peak value of applied pulse, volts
   N_p = number of primary turns
   t = time duration of pulse, sec

Instead of analyzing a pulse transformer on an equivalent frequency response basis, the performance can be established by determining the transformer transient response[1] for the front and trailing edges of the pulse. Also, the response of the transformer during the flat top of the pulse can be determined by analyzing the amount of voltage decay across the primary shunting inductance in the low-frequency equivalent circuit. In addition, this method permits the determination of the pulse trailing edge waveform in those applications where the source or load impedance becomes very large immediately following the pulse.

In general, considerable attention must be given to maximizing the primary inductance and minimizing the leakage inductance and distributed capacitance. Such stringent requirements dictate that the windings be interleaved and the core material have high permeability. At very high frequencies the magnetic flux will not penetrate the core laminations to any appreciable depth and, consequently, flows on the surface of the laminations. Therefore, the core laminations should be as thin as is practicable to present the maximum surface area to the flux. The effect of the thinner laminations is to increase the effective core permeability.

Ordinarily, the power loss in a pulse transformer is due primarily to the core loss since there are very few primary and secondary turns. Core-loss figures for pulse transformer applications are not normally supplied by the manufacturers of core materials. However, for 2-mil grain oriented steel, Reuben Lee has given a core-loss figure of 0.012 watt/lb of core material based on a single 2-μsec pulse/sec which causes a maximum core flux density of 6,000 gausses (38,700 lines/sq in.).

Since the physical size of a pulse transformer is usually minimized to reduce the leakage inductance and distributed capacitance, a typical pulse transformer will have a higher temperature rise than a conventional power transformer having the same power dissipation. Consequently, it is frequently necessary to use high-temperature insulating materials in high-power pulse transformers.

---

[1] For an excellent treatment of this type of analysis, see Moskowitz and Racker, "Pulse Techniques," Sec. 3.5, Prentice Hall, Inc., Englewood Cliffs, N. J., 1951. A more thorough treatment of pulse transformers is given in "Pulse Generators," vol. 5, Chaps. 12–15, Radiation Laboratory Series, McGraw-Hill Book Company, Inc., New York, 1948.

Power Supplies

15.1. Characteristics of Ideal Rectifiers with Infinite Inductance Choke-input Filters. It is possible to tabulate the transformer winding currents and the amplitudes and frequencies of the output ripple components for rectifiers which are followed by infinite inductance choke-input filters provided the transformer leakage inductance and the resistances of the transformer windings and rectifier tubes are assumed to be equal to zero. This has been done in Table 15.1. In an actual rectifier with a choke-input filter, these conditions are not satisfied, but the error in using Table 15.1 is usually negligible. To provide a means of comparing the magnitudes of the various rectifier voltages and currents in the tabulated data, the filtered d-c output voltage and current have each been assumed to have a value of unity.

15.2. Critical Inductance of the Input Choke to a Filter. Considering the input choke of a choke-input filter, *critical inductance* is that value above which the pulsating current from the rectifier becomes continuous and, consequently, ensures that the
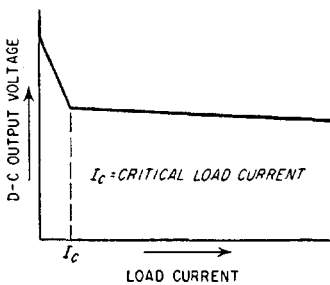
FIG. 15.1. D-c output voltage versus load current for a constant inductance filter input choke.
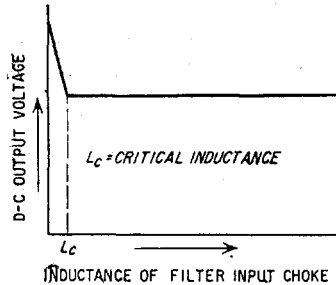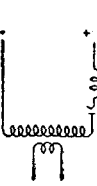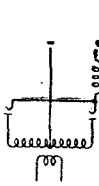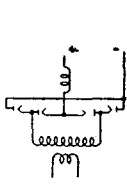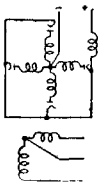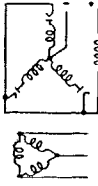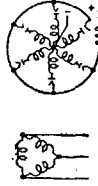
FIG. 15.2. D-c output voltage versus inductance of filter input choke for a constant load current.

current into the filter does not go to zero for any portion of the cycle. A close approximation to this value can be made by assuming that the pulsating component of the current from the rectifier is due entirely to the lowest harmonic present. The errors due to this approximation are small and are amply allowed for by the use of the optimum inductance. *Optimum inductance* is generally defined as twice the critical inductance. For conservative design, general practice is to make the inductance of the input choke equal to the optimum inductance.

For a given power supply there exists a value of critical inductance for every value of load current.[1] Figure 15.1 illustrates the variation in output voltage versus load current for a choke-input power supply. As the load current is reduced, the associated critical inductance becomes equal to the inductance of the input choke at a critical load current I_c. As the current is further decreased the critical inductance becomes higher than that of the input choke. The result is that the input choke no longer filters properly. Frequently a bleeder resistor is added to the power-supply

---

[1] With a half-wave rectifier the value of critical inductance is infinite since there is no value of choke inductance which will ensure that the rectifier current does not go to zero. For this reason and the fact that the volt-ampere rating of the transformer would be much higher than the rating of a transformer operating into a full-wave rectifier and delivering the same d-c output power, choke-input filters are seldom used with half-wave rectifiers.

Tabulated data based on zero tube and transformer resistance and zero transformer leakage inductance	Single-phase half wave	Single-phase full wave	Single-phase full-wave bridge	2-phase full wave	3-phase star half wave	3-phase star full wave	6-phase star half wave
D-C load voltage.........................	*	1.000	1.000	1.000	1.000	1.000	1.000
Secondary rms voltage per leg......................	*	1.111	1.111	0.785	0.855	0.428	0.740
Freq. of lowest harmonic in terms of line freq. (f)..	f	$2f$	$2f$	$4f$	$3f$	$6f$	$6f$
Rms voltage of 1st harmonic......................	*	0.471	0.471	0.0944	0.177	0.0405	0.0405
Rms voltage of 2nd harmonic.....................	*	0.0944	0.0944	0.0224	0.041	0.0099	0.0099
Rms voltage of 3rd harmonic.....................	*	0.0405	0.0405	0.0099	0.018	0.0043	0.0043
Ripple peaks with reference to d-c axis:							
Positive peak..................................	*	0.363	0.363	0.111	0.209	0.0472	0.0472
Negative peak..................................	*	0.637	0.637	0.215	0.395	0.0930	0.0930
Peak inverse tube voltage.......................	*	3.14	1.57	2.22	2.09	1.05	2.09
D-c load current...............................	*	1.000	1.000	1.000	1.000	1.000	1.000
Secondary rms current per leg.....................	*	0.707	1.000	0.500	0.577	0.816	0.408
Peak current per anode / D-C load current	*	1.000	1.000	1.000	1.000	1.000	1.000
D-C output, watts............................	*	1.000	1.000	1.000	1.000	1.000	1.000
Secondary volt-amperes.........................	*	1.57	1.11	1.57	1.48	1.05	1.81
Primary volt-amperes...........................	*	1.11	1.11	1.11	1.21	1.05	1.28
Average of primary and secondary volt-amperes....	*	1.34	1.11	1.34	1.35	1.05	1.55
Line power factor.............................	*	0.90	0.90	0.90	0.826	0.955	0.781

* Values vary with the ratio of load resistance to series inductance. For details see M. B. Stout, Behavior of Half-wave Rectifiers, *Electronics*, September, 1939.

output to increase the minimum load current and reduce the value of the critical inductance, which in turn reduces the required inductance of the input filter choke.

For a given load current, Fig. 15.2 shows how the output voltage increases when the inductance of the input choke is reduced below that of the critical inductance.

Neglecting the transformer and rectifier resistances, the critical inductance can be calculated from Eq. (15.1).

$$L_c = \frac{R}{A} \tag{15.1}$$

where L_c = critical inductance, henrys

R = load resistance (resistance of choke plus bleeder resistance in parallel with external load)

A = a constant determined from Fig. 15.3
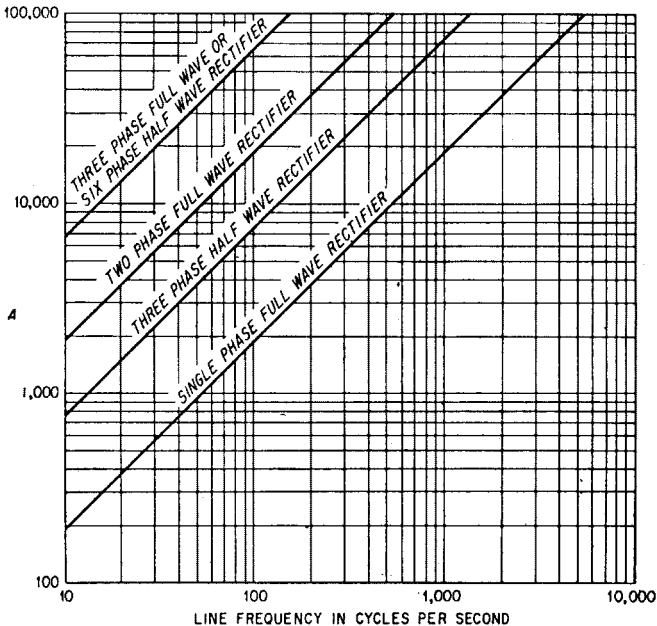
Fig. 15.3. Plot of A versus line frequency for several types of rectifiers.

An input choke whose inductance varies over wide limits as an inverse function of the d-c load current is defined as a swinging choke (see Sec. 14.4g). In a choke-input-filter type of power supply, where the load current varies, a properly designed swinging choke is often used since its inductance will remain greater than the critical inductance at small values of load current.

Example 15.1

Determine the optimum inductance for a single-phase full-wave choke-input power supply operating from a 60-cycle line if the d-c output voltage is 300 volts, the minimum load current is 100 ma, the bleeder is 20,000 ohms, and the choke resistance is 100 ohms.

Solution

1. The external load resistance at the minimum load current is

$$\frac{\text{Output voltage}}{\text{Minimum load current}} = \frac{300}{0.100} = 3,000 \text{ ohms}$$

2. The external load resistance in parallel with the bleeder resistance is equal to

$$\frac{3,000 \times 20,000}{3,000 + 20,000} = 2,610 \text{ ohms}$$

3. The total load resistance is equal to the choke resistance plus the bleeder in parallel with the external load resistance

$$R = 100 + 2,610 = 2,710 \text{ ohms}$$

4. Determine A from Fig. 15.3.
Since the line frequency is 60 cps and the rectifier is single-phase full wave, $A = 1,130$.
5. Determine the critical inductance L_c.

$$L_c = \frac{R}{A} = \frac{2,710}{1,130} = 2.4 \text{ henrys}$$

6. Determine the optimum inductance L_o.

$$L_o = 2L_c = 2 \times 2.4 = 4.8 \text{ henrys}$$

15.3. Power Supplies Having Choke-input Filters. This section contains design information for determining both the d-c output voltage and the per cent ripple voltage in the output of power supplies having single-section choke-input filters.[1] The inductance of the input filter choke L is assumed to be equal to or greater than the critical inductance, and the reactance of the filter capacitor C at the lowest ripple frequency is assumed to be appreciably less than the load resistance.

15.3a. Determination of D-C Output Voltage. Data given in Table 15.1 simplify this determination. The data are based on there being no resistance in the transformers, rectifiers, or choke. In a physical system these components have resistances which must be considered. The product of the d-c output current and the sum of the component resistances is equal to the d-c voltage drop within the power supply. In a practical system the sum of this d-c voltage drop and the d-c output voltage is equal to the d-c load voltage as determined from Table 15.1. For example, in the case of a single-phase full-wave rectifier operating into a choke-input filter, Table 15.1 indicates that the value of one-half the secondary rms voltage is equal to 1.111 times the d-c load voltage. This is based on zero resistance in the transformer, rectifiers, and choke. Therefore, if the terminal voltage of one-half the transformer secondary under load were arbitrarily assumed to be equal to 700 volts, the d-c load voltage from Table 15.1 would be equal to 700/1.111, or 630 volts. The actual d-c output voltage would be equal to 630 volts less the sum of the IR drops in the rectifier tube and filter choke. The transformer IR drop need not be determined since the transformer terminal voltage under load has been specified.[2] Table 15.1 gives the rectifier and transformer operating currents.

---

[1] If additional filter sections of either the LC or RC type are to be added, refer to Sec. 15.5.
[2] The problem becomes more complicated if only the transformer no-load voltage is known. The transformer terminal voltage at full load can be calculated from the following relationship:

$$E_s = \left[(\text{pri. rms voltage} - 1.11 I_p R_p) \left(\frac{\frac{1}{2} \text{ sec. no-load voltage}}{\text{pri. voltage}} \right) \right] - 1.11 R_s I_{dc}$$

where E_s = terminal voltage of $\frac{1}{2}$ secondary
R_s = resistance of $\frac{1}{2}$ secondary
I_p = peak primary current
I_{dc} = d-c load current
R_p = resistance of primary

15.3b. Determination of the Per Cent Ripple. A low-pass filter section for the output of a power-supply rectifier is shown in Fig. 15.4. The per cent ripple voltage[1] in the filter d-c output voltage can be determined from the curves shown in Fig. 15.5 provided the *LC* product, the line frequency, and the type of rectifier are known. It is also possible to solve for the *LC* product necessary to obtain a specified per cent ripple voltage for a given line frequency and type of rectifier.
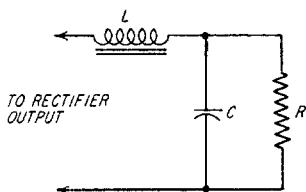
FIG. 15.4. Single-section choke-input filter with load resistance.

The curves shown in Fig. 15.5 are based on the assumption that the only a-c component in the rectifier output voltage is the lowest-frequency harmonic. This introduces a very small error since the filter has a much higher transmission loss for the higher-order harmonics, and, secondly, these higher order harmonics have much smaller amplitudes than the lowest-frequency harmonic.

15.3c. Minimum Size Limitations of the Filter Choke. The value of *L* must be equal to or greater than the critical inductance. It is considered good practice, however,
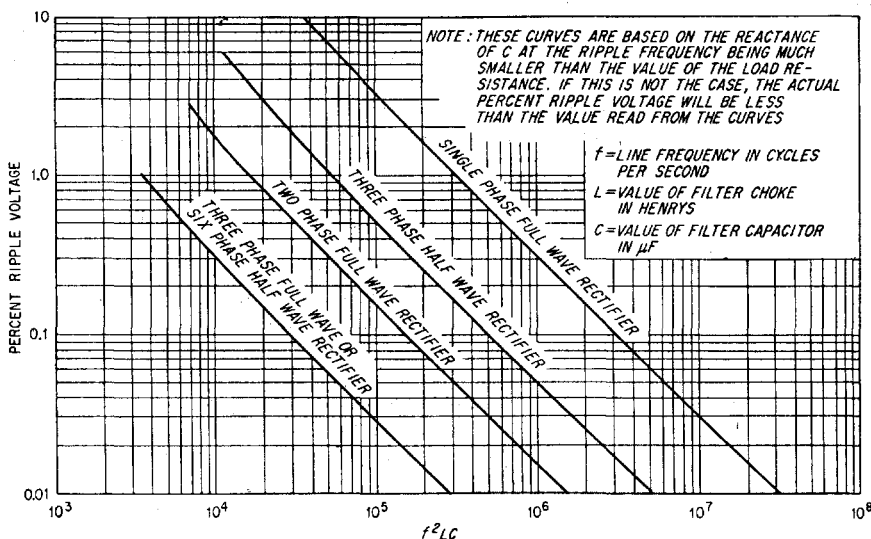
FIG. 15.5. A plot of the per cent ripple voltage versus f^2LC for power supplies having single-section choke-input filters.

to make *L* equal to or greater than the optimum inductance. The optimum inductance is twice the critical inductance as defined in Sec. 15.2.

In unregulated power supplies where very low values of ripple are required, i.e., 50 mv or less, it is suggested that the filter choke or chokes be placed in either the

---

[1] Per cent ripple voltage can be expressed as:

$$\% \text{ ripple} = \frac{E_{ac}}{E_{dc}} \times 100$$

where E_{ac} = rms value of ripple voltage
E_{dc} = d-c value of output voltage

positive or negative lead depending on whether the supply is a positive or negative voltage power supply, respectively.[1] If this practice is not observed, the transformer secondary capacitance to ground will shunt the input choke.

15.3d. Minimum Size Limitations of C. There are two limiting conditions for the minimum size of C. They are:

1. The curves shown in Fig. 15.5 are based on the value of C being sufficiently large that its reactance at the lowest ripple frequency f_1 will be much less than the resistance of the load R. Equation (15.2) is based on the reactance of C at f_1 being no greater than one-fifth the value of the minimum load resistance.

$$C \geq \frac{796,000}{f_1 R_{\min}} \tag{15.2}$$

where $R_{\min}$ = minimum d-c load resistance, ohms
$\quad f_1$ = frequency in cps of lowest harmonic from the rectifier (see Table 15.1 to determine the frequency of the lowest harmonic)
$\quad C$ = filter capacitor, μf

2. The requirements of the load frequently dictate that the power supply shall have an a-c output impedance less than some prescribed value at a specified frequency. The actual value of the power-supply a-c output impedance is sometimes difficult to determine. The output capacitor, as viewed by the load, is in parallel with the filter choke and rectifier components, and at some frequency they will form a resonant circuit. Therefore, when this type of filter is employed it is possible that the power-supply output impedance will be higher than the reactance of the output capacitor at and near the filter resonant frequency. At frequencies much higher than the resonant frequency, the output impedance will be essentially equal to the reactance of the output capacitor. At frequencies much lower than the resonant frequency, the a-c output impedance will approach the d-c output resistance determined by the choke, rectifier, and transformer resistances. For conservative design the choke and capacitor values should be large enough that their resonant frequency will be far below the lowest frequency of load-current variations. Such a precaution will ensure that the power-supply a-c output impedance will be equal to or less than the reactance of the output capacitor. Therefore, to satisfy a requirement for not exceeding a specified maximum output impedance at some frequency far removed from the filter resonant frequency, the value of the output capacitor C must satisfy Eq. (15.3).

$$C \geq \frac{159,000}{f_2 Z_o} \tag{15.3}$$

where Z_o = specified maximum a-c output impedance
$\quad f_2$ = frequency at which Z_o is specified (usually lowest frequency of load current variations)
$\quad C$ = output capacitor, μf

Example 15.2

Determine the d-c output voltage for a power supply with the following characteristics: Single-phase full-wave bridge rectifier:

Voltage drop across each rectifier............................... 15 volts
Voltage drop across filter choke................................ 12 volts
Terminal voltage of transformer secondary under full load........ 1,450 volts (rms)

Solution

1. Determine the d-c output voltage on the basis that there is no voltage drop across the rectifier tubes and filter choke.

---

[1] See *Proc. IRE*, vol. 22, p. 1040, 1934.

From Table 15.1

$$\text{Sec. rms voltage} = 1.111 \text{ (d-c output voltage)}$$
$$\text{D-C output voltage} = 0.9 \text{ (sec. rms voltage)}$$
$$= 0.9 \times 1,450 = 1,305 \text{ volts}$$

2. Determine the power-supply d-c output voltage.

$$\begin{pmatrix} \text{Power-supply d-c} \\ \text{output voltage} \end{pmatrix} = 1,305 - \begin{pmatrix} IR \text{ drop in} \\ 2 \text{ rectifiers} \end{pmatrix} - \begin{pmatrix} IR \text{ drop in} \\ \text{choke} \end{pmatrix}$$
$$= 1,305 - (2 \times 15) - (12)$$
$$= 1,263 \text{ volts}$$

NOTE: In a single-phase full-wave bridge rectifier the d-c load current flows through two rectifiers simultaneously.

Example 15.3

Determine the values of L and C which will permit no more than 0.4 per cent ripple voltage in the d-c output voltage of a three-phase half-wave 60-cycle rectifier, having an output voltage of 3,000 and a load current which varies between 1 and 2 amp. Assume that the characteristics of the load are such that the power-supply output impedance is unimportant.

Solution

1. Determine the maximum and minimum load resistances.

$$R_{\max} = \frac{\text{d-c output voltage}}{\text{min load current}} = \frac{3,000}{1} = 3,000 \text{ ohms}$$
$$R_{\min} = \frac{\text{d-c output voltage}}{\text{max load current}} = \frac{3,000}{2} = 1,500 \text{ ohms}$$

2. Determine the minimum possible value of L [refer to Eq. (15.1) and Fig. 15.3].

$$L_c = \frac{R_{\max}}{A} = \frac{3,000}{4,500} = 0.667 \text{ henry}$$

It is recommended that the inductance be made equal to or greater than the optimum inductance, i.e., $2L_c$. Therefore,

$$L_o = 2 \times 0.667 = 1.33 \text{ henrys}$$

3. Determine the minimum possible value of C.
From Eq. (15.2)

$$C \geq \frac{796,000}{f_1 R_{\min}} = \frac{796,000}{180 \times 1,500} = 2.95 \ \mu\text{f}$$

where $f_1 = 3 \times$ line freq. (see Table 15.1)

4. Determine the product of L and C which will allow no more than 0.4 per cent ripple voltage in the d-c output voltage.

From Fig. 15.5 it can be seen that $f^2 LC$ must be equal to or greater than 125,000. Since $f = 60$, f^2 is equal to 3,600. The product of L and C must therefore be equal to or greater than 35.

5. Determine the values of L and C.

If L were arbitrarily made equal to 1.33 henrys, then C must be equal to or greater than $35/1.33$, or 26.3 μf. Considerations of space, weight, and cost can dictate a different ratio of L and C. Possibly a more economical choice might be to let C be equal to 6 μf, and then L would have to be equal to or greater than $35\!\!\not6$, or 5.83 henrys. It should be noted that both choices of L and C satisfy the limiting values determined in steps 2 and 3.

15.4. Power Supplies Having Capacitor-input Filters. Half-wave, full-wave, and voltage doubler types of single-phase[1] rectifiers utilizing single-section capacitor-

---

[1] Multiphase rectifiers are ordinarily used when the power requirements are heavy, and in such instances choke-input filters and gas-filled rectifier tubes usually are employed. The use of gas-filled rectifiers normally prohibits the use of capacitor-input filters since the peak capacitor charging currents may be very high and, consequently, require the use of excessively large rectifiers. For these reasons multiphase rectifiers are not discussed in this section.

input filters[1] are discussed in this section. It is possible to determine accurately the per cent ripple in the d-c output voltage and the relationship between the following: transformer secondary rms voltage and the filter d-c output voltage, average and peak rectifier currents, and the average and rms rectifier currents.

A single-section capacitor-input filter for the output of a power-supply rectifier is shown in Fig. 15.6. The information in Figs. 15.7 to 15.12 permits a very thorough analysis of half-wave, full-wave, and voltage doubler types of rectifiers utilizing capacitor-input filters. Figures 15.8 to 15.12 were derived by O. H. Schade.[2] The curves are self-explanatory;

FIG. 15.6. Single-section capacitor-input filter with load resistance.

however, a few terms should be reviewed. The source resistance R_s in a power supply having a capacitor-input filter is the sum of the resistances through which the capacitor charging current must flow, i.e., the rectifier resistance, any current-limiting
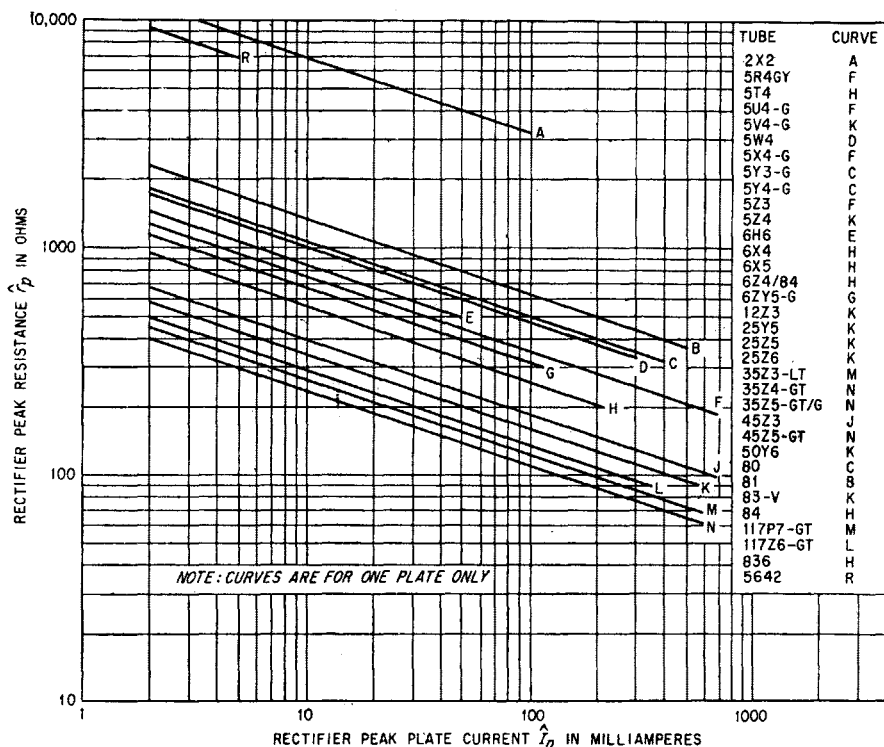
FIG. 15.7. Average characteristics of several different rectifiers.

TUBE	CURVE
2X2	A
5R4GY	F
5T4	H
5U4-G	F
5V4-G	K
5W4	D
5X4-G	F
5Y3-G	C
5Y4-G	C
5Z3	F
5Z4	K
6H6	E
6X4	H
6X5	H
6Z4/84	H
6ZY5-G	G
12Z3	K
25Y5	K
25Z5	K
25Z6	K
35Z3-LT	M
35Z4-GT	N
35Z5-GT/G	N
45Z3	J
45Z5-GT	N
50Y6	K
80	C
81	B
83-V	K
84	H
117P7-GT	M
117Z6-GT	L
836	H
5642	R

resistances, the transformer secondary resistance, and the transformer primary resistance referred to the secondary. The curves show that changes in the source resistance materially affect the over-all circuit performance. Since this is the case, it is necessary to take into consideration the changing value of the rectifier resistance

[1] See Sec. 15.5 if additional filter sections of either the *LC* or *RC* type are to be added.
[2] O. H. Schade. Analysis of Rectifier Operation, *Proc. IRE*, vol. 31, pp. 341–361, July, 1943.
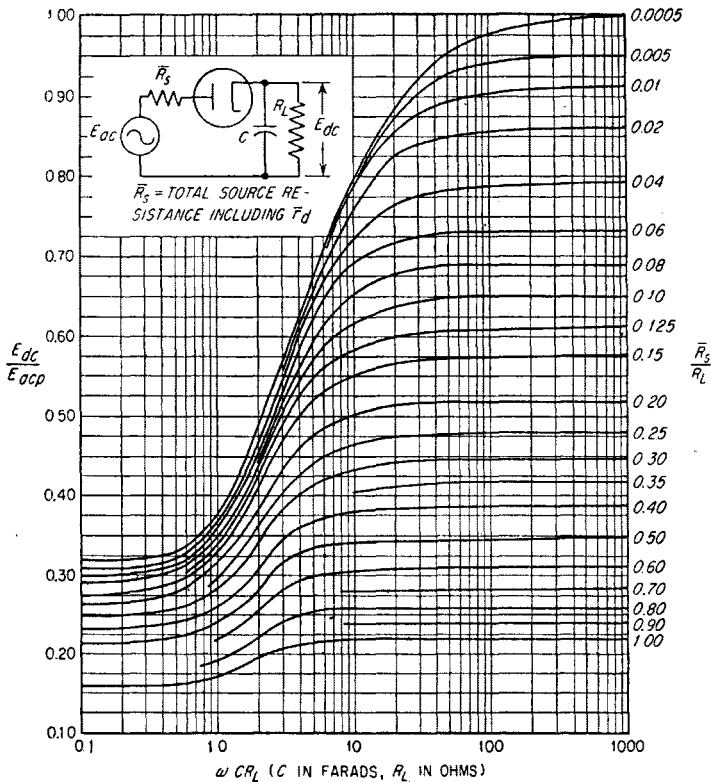
Fig. 15.8. Relation of applied alternating peak voltage to direct output voltage in half-wave capacitor-input circuits. (*From O. H. Schade, Proc. IRE, July, 1943, p. 343.*)
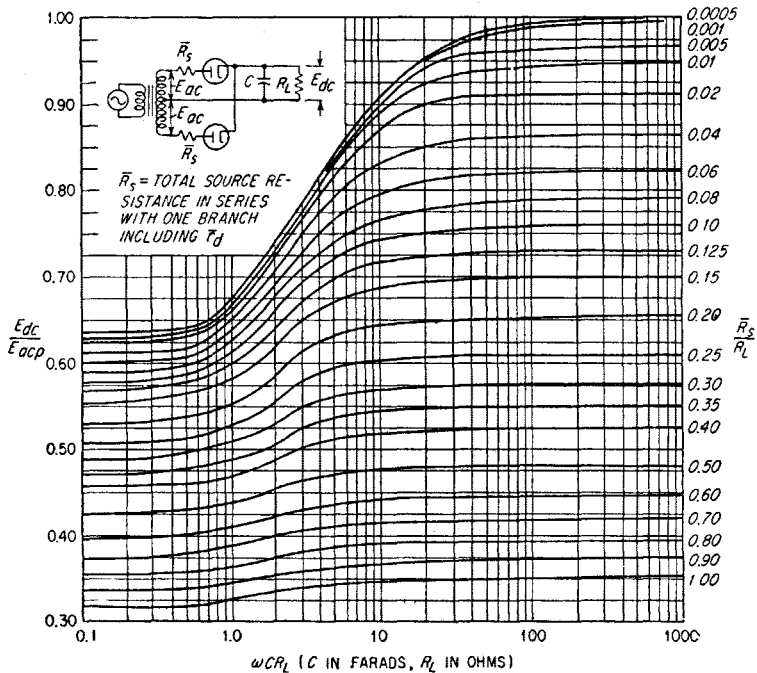
Fig. 15.9. Relation of applied alternating peak voltage to direct output voltage in full-wave capacitor-input circuits. (*From O. H. Schade, Proc. IRE, July, 1943, p. 344.*)
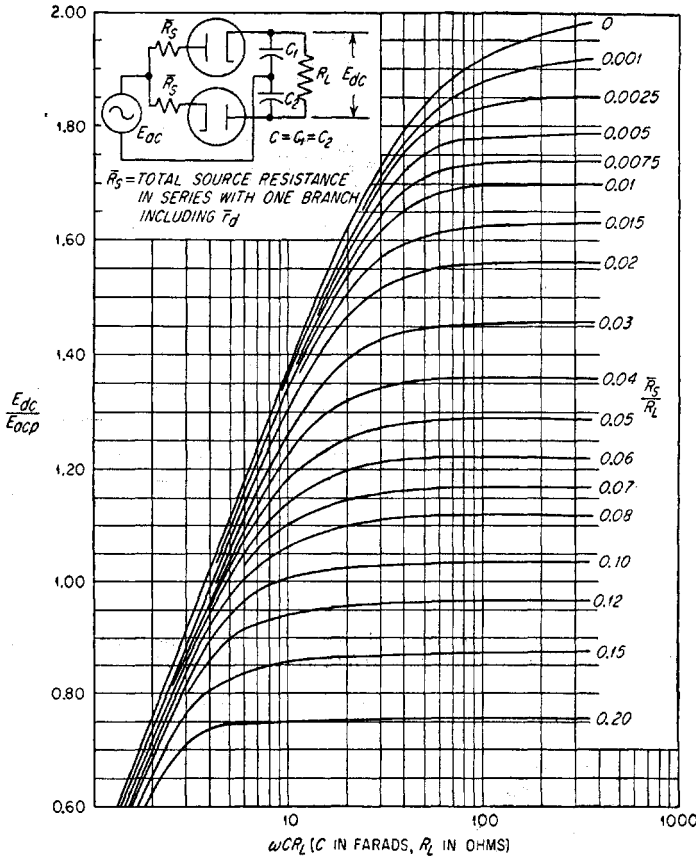
FIG. 15.10. Relation of applied alternating peak voltage to direct output voltage in capacitor-input voltage-doubling circuits. (*From O. H. Schade, Proc. IRE, July, 1943, p. 345.*)

during the conduction period. The rectifier resistances whch must be known are the peak, rms, and average values. Since the source resistance is made up in part by the rectifier resistance, there are peak, rms, and average values of the source resistance. The different rectifier resistances are defined as follows:

$\hat{r}$ = *rectifier peak resistance,* the resistance at the peak rectifier current $\hat{I}_p$

$|r|$ = *rectifier rms resistance,* the resistance which when multiplied by the square of the rectifier rms current $|I_p|$ will equal the plate power dissipated within the rectifier

$\bar{r}$ = *rectifier average resistance,* the resistance which when multiplied by the average rectifier current $\bar{I}_p$ during conduction will equal the average drop across the rectifier during conduction

Schade has shown that the relationships

$$1.14\hat{r} = \bar{r} \tag{15.4}$$
$$1.08\hat{r} = |r| \tag{15.5}$$

are representative within ±5 per cent for capacitor-input circuits containing high-vacuum rectifiers. The peak rectifier resistance $\hat{r}$ is determined from Fig. 15.7, and from this value the rms and average values are calculated.

$n\omega CR_L$

$\dfrac{|I_p|}{I_p} = \dfrac{\text{RMS PLATE CURRENT}}{\text{D-C PLATE CURRENT}}$ (ONE PLATE)

10

1.584

1

0.0002
0.0005
0.001
0.002
0.005
0.01
0.02
0.05
0.10
0.30
1.00

$\dfrac{|R_s|}{nR_L}$

80

$n = \begin{cases} 1 \text{ FOR HALF-WAVE SINGLE-PHASE RECTIFIER CIRCUITS} \\ 2 \text{ FOR FULL-WAVE SINGLE-PHASE RECTIFIER CIRCUITS} \\ \tfrac{1}{2} \text{ FOR VOLTAGE-DOUBLING CIRCUITS} \end{cases}$

C IN FARADS
R_L IN OHMS
$\omega = 2\pi f$ WHERE $f =$ LINE FREQUENCY
$|R_s| =$ RMS EQUIVALENT

$\dfrac{|I_p|}{I_p} = \dfrac{\text{PEAK PLATE CURRENT}}{\text{D-C PLATE CURRENT}}$ (ONE PLATE)

10

3.141

1.0

0.0002
0.0005
0.001
0.002
0.005
0.01
0.02
0.05
0.10
0.30
1.00

$\dfrac{\hat{R_s}}{nR_L}$

$n\omega CR_L$

FIG. 15.11. Relation of peak, average, and root-mean-square diode current in capacitor-input circuits. (*From O. H. Schade, Proc. IRE, July, 1943, p. 346.*)

100

CIRCUIT   PARAMETER

$\dfrac{|R_s|}{R_L}$

HALF-WAVE
0.001
0.010
0.10
0.30

VOLTAGE-DOUBLER
0.001
0.10

FULL-WAVE
0.001
0.010
0.10
0.30

$\% \dfrac{\text{RIPPLE VOLTS}}{E_{dc}}$

10

1.0

0.1

ωCR_L (C IN FARADS, R_L IN OHMS)
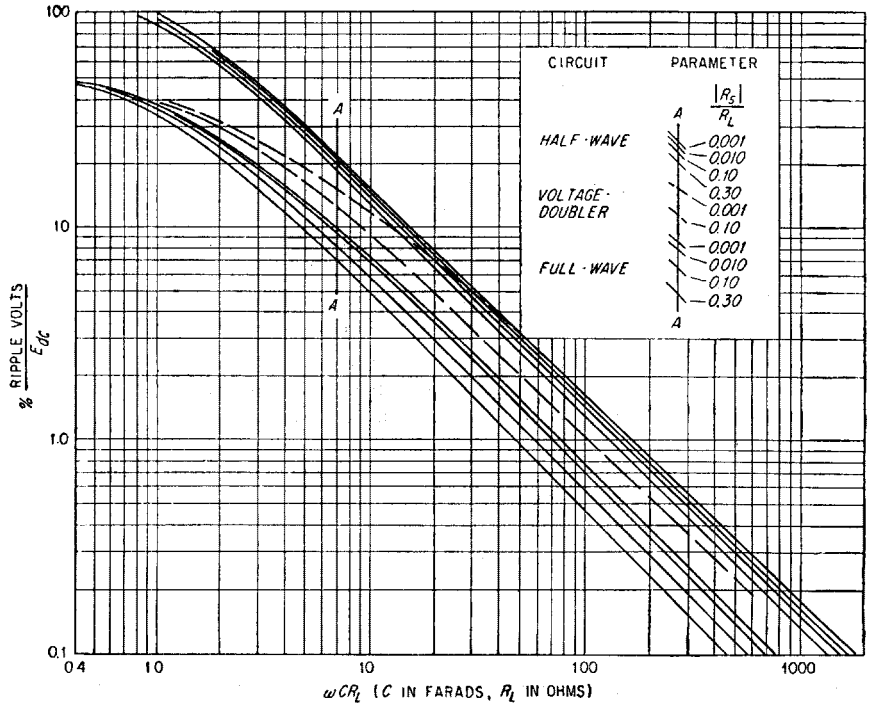
FIG. 15.12. Root-mean-square ripple voltage of capacitor-input circuits. (*From O. H. Schade, Proc. IRE, July, 1943, p. 347.*)

The per cent ripple voltage is defined as the output rms ripple voltage expressed as a percentage of the d-c output voltage.

In a given power supply it is important to note that, to obtain the best possible regulation, the value of $\omega C R_L$ should be to the right of the knee of the appropriate curve in Figs. 15.8 to 15.10.

Example 15.4

In the power supply shown in Fig. 15.13, determine the average, peak, and rms currents per rectifier plate; the d-c output voltage; and the per cent ripple in the d-c output voltage. The circuit constants are as follows:

1. The transformer primary resistance is 4 ohms.

2. The total transformer secondary resistance is 500 ohms.

3. The transformer total secondary no-load voltage rating is 1,500 volts rms with 120 volts rms on the primary.

4. The filter capacitor C is 4 μf.

5. The supply frequency is 50 cps.

6. The d-c load resistance R_L is 3,000 ohms.
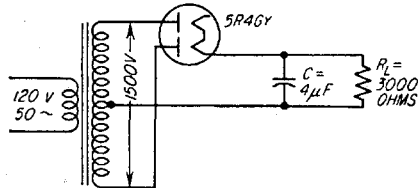
FIG. 15-13. Single-phase full-wave rectifier with a capacitor-input filter.

Solution

It is necessary to know the peak resistance of each rectifier plate. The value of the rectifier peak resistance is obtained from Fig. 15.7 through the knowledge of the rectifier peak current. Since the rectifier peak current is not yet known, it is necessary to make an assumption as to the value of the rectifier peak resistance. On the basis of the assumed peak resistance, the associated rectifier rms and average resistances can be calculated from Eqs. (15.4) and (15.5). Using these tentative values of resistance, the circuit analysis can be completed. The resulting analysis includes the determination of a value for the rectifier peak current, and from Fig. 15.7 the associated rectifier peak resistance can be determined. If the value of the rectifier peak resistance taken from Fig. 15.7 is reasonably close to the assumed value, the analysis can be considered valid. If the value from Fig. 15.7 differs greatly from the assumed value, it is necessary to use the value established from Fig. 15.7 and repeat the circuit analysis, thereby redetermining the rectifier peak resistance. It is relatively certain that this new value of rectifier peak resistance will be reasonably close to the value of rectifier peak resistance obtained in the first circuit analysis.

1. Assume a value for the rectifier peak resistance.

Assume that the 5R4GY peak resistance $\hat{r}$ is equal to 200 ohms per plate (see Fig. 15.7).

2. Determine the rectifier rms resistance $|r|$ and the rectifier average resistance $\bar{r}$ from Eqs. (15.4) and (15.5).

$$|r| = 1.08\hat{r} = 1.08 \times 200 = 216 \text{ ohms}$$
$$\bar{r} = 1.14\hat{r} = 1.14 \times 200 = 228 \text{ ohms}$$

3. Determine the peak $\hat{R}_s$, rms $|R_s|$, and average $\bar{R}_s$ source resistances.

R_s = pri. resistance referred to $\frac{1}{2}$ sec.* + $\dfrac{\text{total sec. resistance}}{2}$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ + resistance of one plate of rectifier

$\quad = \left[\text{pri. resistance} \times \left(\dfrac{\frac{1}{2} \text{ sec. volt.}}{\text{pri. volt.}} \right)^2 \right] + \dfrac{500}{2}$ + resistance of one plate of rectifier

$\quad = [(4) \times (^{750}\!/_{120})^2] + 250$ + resistance of one plate of rectifier

$\quad = 406$ + resistance of one plate of rectifier

Therefore

$$\hat{R}_s = 406 + \hat{r} = 406 + 200 = 606 \text{ ohms}$$
$$|R_s| = 406 + |r| = 406 + 216 = 622 \text{ ohms}$$
$$\bar{R}_s = 406 + \bar{r} = 406 + 228 = 634 \text{ ohms}$$

---

* The resistance of the prime voltage source, if appreciable, should be added to the primary resistance.

4. Determine the values of $\hat{R}_s/nR_L$, $|R_s|/nR_L$, $|R_s|/R_L$ and $\bar{R}_s/R_L$ where $n = 2$ for single-phase full-wave rectifiers. See Fig. 15.11 for values of n when other types of rectifier circuits are used.

$$\frac{\hat{R}_s}{nR_L} = \frac{606}{2 \times 3{,}000} = 0.101$$

$$\frac{|R_s|}{nR_L} = \frac{622}{2 \times 3{,}000} = 0.104$$

$$\frac{|R_s|}{R_L} = \frac{622}{3{,}000} = 0.207$$

$$\frac{\bar{R}_s}{R_L} = \frac{634}{3{,}000} = 0.211$$

5. Determine ωCR_L and $n\omega CR_L$ (where $\omega = 2\pi f$, f = line frequency, C is in farads, and R_L is in ohms).

$$\omega CR_L = 2\pi f CR_L$$
$$= 3.77$$
$$n\omega CR_L = 2 \times 3.77 = 7.54$$

6. Determine from Fig. 15.9 the value of E_{dc}/E_{acp} where E_{acp} is the peak value of one-half of the secondary rms voltage E_{ac}.
Since $\omega CR_L = 3.77$ and $\bar{R}_s/R_L = 0.211$, interpolation from Fig. 15.9 shows that E_{dc}/E_{acp} is approximately equal to 0.615.

7. Determine E_{dc}.
From step 6, E_{dc} was found to be $0.615 \times E_{acp}$.
Therefore,

$$E_{dc} = 0.615 \times E_{acp}$$
$$= 0.615 \times 1.414 \times E_{ac}$$
$$= 0.615 \times 1.414 \times 750$$
$$= 652 \text{ volts}$$

8. Determine the load current I_L and the average current per rectifier plate $\bar{I}_p$.

$$I_L = \frac{\text{d-c output voltage}}{\text{load resistance}} = \frac{652}{3{,}000} = 0.217 \text{ amp}$$

$$\bar{I}_p = \frac{I_L}{\text{number of rectifier plates}} = \frac{0.217}{2} = 0.109 \text{ amp}$$

9. Determine the value of $|I_p|/\bar{I}_p$ and $|I_p|$, where $|I_p|$ is the rectifier rms plate current (see Fig. 15.11).

$$\frac{|R_s|}{nR_L} = 0.104 \qquad \text{and} \qquad n\omega CR_L = 7.54$$

as determined in steps 4 and 5, respectively. From Fig. 15.11

$$\frac{|I_p|}{\bar{I}_p} = 2.0 \qquad \text{and} \qquad |I_p| = 2\bar{I}_p = 0.218 \text{ amp}$$

10. Determine the value of $\hat{I}_p/\bar{I}_p$ and $\hat{I}_p$, where $\hat{I}_p$ is the rectifier peak plate current (see Fig. 15.11).

$$\frac{\hat{R}_s}{nR_L} = 0.101 \qquad \text{and} \qquad n\omega CR_L = 7.54$$

as determined in steps 4 and 5. From Fig. 15.11,

$$\frac{\hat{I}_p}{\bar{I}_p} = 5.1 \qquad \text{and} \qquad \hat{I}_p = 5.1\bar{I}_p = 0.556 \text{ amp}$$

11. Determine the per cent ripple voltage in the d-c output voltage (see Fig. 15.12).
Since $\omega CR_L = 3.77$, $|R_s|/R_L = 0.207$, and the rectifier circuit is full-wave, it can be determined from Fig. 15.12 that the ripple voltage is approximately 13.5 per cent.

Summary

In step 10 the rectifier peak plate current was found to be 556 ma. Figure 15.7 shows that the rectifier peak resistance is approximately 200 ohms. This value agrees with the original assumed value and hence validates the complete analysis. If the two values were appreciably different, it would be necessary to repeat the analysis using the rectifier peak resistance value obtained from Fig. 15.7 based on the rectifier peak current obtained in step 10.

Example 15.5

Design a 60-cycle power supply utilizing a half-wave rectifier and a capacitor-input filter which will have a d-c output voltage of 500 volts at 75 ma and an output ripple voltage which is equal to or less than 2 per cent of the d-c output voltage (Fig. 15.14).

Solution

As in Example 15.4, the rectifier peak resistance must be known. This value is ordinarily obtained from Fig. 15.7 through knowledge of the rectifier peak current. In this particular example the rectifier peak current is not known, but the rectifier average current is given in the statement of the problem. O. H. Schade[1] has suggested that in such instances the rectifier

FIG. 15.14. Power supply with a half-wave rectifier and capacitor-input filter.

peak current be assumed to be four times the rectifier average current. On the basis of the rectifier peak resistance value obtained from the assumed peak current, the associated rectifier rms and average resistances can be calculated from Eqs. (15.4) and (15.5). Assuming that these calculated values are correct, the analysis can be completed. This includes the determination of a value for the rectifier peak current, and from Fig. 15.7 the associated rectifier peak resistance can be determined. The rectifier peak resistance obtained should be compared to the assumed value. If the two values differ greatly, it is necessary to use the value from Fig. 15.7 and repeat the analysis. It is reasonably certain that this new value of the rectifier peak resistance will closely agree with the value obtained in the initial analysis.

1. Estimate the rectifier peak current $\hat{I}_p$ and determine the associated value of peak resistance $\hat{r}$ from Fig. 15.7.

Let $\hat{I}_p = 4 \times \bar{I}_p = 4 \times 0.075 = 0.300$ amp, or 0.150 amp per plate. From Fig. 15.7, the peak resistance per plate is established as being approximately 300 ohms, and, for two plates in parallel, approximately 150 ohms.

2. Determine the rms resistance $|r|$ and the average resistance $\bar{r}$ from Eqs. (15.4) and (15.5).

$$\bar{r} = 1.14\hat{r} = 1.14 \times 150 = 171 \text{ ohms}$$
$$|r| = 1.08\hat{r} = 1.08 \times 150 = 162 \text{ ohms}$$

3. Determine the load resistance R_L.

$$R_L = \frac{\text{d-c output voltage}}{\text{d-c output current}} = \frac{500}{0.075} = 6{,}670 \text{ ohms}$$

4. Determine the values of the source resistance R_s.

There are an infinite number of combinations of E_{acp} and R_s which will give the specified d-c output voltage and current. The actual value of R_s is somewhat difficult to determine since the resistances of the transformer windings are not known.

The effective transformer resistance which is in series with the capacitor charging current through the rectifier is termed R_t. The sum of R_t and the rectifier resistance r_d is the source resistance R_s. In the case of a half-wave rectifier, the effective transformer resistance is equal to the total secondary resistance plus the primary resistance referred to the secondary, i.e., secondary resistance $+ R_p N_s^2/N_p^2$. It follows that the rms source resistance $|R_s|$ is equal to the rectifier rms resistance $|r|$ plus R_t.

For every value of R_t, it is possible to calculate the associated values of the rms rectifier current $|I_p|$, the transformer no-load peak voltage E_{acp}, and the copper loss in the transformer $|I_p|^2 R_t$. In a practical design problem, values of R_t are usually assumed and the other associated values are then determined. Figure 15.15 can be useful in indicating
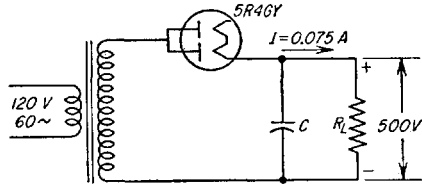
---

[1] *Loc. cit.*

when practical values of R_t have been assumed since the calculated values of $|I_p|^2 R_t$, i.e., total transformer copper loss,[1] can be compared to the total copper loss of a typical transformer for a power supply having a capacitor-input filter. The final step is to choose a transformer whose ratings for E_{acp}, R_t, and secondary rms current are compatible with the ratings associated with one of the assumed values of R_t. It is possible to utilize a transformer having an excessive value of E_{acp} for the associated value of R_t provided resistance is added in the lead between the filter capacitor and the transformer secondary to increase the resistance in the charging path for the capacitor.
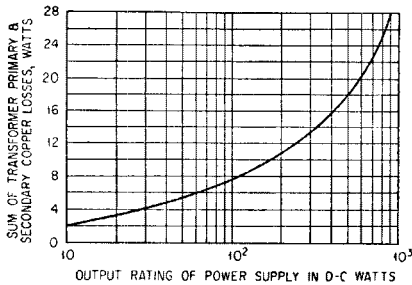
FIG. 15.15. Plot of power-supply d-c output wattage versus power transformer copper losses for a typical power supply with a capacitor-input filter.

If, in this example, R_t is assumed equal to 130 ohms, the value of $|R_s|$ will be 130 + 162, or 292 ohms. For best regulation, $n\omega C R_L$ should be made sufficiently large to ensure operation above the knee of the curve in Fig. 15.11. In this example $|R_s|/R_L$ will be equal to 0.044, therefore the value of $|I_p|/\bar{I}_p$ as determined from Fig. 15.11 is approximately equal to 2.5, which makes $|I_p|$ approximately equal to 0.188 amp. The total transformer copper loss is equal to $|I_p|^2 R_t$, or 0.188^2 × 130, or 4.6 watts. With reference to Fig. 15.15, it can be seen that the assumed value of R_t is probably typical since the calculated copper loss is typical of transformers used in power supplies with capacitor-input filters and having a d-c output rating of 37.5 watts, i.e., 500 × 0.075.

From the preceding discussion, the following values are used in the remaining steps:

$$R_t = 130$$
$$|R_s| = 292$$
$$\frac{|R_s|}{R_L} = 0.044$$

5. Determine the average and peak values of the source resistance.

$$\bar{R}_s = R_t + \bar{r} = 130 + 171 = 301 \text{ ohms}$$
$$\hat{R}_s = R_t + \hat{r} = 130 + 150 = 280 \text{ ohms}$$

6. Determine the value of $\bar{R}_s/R_L$.

$$\frac{\bar{R}_s}{R_L} = \frac{301}{6,670} = 0.045$$

7. Determine the value of E_{dc}/E_{acp} from Fig. 15.8 on the basis that $\omega C R_L$ is large enough to ensure operation on the flat part of the curve for best regulation.

$$\frac{E_{dc}}{E_{acp}} = 0.78$$

8. Determine the secondary rms voltage E_{ac}.

$$E_{dc} = 0.78\, E_{acp}$$
$$500 = 0.78 \times \sqrt{2} \times E_{ac}$$
$$E_{ac} = 453 \text{ volts}$$

9. Determine $\hat{R}_s/nR_L$, $|R_s|/nR_L$, and $|R_s|/R_L$.

$$n = 1 \qquad \text{(from Fig. 15.11)}$$
$$\frac{\hat{R}_s}{nR_L} = \frac{280}{1 \times 6,670} = 0.042$$
$$\frac{|R_s|}{nR_L} = \frac{292}{1 \times 6,670} = 0.044$$
$$\frac{|R_s|}{R_L} = 0.044 \qquad \text{(from step 4)}$$

---

[1] In the case of either the voltage doubler or full-wave rectifier, the total transformer copper losses are equal to $2|I_p|^2 R_t$.

10. Determine the value of C necessary to obtain 2 per cent or less ripple.

From step 4, the value of $|R_s|/R_L$ was found to be 0.044. From Fig. 15.12 note that $\omega C R_L$ must be equal to or greater than 74 for the ripple to be equal to or less than 2 per cent. Therefore,

$$\omega C R_L = 74 \qquad \text{(where } \omega = 2\pi f \text{ and } f = \text{line frequency)}$$

$$C = \frac{74}{\omega R_L}$$

$$= 29.5 \times 10^{-6} \text{ farads, or } 29.5 \ \mu f$$

NOTE: In steps 4 and 7 it was assumed that $\omega C R_L$ was to be large enough to permit operation on the flat portions of the curves shown in Figs. 15.8 and 15.11. Since $\omega C R_L = 74$, the assumption is satisfied. If $\omega C R_L$ had not been large enough to permit operation on the flat portions of the curves, the value of C could be increased so that $\omega C R_L$ would be sufficiently large. This would result in a further decrease in the ripple voltage.

11. Determine the values of $\hat{I}_p/\bar{I}_p$ and $\hat{I}_p$.

$$\frac{\hat{R}_s}{n R_L} = 0.042 \qquad \text{(from step 9)}$$

and

$$n \omega C R_L = 74 \qquad \text{(from step 10)}$$

Therefore from Fig. 15.11

$$\frac{\hat{I}_p}{\bar{I}_p} = 6.8$$

and

$$\hat{I}_p = 6.8 \times 0.075 = 0.51 \text{ amp}$$

NOTE: $\hat{I}_p$ should be checked with the rectifier tube peak current rating.

Summary

In step 11, $\hat{I}_p$ was found to be 510 ma, or 255 ma per rectifier plate. Referring to Fig. 15.7 it can be seen that the rectifier peak resistance $\hat{r}$ for the two plates in parallel is approximately 125 ohms. This value is reasonably close to the original estimate, hence validates this portion of the analysis.

The results of steps 4 and 8 specify that the transformer should have the following rating:

Sec. rms voltage = 453 volts
Sec. rms current = 188 ma
Sum of sec. resistance and pri. resistance referred to sec. = 130 ohms

If the transformer contributes appreciably less than 130 ohms to the source resistance and the result is too high a d-c output voltage, a satisfactory solution is to insert the necessary resistance in series with the transformer secondary winding. If the transformer contributes appreciably more than 130 ohms to the source resistance and the result is too low a d-c output, it is necessary to increase the transformer secondary voltage.

15.5. Power-supply LC and RC Low-pass Filters. Sections 15.3 and 15.4 permit the calculation of the d-c output voltage and the per cent ripple voltage for power
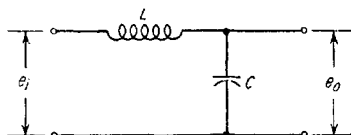
FIG. 15.16. LC low-pass filter section.   FIG. 15.17. RC low-pass filter section.
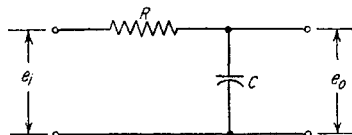
supplies having single-section choke-input or capacitor-input filters. The material contained herein is applicable to the addition of one or more LC or RC filter sections of the types shown in Figs. 15.16 and 15.17 and enables the determination of the increased ripple attenuation.

Ordinarily one or more filter sections are employed to reduce the ripple in the output voltage of a power supply. If the total values of L and C in cascaded LC filter sections are held constant or if the total values of R and C in cascaded RC filter sections are held constant, maximum attenuation is achieved when all sections have identical

(WHERE L IS IN HENRYS, C IS IN μF, AND f IS RIPPLE FREQUENCY IN CYCLES PER SECOND)
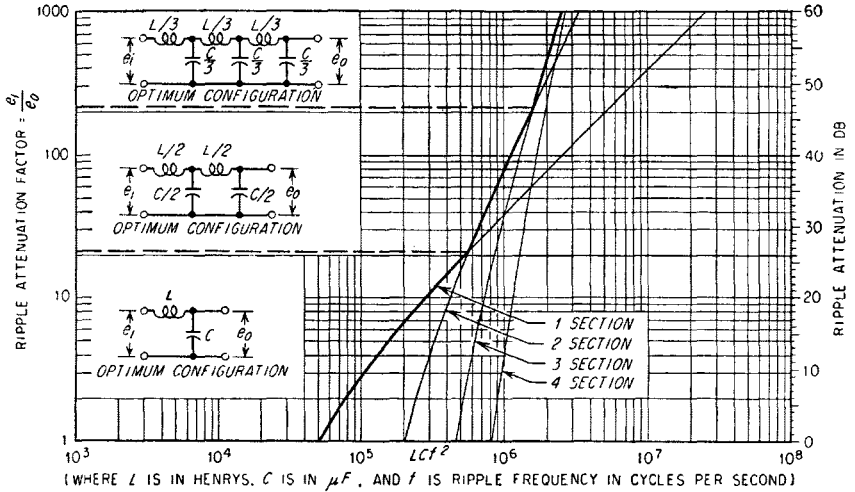
FIG. 15.18. Plot of ripple attenuation versus LCf^2 for several LC filter configurations.

(WHERE R IS IN THOUSANDS OF OHMS, C IS IN μF, AND f IS RIPPLE FREQUENCY IN CYCLES PER SECOND)
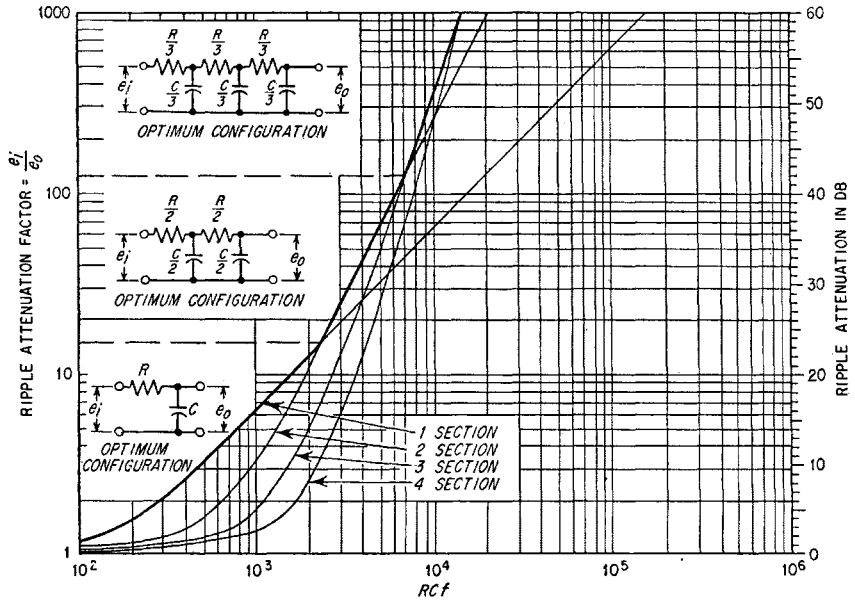
FIG. 15.19. Plot of ripple attenuation versus RCf for several RC filter configurations.

component values. In addition there is an optimum number of these identical filter sections for maximum attenuation based on the product of L, C, and f^2 in LC filters and the product of R, C, and f in RC filters. The values of C, L, and R represent the sums of the values in the cascaded sections, and f is the frequency of the ripple voltage. The optimum number of filter sections can be established from Figs. 15.18 and 15.19.

RC filters are frequently used when the power requirements are low and the regulation requirements are not stringent. LC filters are usually employed when the power requirements are high and when lower voltage drops and better regulation are desired.

The output impedance of unregulated power supplies having LC or RC filter sections varies as a function of frequency; hence, precautions should be taken to ensure that the output impedance is not excessive at the load current frequency.

The output capacitor in cascaded LC filter sections, as viewed from the load, is in parallel with other filter-section elements and at one or more frequencies will resonate with those elements. Therefore, the power-supply output impedance can be expected to be considerably higher than the reactance of the output capacitor at certain frequencies because of resonance within the filter. This may be of prime importance when the load is an amplifier since such applications often require that the power-supply output impedance be very low over a wide band of frequencies. The actual determination of the output impedance of LC filters can be quite complicated. Standard practice is to employ the proper values of L and C for ripple reduction and an output capacitor whose reactance at the lowest load-current frequency is below the maximum output impedance allowable for the power supply. In those few instances where the output impedance is excessive because of resonance within the filter, it is necessary to further increase the value of the output capacitor.

If RC filters are used, the power-supply output impedance will be equal to or lower than the reactance of the output capacitor since filter resonance is nonexistent.

Figures 15.18 and 15.19 contain the data necessary to determine the ripple attenuation in LC and RC filters and in addition indicate the optimum number of sections as a function of total attenuation. These curves are based on the shunt capacitor C having a value larger than any one of the three critical values established in Eqs. (15.6), (15.7), and (15.8).

First and Intermediate Filter Sections. Figures 15.18 and 15.19 are based on the assumption that the reactance of C at the ripple frequency is less than one-fifth the value of the resistor in the following RC filter section or one-fifth the reactance of the inductance in the following LC filter section. The relationships are expressed by the equations

$$C \geq \frac{796,000}{f_1 R_f} \tag{15.6}$$

and
$$C \geq \frac{127,000}{(f_1)^2 L_f} \tag{15.7}$$

where R_f = resistance in ohms of resistor in following RC filter section
  L_f = inductance in henrys of choke in following LC filter section
  C = filter capacitor, μf
  f_1 = frequency of lowest harmonic from rectifier (see Table 15.1)

End Filter Sections. The curves shown in Figs. 15.18 and 15.19 are also based on the assumption that the reactance of the output capacitor at the ripple frequency is less than one-fifth the d-c load resistance across the power-supply output. This relationship is expressed by Eq. (15.8).

$$C_o \geq \frac{796,000}{f_1 R_{\min}} \tag{15.8}$$

where $R_{\min}$ = minimum d-c resistance of load, ohms
  C_o = output filter capacitor, μf

When multisection filters are composed of dissimilar sections, the *over-all attenuation in decibels* can be determined by adding the individual attenuations in decibels as read from Figs. 15.18 and 15.19. The *over-all ripple-reduction factor* for a multisection filter composed of dissimilar sections is equal to the product of the individual reduction factors as read from Figs. 15.18 and 15.19.

Example 15.6

Design a filter for addition to an existent 60-cycle power supply having a full-wave rectifier and a choke-input filter in which the input choke has a value of 10 henrys and the shunting capacitor has a value of 8 μf. The power supply has a d-c output voltage of 450 volts at 0.4 to 0.5 amp with 4.5 volts rms ripple. The additional filter is to be composed of the optimum number of sections and should reduce the ripple voltage by a factor of 100 with a minimum reduction in the d-c output voltage.

Solution

1. Determine the advisability of using LC or RC filter sections.
Since the current through the filter is comparatively high, an LC filter is recommended.
2. Determine the minimum load resistance $R_{\min}$.
Since the load on the power supply varies between 0.4 and 0.5 amp, the minimum resistance of the load is determined at maximum current as follows:

$$R_{\min} = \frac{\text{d-c output voltage}}{\text{max d-c output current}} = \frac{450}{0.5} = 900 \text{ ohms}$$

3. Determine the minimum allowable size of C in the end filter section.
From Eq. (15.8)

$$C_o \geq \frac{796,000}{f_1 R_{\min}} = \frac{796,000}{120 \times 900} = 7.37 \ \mu\text{f}$$

Therefore let $C_o = 8 \ \mu$f.
4. Determine from Fig. 15.18 the optimum number of filter sections and the required value of LCf^2.
For an attenuation of 100 it can be seen from Fig. 15.18 that a two-section filter is optimum and $LCf^2 = 1,125,000$.
5. Determine LC.
The ripple frequency f is 120 cycles (see Table 15.1). Therefore

$$LC = \frac{1,125,000}{120^2} = 78.1$$

6. Determine the values of L and C in each filter section.
In step 3 it was determined that the output capacitor must be equal to 8 μf. If the filter sections are to have equal capacitance, the total value of C will be 16 μf. The total value of L is therefore equal to 78.1/16, or 4.9 henrys. Equal division of the total L between the two sections will make the inductance in each section equal to 2.45 henrys.
7. Determine by Eq. (15.7) the value of C necessary in the first and intermediate filter sections so that the reactance of C at the ripple frequency will be less than one-fifth the reactance of the inductance in the following filter section. If the proper relationship does not exist, it will be necessary to increase either L or C, whichever is the more economical.

$$C \geq \frac{127,000}{(f_1)^2 L_f} = \frac{127,000}{120^2 \times 2.45} = 3.6 \ \mu\text{f}$$

Since C is equal to 8 μf in both the first and intermediate filter sections, it is not necessary to increase either L or C.
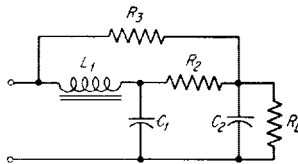
FIG. 15.20. *LC* cancellation filter.

15.6. Cancellation and Resonant Filters. Cancellation and resonant filters are most effective in filtering a specific ripple frequency.

15.6a. Cancellation Filters. Cancellation filters have the disadvantages of requiring careful adjustment of circuit values and have the characteristic of being rather ineffective at frequencies somewhat removed from the frequency at which maximum cancellation occurs.

In Fig. 15.20 the ripple-frequency current passing through the $L_1 C_1$ filter section and R_2 is shifted nearly 180° and, hence, tends to cancel the ripple-frequency current passing through R_3. If the currents are of the same amplitude and 180° out of phase,

the cancellation will be complete. The requirement that the two signal amplitudes be identical is satisfied for the circuit shown in Fig. 15.20 provided

$$\frac{R_3}{R_2} = \text{reduction factor for the undesired ripple frequency through } L_1 C_1$$

To approximate the desired 180° phase shift through $L_1 C_1$, the Q of L_1 must be fairly high. Most power-supply filter chokes will have a Q less than 6 to 10; consequently it may be necessary to use the configuration shown in Fig. 15.21. (For a single LC filter section, see the attenuation curve shown in Fig. 15.18 for the value of the reduction factor.)
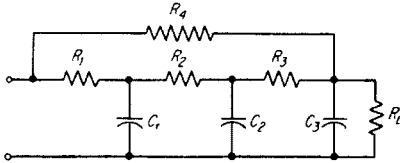
FIG. 15.21. *RC* cancellation filter.   FIG. 15.22. Tapped choke cancellation.
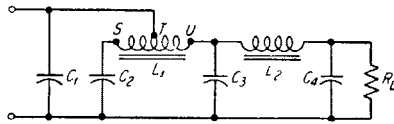
In Fig. 15.21 the phase shift of the ripple frequency is acquired in two *RC* sections instead of a single *LC* section. The reactances of C_1 and C_2 at the ripple frequency should be very small compared to the values of R_1 and R_2, respectively, in order to approach 180° phase shift in the two sections.[1] In addition, the reactance of C_1 should be small compared to the value of R_2. To obtain amplitude cancellation, the following relationship must hold for Fig. 15.21:

$$\frac{R_4}{R_3} = \text{over-all reduction factor for undesired ripple frequency through } R_1 C_1 \text{ and } R_2 C_2$$

(See Fig. 15.19 to determine reduction factor.)

A tapped choke cancellation filter section is shown in Fig. 15.22. The necessary relationships for ripple cancellation are given by Eq. (15.9).

$$X_{c2} = (n + 1) X_{ST} \tag{15.9}$$

[1] It should be noted that 180° phase shift cannot actually be obtained in a single LC section unless L and C have infinite Q's. In two RC sections, 180° phase shift cannot be obtained except with infinite capacitance or at infinite frequency. The amount of cancellation obtained with a few degrees less than 180° phase shift is usually sufficient and hence obviates the need of filter sections which are critical in both amplitude and phase shift; however, it is possible to modify both Figs. 15.20 and 15.21 to provide complete cancellation. With reference to Fig. 15.20, assume that a capacitor C_3 is added in series with R_3. Ordinarily, values for L_1, C_1 and R_3 are assumed, and the required values of R_2 and C_3 for complete cancellation are obtained from the following equations:

$$R_2 = \frac{R_3}{\omega^2 L_1 C_1 - 1}$$

$$C_3 = \frac{1}{\omega^2 L_1}$$

where $\omega = 2\pi f$
f = ripple frequency
With reference to Fig. 15.21, assume that a capacitor C_4 is added in series with R_4. If values are assumed for R and C, the required values of R_4 and C_4 for complete cancellation can be obtained from the following equations:

$$R_4 = \omega^2 R^3 C^2 - 3R$$

$$C_4 = \frac{1}{4\omega^2 R^2 C}$$

where $R = R_1 = R_2 = R_3$
$C = C_1 = C_2$

where X_{c2} = reactance of C_2 at undesired ripple frequency

X_{ST} = reactance between terminals S and T at undesired ripple frequency

n = ratio of turns in section TU to turns in section ST

If cancellation is complete, there is no ripple current through TU since the induced voltage in section TU due to the flow of ripple current through section ST will be equal and opposite to the applied ripple voltage. C_1, C_3, L_2, and C_4 provide additional filtering at frequencies other than the frequency at which cancellation occurs.
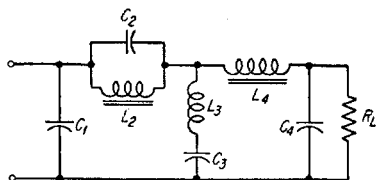
FIG. 15.23. Filter containing resonant series and shunt branches.

15.6b. *Resonant Filters.* In a filter containing resonant series and shunt sections such as L_2C_2 and L_3C_3 shown in Fig. 15.23, it is necessary that L and C in each branch resonate at the undesired ripple frequency. The necessary relationship is as follows:

$$LC = \frac{25{,}300}{f^2} \tag{15.10}$$

where f = ripple frequency, cps

C = capacitance, μf

L = inductance, henrys

Commercial tolerances on inductors and capacitors usually make improbable the satisfaction of Eq. (15.10) with production values. Another consideration is the fact that the inductance of an iron-core reactor is a function of both the direct current through the winding and the ripple voltage across the winding; hence it is advisable to be able to adjust the value of C for each inductor under operating conditions.

Example 15.7

Design a 120-cycle cancellation filter of the type shown in Fig. 15.21 for a load current of 20 ma and a d-c voltage drop across the filter of 150 volts.

Solution

1. Determine the sum of R_1, R_2, and R_3.

$$R_1 + R_2 + R_3 = \frac{\text{specified voltage drop}}{\text{load current}}$$

$$= \frac{150}{0.02} = 7{,}500 \text{ ohms}$$

2. Determine the values of R_1, R_2, and R_3.

Arbitrarily let both R_1 and R_2 equal 3,500 ohms. This choice is based on the desire to make R_1 and R_2 large to achieve maximum ripple attenuation in R_1C_1 and R_2C_2. The value of R_3 directly affects the size of R_4; consequently the sum of R_1 and R_2 should not be excessively large, otherwise the value of R_3 must be very small. Whether the size of R_3 is satisfactory will not be known until R_4 has been determined in step 5. R_4 should be much larger than the sum of R_1, R_2, and R_3 in order not to upset the voltage-drop calculations which were dependent on the load current flowing through R_1, R_2, and R_3.

3. Determine the values of C_1, C_2 and C_3.

The reactances of C_1 and C_2 at the ripple frequency should be much less than the values of the resistors R_1 and R_2 to achieve the maximum phase shift in R_1C_1 and R_2C_2. R_1 and R_2 are each equal to 3,500 ohms, hence the reactances of C_1 and C_2 might arbitrarily be required to be equal to or less than one-fiftieth of 3,500 ohms. This will cause a total phase shift of approximately 178°. Therefore,

$$\frac{3{,}500}{50} \geq \frac{1}{2\pi f C}$$

$$70 \geq \frac{1}{2 \times 3.14 \times 120 \times C}$$

$$C \geq 18.9 \ \mu\text{f}$$

The minimum value for both C_1 and C_2 is 18.9 μf. In Sec. 15.5 it was pointed out that the maximum ripple reduction is realized if the two filter sections are identical. For this reason, let both C_1 and C_2 equal 20 μf. C_3 lowers the power-supply output impedance and attenuates any uncanceled ripple or additional frequency components which might be present. Since there is no requirement on C_3 in this example, it could arbitrarily be made equal to C_1 and C_2.

4. Determine the ripple attenuation through R_1C_1 and R_2C_2 from Fig. 15.19.

$$R_1C_1f = 3.5 \times 20 \times 120 = 8,400$$

where R_1 is in thousands of ohms, C is the μf, f is ripple frequency. From the one section curve, the ripple-reduction factor is found to be 53. For the two cascaded sections the ripple reduction factor is equal to $(53)^2$, or 2,810.

5. Determine the value of R_4.

$$\frac{R_4}{R_3} = \text{over-all reduction factor in ripple through } R_1C_1 \text{ and } R_2C_2 \text{ where } R_3 = 500 \text{ ohms}$$

$$\frac{R_4}{500} = 2,810$$

$$R_4 = 1,405,000 \text{ ohms}$$

To allow for variations in resistor and capacitor values, R_4 can be made variable with a maximum value somewhat larger than the calculated value. Since the calculated value of R_4 is equal to 1.405 megohms, a variable resistor with a maximum value of 2 megohms would be satisfactory. R_4 should be adjusted for maximum cancellation.

15.7. Gas-tube Voltage Regulators

15.7a. Fundamental Voltage-regulator Circuit. The basic gas-tube voltage-regulator circuit is shown in Fig. 15.24. The regulator tube in parallel with a varying load will tend to maintain an essentially constant load voltage provided the operating
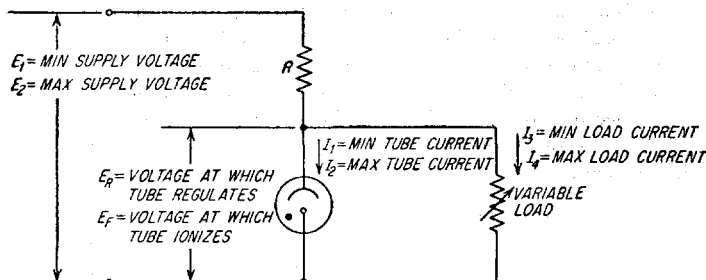
Fig. 15.24. Gas-tube voltage regulator.

limits of the tube are not exceeded. Since the voltage across the load and regulator tube is essentially constant, even in the presence of a varying load current, the voltage drop across the source resistance R will also remain essentially constant for a given source voltage. A constant voltage drop across R stipulates that the current through R is constant, hence it is apparent that any increase in load current must be accompanied by an equal decrease in tube current and a decrease in load current must be accompanied by an equal increase in tube current.

To realize the voltage-regulating capabilities of a gas tube, it is necessary to ionize the gas by increasing the voltage gradient in the deionized tube until the gas molecules become ionized. Ionization is maintained in the tube by the flow of electrons from the cathode to the plate. The positive ions neutralize most of the negative space charge within the tube, and the voltage drop across the tube decreases because of the reduced tube resistance. The voltage at which the tube ionizes is called the firing voltage E_F, and the voltage during ionization is termed the regulating voltage E_R.

A gas tube is not an absolutely stable voltage reference. If certain tubes are alternately ionized and deionized, the tube voltage during any ionization period may differ as much as 1 per cent from that of the preceding ionization period. Also, the voltage drop across certain gas tubes over long periods of continued ionization can show the same degree of instability for no apparent reason. In addition, gas tubes are light-sensitive, consequently their ionizing characteristics are dependent to some extent on the ambient lighting. Variations in the current through a gas tube will also cause small variations in its terminal voltage.

The minimum current through a voltage-regulator tube must be greater than that current required to sustain ionization. The maximum current must be less than that which causes either excessive heat dissipation within the tube or damage to the cathode because of positive ion bombardment.

15.7b. Surge Currents Due to Shunting Capacitance. Frequently there is a requirement for a capacitor across the regulator tube to provide additional filtering of the source voltage, to suppress noise generated within the gas tube, or to lower the output-voltage source impedance as seen by the load. The presence of the capacitor across the regulator tube has the disadvantage of causing a surge current, which is many times greater than the rated tube current, to flow through the tube at the instant of ionization. At the moment the gas tube ionizes, the capacitor is charged to the ionizing potential. After the tube ionizes, the tube voltage drops to the regulating voltage, and at the same time the capacitor partially discharges through the tube. The larger the capacitor, the larger the discharge current through the tube and hence the increased possibility of damage. The manufacturer usually specifies the maximum allowable value for this capacitor.

15.7c. Resistance of Voltage-regulator Tubes. Figure 15.25 shows the voltage-current relationship in a typical gas tube. The slope of the curve at any given point, is equal to the dynamic resistance of the tube at that point. In the regulator circuit shown in Fig. 15.24, a change in the load current will cause a change in the tube current and the voltage across the tube. The change in voltage will be equal to the product of the change in tube current and the dynamic resistance of the tube. The dynamic resistance of most gas tubes remains essentially the same at very low frequencies and conventional power-supply ripple frequencies but increases appreciably at higher frequencies.
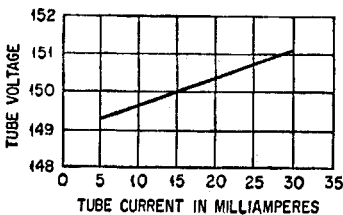
FIG. 15.25. Measured character-istics of a sample *OA2.*
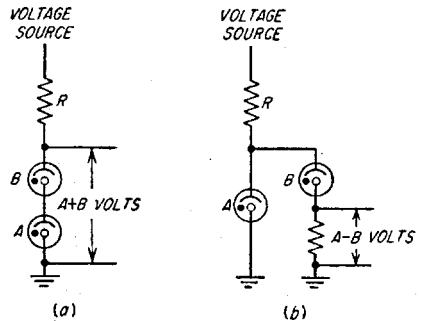
FIG. 15.26. Methods of obtaining regulated voltages which are the sum and difference of the regulator tubes' voltages.

15.7d. Special Circuit Configurations. Often there is a requirement for a regulated voltage which is greater or less than that which can be obtained by any single voltage-regulator tube available. In such cases it is frequently possible to obtain the desired voltage by one of the methods shown in Fig. 15.26.

15.7e. Oscillation in Voltage-regulator Circuits. If the regulator tube is shunted by capacitance and the source resistance R is so large that the current is too small to maintain ionization in the tube, the circuit will oscillate. The oscillation cycle consists of the capacitor charging through R to the tube firing voltage, ionization of the tube and partial discharging of the capacitor, deionization due to the current through R being too small to sustain ionization, recharging of the capacitor to the tube firing voltage, etc. The frequency of oscillation is a function of the values of the dropping resistor, the shunting capacitance, the difference between the deionizing and ionizing voltages, and the source voltage.

If the value of the shunting capacitance is too large, oscillation sometimes exists in a voltage regulator circuit even though R is small enough to maintain ionization in the tube in the absence of the shunting capacitance. This type of oscillation is only possible if some portion of the curve for the regulating voltage versus tube current has a slope which is negative. In this case the oscillation cycle is initiated by an excessive discharge current, at the instant of tube ionization, from the large capacitor. This current passes through the tube and causes the terminal voltage of the tube to drop until the tube extinction potential is reached. The tube then deionizes since the capacitor C is sufficiently large that it cannot rapidly be charged back to the proper tube voltage for steady-state regulation. After the tube deionizes, the capacitor again charges to the gas-tube firing voltage, and the cycle repeats itself.

15.7f. Calculation of Circuit Values. The first step in the design of a voltage-regulator circuit of the type shown in Fig. 15.24 is to assume a value for the minimum regulator tube current I_1. The assumed current I_1 must satisfy only Eq. (15.11) if the minimum supply voltage E_1 has not been specified and is to be established. The minimum possible value for E_1 which will permit satisfactory regulation for the assumed value of I_1 can then be determined from Eq. (15.13). If E_1 has been specified, the assumed value of I_1 must satisfy Eq. (15.12).

Method A for establishing I_1 when E_1 is to be established:

The assumed value of I_1 must satisfy the following two equations:

$$I_1 \geq \text{tube's minimum current rating}$$

and $$I_1 > \frac{(E_F - E_R)I_4}{E_R} \qquad (15.11)$$

Refer to Fig. 15.24 for definition of terms.

Method B for establishing I_1 when E_1 is specified:

The assumed value of I_1 must satisfy the following two equations:

$$I_1 \geq \text{tube's minimum current rating}$$

and $$I_1 \geq I_{\min} = \frac{E_1 I_4 (E_F - E_R)}{E_R (E_1 - E_F)} \qquad (15.12)$$

where $I_{\min}$ = minimum allowable tube current for specified value of E_1
Refer to Fig. 15.24 for definition of other terms.

Equation (15.12) states that the actual minimum tube current I_1 must be equal to or greater than the minimum allowable tube current $I_{\min}$. If Method A is used in the establishment of I_1, it is necessary to employ Eq. (15.13) in the determination of E_1. It can be seen that the minimum allowable value of E_1 is a function of the assumed value of I_1.

$$E_1 \geq E_{\min} = \frac{E_R E_F I_1}{E_R (I_1 + I_4) - E_F I_4} \qquad (15.13)$$

where $E_{\min}$ = minimum allowable supply voltage for specified value of I_1
Refer to Fig. 15.24 for definition of other terms.

Equation (15.13) states that the actual minimum supply voltage E_1 must be equal to or greater than the minimum allowable supply voltage E_{min} for the assumed value of I_1.

The remaining calculations using Eqs. (15.14), (15.15), and (15.16) are the same for either the A or B method.

$$R = \frac{E_1 - E_R}{I_1 + I_4} \qquad\qquad (15.14)$$

$$K = \frac{E_2}{E_1} \qquad\qquad (15.15)$$

$$I_2 = \frac{E_2 - E_R}{R} - I_3 \qquad\qquad (15.16)$$

The minimum possible value for I_2 can be determined from

$$I_2 \geq K(I_1 + I_4) - I_3 \qquad\qquad (15.17)$$

Refer to Fig. 15.24 for definition of terms.

The value of the maximum tube current I_2, determined by Eq. (15.16), should be equal to or less than the maximum current rating of the tube. If the value of I_2 is excessive, it sometimes can be reduced sufficiently by increasing both the minimum supply voltage E_1 and the maximum supply voltage E_2. The absolute-minimum possible value for I_2 is given by Eq. (15.17). This minimum occurs, however, when the value of the supply voltage is infinite. *Equation (15.17) therefore permits the recognition of the impossible regulation requirement where the calculated minimum possible value of I_2 is larger than the tube maximum current rating.*

Example 15.8

If the unregulated voltage supply for a gas-tube voltage regulator is expected to vary 20 per cent, i.e., $E_2 = 1.20\,E_1$, determine satisfactory values for E_1, E_2, and R for a regulator which has the following load requirements:

    a. Output voltage to be 150 volts (use an OA2-type tube)

    b. Load current to vary between 20 and 30 ma

Solution

1. Determine the minimum possible value for the minimum tube current I_1 (use Method A).

From Eq. (15.11)

$$I_1 > \frac{(180 - 150) \times 0.030}{150} = 0.006 \text{ amp, or 6 ma}$$

where E_F = 180 volts (firing voltage of an OA2)

     E_R = 150 volts

     I_4 = 0.030 amp

therefore I_1 might arbitrarily be made equal to 8 ma.

2. Determine the minimum supply voltage E_1 based on the value of I_1 determined in step 1.

From Eq. (15.13)

$$E_1 \geq E_{min} = \frac{150 \times 180 \times 0.008}{150(0.008 + 0.030) - (180 \times 0.030)}$$

$$E_1 \geq E_{min} = 720 \text{ volts}$$

therefore let E_1 = 725 volts.

3. Determine E_2.

$$E_2 = 1.20E_1 = 1.20 \times 725$$
$$= 870 \text{ volts}$$

4. Determine R.
From Eq. (15.14)

$$R = \frac{725 - 150}{0.008 + 0.030}$$

$$= 15{,}130 \text{ ohms}$$

5. Determine the maximum tube current I_2.
From Eq. (15.16)

$$I_2 = \frac{870 - 150}{15{,}130} - 0.020$$

$$= 0.028 \text{ amp, or } 28 \text{ ma}$$

(which is within the maximum current rating of 30 ma).

Note: E_1 and E_2 could both be reduced if a larger value of I_1 had been assumed in step **1**. However, an incremental increase in I_1 means an even larger incremental increase in I_2.

Example 15.9

Determine the circuit values for a regulator which will deliver approximately 150 volts to a load in which the current varies between 25 and 30 ma. The regulator circuit is to operate from a supply voltage which varies between 450 and 525 volts. The regulator tube is to be an OA2.

Solution

1. Determine the minimum allowable value for the minimum tube current I_1 (use Method B).
Since E_1 has been specified, the minimum allowable value of I_1 is determined from Eq. (15.12).
From Eq. (15.12)

$$I_1 \geq I_{\min} = \frac{450 \times 0.030(180 - 150)}{150(450 - 180)} = 0.010 \text{ amp, or } 10 \text{ ma}$$

therefore I_1 might arbitrarily be made equal to 12 ma.
2. Determine R.
From Eq. (15.14)

$$R = \frac{450 - 150}{0.012 + 0.030}$$

$$= 7{,}140 \text{ ohms}$$

3. Determine the maximum tube current I_2.
From Eq. (15.16)

$$I_2 = \frac{525 - 150}{7{,}140} - 0.025$$

$$= 0.028 \text{ amp, or } 28 \text{ ma}$$

15.8. Vacuum-tube Voltage Regulators. The three general types of electronic voltage regulators for power supplies are as follows:

1. *Voltage regulators for fixed loads.* A voltage regulator for a fixed load will maintain a constant voltage across the load even though the input voltage may vary.

2. *Degenerative voltage regulators.* Degenerative regulators have the characteristic of attempting to maintain a constant output voltage even though both the load current and input voltage vary. Because of the principle by which they operate, degenerative regulators are never able to maintain an absolutely constant output voltage. However, the majority of all regulation requirements are more than satisfied by this type of regulator.

3. *Compensated degenerative voltage regulators.* Compensated degenerative regulators offset some of the inabilities of degenerative regulators to give ideal performance. Compensated degenerative regulators can be designed which will have zero a-c output impedance (within certain frequency limits) and zero d-c output impedance as well as other desirable characteristics. This type of regulator has the undesirable

characteristic of requiring careful adjustment of the compensation, and, in addition, a drift in component parts may cause a more severe loss in regulation than in a degenerative regulator.

15.8a. Voltage Regulators for Fixed Loads. Each of the circuits shown in Figs. 15.27 and 15.28 will maintain a constant output voltage for a fixed load even though the input voltage varies. Since these circuits are not intended to regulate the output voltage for load current changes, it is recommended that they be used only if the load is constant. If the sources which supply the unregulated voltages to the regulators are not grounded in any way, both types of regulators can serve as either a positive or a negative supply. Detailed descriptions of the circuits are given in the following paragraphs.

Shunt Regulator for a Fixed Load. The shunt regulator shown in Fig. 15.27 will supply a constant voltage to a fixed load with a varying input voltage. For proper operation, a fraction of the input voltage changes is applied to the grid of $V1$, thereby causing current variations through $V1$ and R_5, creating a change in the voltage drop across R_5 equal to the input voltage change. The net result is that the plate voltage of $V1$ does not change. Since the fraction of the input voltage variations applied to the grid of $V1$ causes only the plate current to change, it is apparent that tubes with large values of g_m require smaller frac-

(a) SERIES REGULATOR FOR A FIXED LOAD

(b) IMPROVED SERIES REGULATOR FOR A FIXED LOAD

FIG. 15.28. Series regulators for fixed loads.

FIG. 15.27. Shunt regulator.

tions of the input voltage variations to be applied to their grids to achieve regulation. Example 15.10 illustrates a typical design procedure.

Series Regulators for Fixed Loads. The series regulators shown in Fig. 15.28a and b will supply a constant voltage to a fixed load from a varying input voltage. In each circuit the voltage changes across R_2, caused by changes in the input voltage, are amplified by $V1$. The polarity of the amplified signals at the plate of the tube is the same as the polarity of the input-voltage changes. Therefore, if the amplified signals are of the same amplitude as the changes in the input voltage, the voltage across R_L will remain constant. Since the fraction of the input-voltage variations applied to the grid of $V1$ causes only the plate voltage to change, it is apparent that tubes with large values of μ require smaller fractions of the input voltage to be applied to their grids to achieve regulation.

The circuit shown in Fig. 15.28a has limited application. In many instances the circuit cannot be utilized since there is a conflict in the operational requirements. First there is the requirement for a certain ratio of R_2 to $R_1 + R_2$ for proper regulation of the input-voltage changes, and secondly there is the requirement that the d-c voltage drop across R_2, because of the voltage divider action of R_1 and R_2 across

the input voltage, be considerably less[1] than the bias required on $V1$ to establish the desired d-c output voltage across the load. If the d-c voltage drop across R_2, due only to the divider action, approaches or is greater than the allowable bias on $V1$, the circuit in Fig. 15.28a cannot be used. Figure 15.28b obviates this limitation.

Example 15.11 illustrates the details to be considered in the design of the two types of series regulators.

Example 15.10

Design a regulator of the type shown in Fig. 15.27 which will provide a regulated output of 260 volts at 10 ma from a voltage source which varies from 365 to 435 volts.

NOTE: In the detailed analysis which follows, the voltages and currents associated with the mid-value of the source voltage are referred to as the nominal values.

Solution

1. Determine the output voltage E_o and the load current I_L.
From statement of problem

$$E_o = 260 \text{ volts}$$
$$I_L = 0.010 \text{ amp}$$

2. Determine the value of I_2.
The value of R_4 is not critical, but its value cannot be so great that it is impossible to realize the desired value for R_2 to be determined in step 8. Satisfactory operation can be assured if R_4 has a value which is smaller than the desired value of R_2. Since step 8 is dependent on this step, it is necessary to tentatively assign a value of resistance to R_4. To minimize the current requirements, a tentative value for R_4 might be 50,000 ohms.
Therefore,

$$I_2 = \frac{E_o}{R_4} = \frac{260}{50,000} = 0.0052 \text{ amp}$$

3. Determine the type of tubes to be used.
The lower the operating voltage of the gas tube, the greater will be the voltage drop across $V1$. The choice of a 5644 for the gas tube (nominally regulates at 95 volts) apparently is satisfactory since the fixed voltage drop across $V1$ would be 260 − 95, or 165 volts, which is within the ratings of most small tubes which might be used for $V1$. Arbitrarily assume that $V1$ is a 5902 (triode-connected).

4. Determine the nominal current through $V1$ and $V2$.
The minimum current through $V1$ and $V2$ should be no less than 5 ma because of the minimum rating of $V2$. The maximum allowable current through $V1$ is 24 ma since at 165 volts and 24 ma the tube maximum plate dissipation rating of 4 watts is equaled. Since the maximum current rating of $V2$ is 25 ma, the maximum current is limited by $V1$ rather than $V2$. The nominal or mid-value of current through $V1$ and $V2$ which permits the maximum dynamic range of regulation is therefore $(24 + 5)/2$, or 14.5 ma. This current should flow through $V1$ and $V2$ when the input voltage is equal to the nominal value of 400 volts.

5. Determine the nominal value of the current I_1 through R_5.

$$
\begin{aligned}
I_1 &= I_2 + I_3 + I_L \\
&= I_2 + \text{nominal value of current through } V1 \text{ and } V2 + I_L \\
&= 0.0052 + 0.0145 + 0.010 \\
&= 0.030 \text{ amp}
\end{aligned}
$$

6. Determine the value of R_5.
The value of R_5 is that value required to drop the nominal source voltage of 400 volts down to the output voltage of 260 volts when the nominal current of 30 ma is flowing through R_5.

$$
\begin{aligned}
R_5 &= \frac{\text{nominal source voltage} - \text{output voltage}}{\text{nominal current through } R_5} \\
&= \frac{400 - 260}{0.030} = 4,670 \text{ ohms}
\end{aligned}
$$

---

[1] The reason that this value must be considerably less than the bias required on $V1$ is that the load current flows through R_2, and therefore the d-c voltage drop across R_2 will be greater than that determined only by the divider action of R_1 and R_2 across the source voltage.

7. Determine the input-voltage limits over which the circuit will regulate.

The current through $V1$ and $V2$ can increase or decrease 9.5 ma from the nominal value of 14.5 ma as determined in step 4. The possible source-voltage deviation from the nominal value is therefore equal to $R_5 \times 9.5$ ma, or approximately 44 volts. This circuit will therefore regulate from a source voltage of 400 plus or minus 44 volts, or 356 to 444 volts.

8. Determine the value of R_1 and R_2.

The required relationship between R_1 and R_2 is given by the equation

$$\frac{R_1}{R_2} = g_m R_5 - 1$$

$$= \frac{\mu R_5}{r_p} - 1$$

$$= \frac{15 \times 4,670}{4,000} - 1 = 16.5$$

where g_m = transconductance of $V1$
r_p = dynamic plate resistance of $V1$
μ = amplification factor of $V1$

As mentioned in step 2, the value of R_2 cannot safely be made significantly smaller than R_4. In this case R_4 has been set equal to 50,000 ohms (assumed in step 2), therefore R_2 can also be given a value of 50,000 ohms. Since R_1/R_2 is equal to 16.5, R_1 is equal to $16.5 \times 50,000$, or 825,000 ohms. It is suggested that R_3 be a 100,000-ohm potentiometer in the event resistor and tube tolerances require a wider range of adjustment.

Summary

The adjustments of R_3 and R_4 are as follows: With the input voltage set at its nominal value, adjust R_4 to obtain the desired output voltage across the load. Vary the input voltage slowly between its operating limits and note whether or not the output voltage increased or decreased when the input voltage was increased. If the output voltage increased, it is necessary to increase the value of R_3. Conversely, R_3 should be decreased if the output voltage decreased when the input voltage was increased. Associated with each adjustment of R_3 is a readjustment of R_4. R_4 should be readjusted with the same procedure as that used in the initial adjustment.

Note that the current through $V1$ and $V2$ increases if the load on the power supply is removed, but the increase will not be more than the value of the load current. If $V1$ and $V2$ cannot safely carry this increase in current, special precautions should be taken to prevent the removal of the load.

Example 15.11

Design a series-type regulator using one-half of a 12AU7 tube which will operate from a voltage source that varies between 275 and 325 volts and which will provide a regulated output of −100 volts at 10 ma to a fixed load. Use the circuit shown in Fig. 15.28a if possible; otherwise use the circuit shown in Fig. 15.28b.

NOTE: The design is independent of the polarity of the output voltage; however, there is a requirement that the power supply be grounded only at the positive terminal of R_L in order to achieve the negative polarity required in this example. In the detailed analysis which follows, the voltages and currents associated with the center value of the input voltage are referred to as the nominal values.

Solution

1. Determine E_o and I_L.
From statement of problem

$$E_o = 100 \text{ volts}$$
$$I_L = 0.010 \text{ amps}$$

2. Determine the required value of $R_2/(R_1 + R_2)$ for proper regulation.

$$\frac{R_2}{R_1 + R_2} = \frac{1}{\mu + 1}$$
$$= \frac{1}{21}$$

3. Determine the minimum possible value of the voltage drop across R_2 based on the required ratio $R_2/(R_1 + R_2)$ and the minimum supply voltage.

$$\text{Minimum possible value of the d-c voltage drop across } R_2 = \tfrac{1}{2}_1 \times 275$$
$$= 13.1 \text{ volts}$$

4. Assuming the use of the circuit shown in Fig. 15.28a, determine the sum of the voltage drops across the tube and across R_2 at the minimum value of the source voltage.

$$\text{Voltage drop across tube} + \text{voltage drop across } R_2 = \text{minimum input voltage} - E_o$$
$$= 275 - 100$$
$$= 175 \text{ volts}$$

5. For the circuit shown in Fig. 15.28a, determine the required values of tube voltage drop and bias (voltage drop across R_2) necessary to obtain the desired output voltage when the input voltage is at its minimum value.

During proper regulation, the current through $V1$ has a constant value of 10 ma; therefore on the plate characteristics of one triode section of a 12AU7, draw a horizontal line through the ordinate at 10 ma. The operation of the tube will be along this constant current line. To determine the operating point at the minimum value of the input voltage, find a point along the constant current line at which the sum of the plate voltage and the associated value of bias is equal to 175 volts. At a point on the constant current line, a bias of slightly more than 4 volts coincides with a drop across the tube of approximately 171 volts and thus fulfills the required conditions. Therefore the required bias and the tube voltage drop must be equal to 4 and 171 volts, respectively, if the circuit shown in Fig. 15.28a is to be used. In the event the tube characteristics should show that the bias must be positive, it would be necessary to change to another tube type.

6. Compare the minimum possible value of the voltage drop across R_2 as determined in step 3 with the value of the required bias as determined in step 5 to determine whether or not it is possible to use the circuit shown in Fig. 15.28a.

The minimum possible value of the voltage drop across R_2 is equal to 13.1 volts as determined in step 3.

The required value of the bias (voltage drop across R_2) for $V1$ is equal to 4 volts as determined in step 5.

Note that the results of steps 3 and 5 are in conflict; hence it is impossible to use the circuit shown in Fig. 15.28a with this particular tube type and it is necessary to use the circuit shown in Fig. 15.28b.

NOTE: If the circuit shown in Fig. 15.28a could have been used, the values of R_1 and R_2 could have been determined from the following equations.

$$R_2 = \frac{|E_c(\mu + 1)| - |\text{minimum input voltage}|}{\mu I_L}$$

where E_c = required value of bias as determined in step 5
  μ = amplification factor of $V1$
  $R_1 = \mu R_2$

If the calculated values of R_1 and R_2 should be abnormally small, causing an unusually high current drain through R_1, it may be more economical to use higher values and employ the circuit shown in Fig. 15.28b. In general, R_1 and R_2 should be variable with maximum values greater than the calculated values to allow for circuit tolerances.

7. Assume a value for the sum of R_1 and R_2 in the circuit shown in Fig. 15.28b.

The sum of R_1 and R_2 is not critical, but since the load current passes through R_2, it is desirable that R_2 be made as small as possible without excessively increasing the current through R_1. In this example arbitrarily assume that the sum of R_1 and R_2 is equal to 50,000 ohms.

8. Determine the values of R_1 and R_2 based on the assumption made in step 7.

$$\frac{R_2}{R_1 + R_2} = \frac{1}{\mu + 1}$$
$$\frac{R_2}{50,000} = \frac{1}{20 + 1}$$
$$R_2 = 2,380 \text{ ohms}$$

and
$$R_1 = 50,000 - 2,380$$
$$= 47,620 \text{ ohms}$$

9. Determine the nominal value of the voltage E_K at the cathode of $V1$ in the circuit of Fig. 15.28b from the following equation.

$$E_K = \frac{\text{nominal input voltage} + (R_2 I_L \mu)}{\mu + 1}$$

$$= \frac{300 + (2{,}380 \times 0.010 \times 20)}{21} = 37 \text{ volts}$$

10. Determine the necessary bias E_c on $V1$ and the grid voltage E_g.

$$\begin{aligned}
\text{Voltage drop across } V1 &= \text{nominal input voltage} - |E_o| - |E_K| \\
&= 300 - 100 - 37 \\
&= 163 \text{ volts}
\end{aligned}$$

From the 12AU7 tube characteristics it can be determined that the necessary bias for a 163-volt tube drop at 10 ma is approximately −4 volts. The grid must therefore be connected to a voltage source equal to $E_K + E_c$, or 33 volts. The voltage source is the potentiometer across the gas tube $V2$.

11. Determine whether or not the circuit will regulate between the specified limits.

The lower and upper input voltage limits are 275 and 325 volts, respectively, and the change in bias on $V1$ is given by the following equation:

$$\text{Change in bias} = \frac{R_2}{R_1 + R_2} \times \text{change in input voltage from nominal value}$$

$$= \frac{2{,}380}{50{,}000} \times (\pm 25)$$

$$= \pm 1.19 \text{ volts}$$

Since the nominal bias is −4 volts, the bias limits are

$$-4 + 1.19 = -2.81 \text{ volts}$$
$$-4 - 1.19 = -5.19 \text{ volts}$$

The bias does not go positive; therefore, the circuit will properly regulate over the required limits. Note that the value of the sum of R_1 and R_2 assumed in step 7 will affect the bias range.

15.8b. Degenerative Voltage Regulators. A characteristic of degenerative-type regulators is that they contain a closed control or feedback loop. Changes in the output voltage caused by changes in either the input voltage or output current are opposed by the action of this loop. It is important to recognize that in the strictly degenerative type of regulator the input impedance to the regulator is always less than infinity and the output impedance of the regulator is always greater than zero; i.e., the input-voltage and load-current changes will always cause a change in the output voltage. In very good regulators the change in output voltage may be very small, but it does exist.

Since a power supply will regulate better for load-current changes if the output impedance is low, there are applications in which minimizing the power-supply output impedance is very important. The output impedance of a degenerative voltage regulator will vary with frequency. This is due to the fact that the reactance values of the circuit shunting capacitances vary with frequency. At high frequencies the small values of reactance reduce the gain of the feedback loop within the regulator and consequently tend to increase the power-supply output impedance. This can, in part, be offset by shunting a capacitor across the power-supply output. The presence of the capacitor tends to lower the output impedance with increasing frequency. If the regulator feedback circuit has only the required gain and bandwidth to maintain a sufficiently low output impedance over a limited frequency range, the value of the output shunting capacitor should be made large enough to ensure that

its impedance is equal to or less than the maximum allowable value for the power supply in the frequency range at and above that at which the regulator feedback circuit gain is appreciably reduced. This will ensure a low power-supply output impedance for those high frequencies at which the regulator is incapable of regulating.

Simple Series Types of Degenerative Regulators. The regulators shown in Fig. 15.29 are of the series type and have configurations similar to those of cathode followers. The three circuits are similar except for the manner in which the grids are biased, which in turn affects both their a-c and d-c regulating characteristics. The preceding discussion in regard to the frequency characteristics of degenerative regulators is not applicable since this type of regulator does not have a separate amplifier tube.

The circuit shown in Fig. 15.29a is one of the simplest types of degenerative regulators. The fraction $R_1/(R_1 + R_2)$ of any output-voltage change ΔE_o appears as a bias change on $V1$. This change in bias effectively puts a generator of voltage $\mu\Delta E_o R_1/(R_1 + R_2)$ in opposition to the change. Optimum regulation occurs when $R_1/(R_1 + R_2)$ is equal to unity; therefore, performance becomes somewhat less than optimum if the desired operating bias for $V1$ is low with respect to the output voltage since the fraction $R_1/(R_1 + R_2)$ would be much less than unity. The output-voltage control is the cathode potentiometer of $V1$. The a-c and d-c voltage-reduction factors k_r and K_r, respectively, i.e., the ratio of input-voltage changes to the resulting output-voltage changes, are given by Eq. (15.18). The regulator output impedance Z_o and output resistance R_o are given by Eq. (15.19).
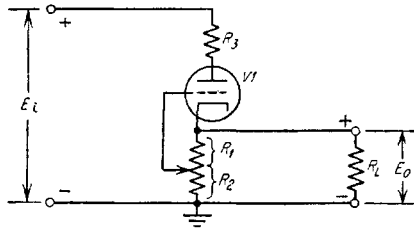
(a) DEGENERATIVE REGULATOR
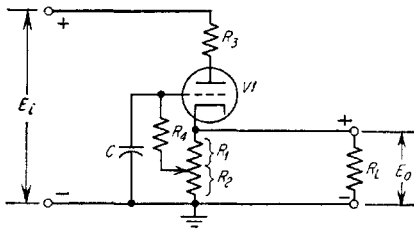
(b) DEGENERATIVE REGULATOR WITH OPTIMUM A-C CHARACTERISTICS
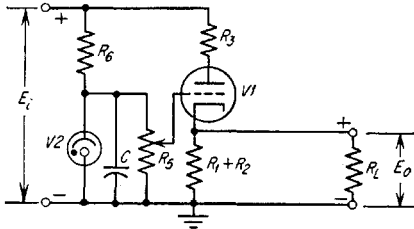
(c) DEGENERATIVE REGULATOR WITH OPTIMUM A-C AND D-C CHARACTERISTICS

FIG. 15.29. Simple series regulators.

$$k_r = K_r = \frac{\Delta e_i}{\Delta e_o} = \frac{\Delta E_i}{\Delta E_o} = \frac{r_p + R_3}{R_k} + \frac{\mu R_1}{R_1 + R_2} + 1 \qquad (15.18)$$

where $R_k = \dfrac{R_L(R_1 + R_2)}{R_1 + R_2 + R_L}$

Δe_i = a-c input voltage (ripple)

Δe_o = a-c output voltage (ripple)

ΔE_i = d-c input-voltage change

ΔE_o = d-c output-voltage change

k_r = ratio of an a-c input voltage to associated a-c output voltage

K_r = ratio of a d-c input-voltage change to associated d-c output-voltage change

$$R_o = Z_o = \frac{(R_1 + R_2)(R_3 + r_p)}{R_1(\mu + 1) + R_2 + R_3 + r_p} \qquad (15.19)$$

The circuit in Fig. 15.29b is the same as that of Fig. 15.29a with the exception of R_4 and C. The purpose of R_4 and C is to stabilize the grid with respect to ground for a-c input voltage or a-c load current changes. For an a-c output voltage change Δe_o, there appears a bias change of Δe_o on $V1$. This change in bias effectively puts a generator of voltage $\mu \Delta e_o$ in opposition to the change. R_4 should be several times greater than the reactance of C at the lowest frequency of input-voltage and load-current variations. Arbitrarily some minimum ratio such as 10:1 could be assumed. The d-c characteristics of the circuit are the same as for the circuit shown in Fig. 15.29a, but the a-c characteristics have been improved and are given by Eqs. (15.20) and (15.21).

$$k_r = \frac{\Delta e_i}{\Delta e_o} = \frac{r_p + R_3}{R_k} + \mu + 1 \tag{15.20}$$

where $R_k = \dfrac{R_L(R_1 + R_2)}{R_1 + R_2 + R_L}$

$$Z_o = \frac{(R_1 + R_2)(r_p + R_3)}{(R_1 + R_2)(\mu + 1) + R_3 + r_p} \tag{15.21}$$

For both alternating and direct current, the circuit shown in Fig. 15.29c has the maximum reduction factors and the minimum output impedances possible for a cathode-follower type of regulator. The output-voltage control is the grid potentiometer R_5. The applicable equations are:

$$k_r = K_r \text{ [use Eq. (15.20)]}$$
$$R_o = Z_o \text{ [use Eq. (15.21)]}$$

If R_3 is small compared to r_p and if $R_1 + R_2$ is large compared to the tube output impedance, the a-c output impedance of the circuit in Fig. 15.29b and the a-c and d-c output impedances of the circuit in Fig. 15.29c will be approximately equal to $r_p/(\mu + 1)$, or very nearly equal to $1/g_m$.

Example 15.12

Assume that one triode section of a 12AU7 tube is to be used in a circuit of the type shown in Fig. 15.29a to supply an output of 150 volts to a load of 25,000 ohms.

a. Determine the circuit values if the source resistance R_3 is equal to 0 ohms and the input voltage is equal to 250 volts.

b. Determine the increase in the output voltage if the input voltage were to be increased to 300 volts.

c. Determine the increase in the output voltage if the load current were to be decreased by 2 ma.

Solution

1. Determine the value of R_K.

$$R_L = 25,000 \text{ ohms}$$

Let $R_1 + R_2 = 250,000$ ohms (for low-current drain). Therefore

$$R_K = \frac{(R_1 + R_2)R_L}{R_1 + R_2 + R_L} = \frac{250,000 \times 25,000}{250,000 + 25,000}$$
$$= 22,700 \text{ ohms}$$

2. Determine the required bias voltage.

The required voltage drop across the tube is equal to the supply voltage minus the output voltage, i.e., $250 - 150$, or 100 volts. The current through the tube is equal to the output voltage divided by R_k, i.e., $150/22,700$, or 6.6 ma. The bias required for the 12AU7 tube is approximately -2 volts.

3. Determine the values of R_1 and R_2.

The required bias is -2 volts, and the cathode voltage is 150 volts. Therefore,

$$R_1 = \frac{|\text{bias}|}{\text{output voltage}} \times (R_1 + R_2)$$
$$= \frac{2}{150} \times 250,000$$
$$= 3,330 \text{ ohms}$$
$$R_2 = 250,000 - R_1$$
$$= 246,700 \text{ ohms}$$

4. Determine the reduction factors k_r and K_r.

$r_p \simeq 10,000$ ohms and $\mu \simeq 20$ for one triode section of a 12AU7 tube at a plate voltage of 100 volts and a plate current of 6.6 ma

From Eq. (15.18)

$$k_r = K_r = \frac{10,000}{22,700} + \frac{20 \times 3,330}{250,000} + 1$$
$$= 1.71$$

5. Determine the output-voltage increase ΔE_o when the input voltage is increased from 250 volts to 300 volts.

$$K_r = \frac{\Delta E_i}{\Delta E_o}$$

therefore

$$\Delta E_o = \frac{\Delta E_i}{K_r} = \frac{50}{1.71}$$
$$= 29.2 \text{ volts}$$

6. Determine the output resistance R_o and the output impedance Z_o

From Eq. (15.19)

$$R_o = Z_o = \frac{250,000 \times 10,000}{(3330 \times 21) + 246,700 + 10,000}$$
$$= 7,650 \text{ ohms}$$

7. Determine the increase in the output voltage if the load current is decreased 2 ma.

$$\text{Output-voltage change} = -\text{load-current change} \times R_o$$
$$= -(-0.002) \times 7,650$$
$$= 15.3 \text{ volts}$$

NOTE: If the circuit shown in Fig. 15.29c had been used, both K_r and k_r would have been equal to 21.4 and both R_o and Z_o would have been equal to 475 ohms. Therefore, a 50-volt increase in the input voltage would have caused only a 2.34-volt increase in the output voltage. A 2-ma decrease in the load current would have caused the output voltage to increase 0.95 volt.

Series Types of Degenerative Regulators Having One Control Tube. A very common type of series regulator is shown in Fig. 15.30a. $V1$ is usually referred to as the series tube, $V2$ as the control tube, and $V3$ as the reference tube. For a given load the bias of $V2$ is adjusted, by R_{17}, to a value such that its plate voltage places the bias on $V1$ necessary to obtain the desired output voltage. Variations in the output voltage due to input-voltage or load-current changes are amplified by $V2$ and applied to the grid of $V1$. The polarities of the signals applied to the grid of $V1$ are such that they oppose the changes in output voltage; hence, the circuit acts as a regulator. The circuit has different a-c and d-c characteristics due to C_1, C_2, and C_3. C_2 reduces the gas-tube noise voltage across $V3$. The maximum value of C_2 is usually specified by the manufacturer of $V3$. A typical value is 0.1 μf. The values of C_1 and R_1 are those necessary to decouple $V2$ from the ripple voltage appearing at the plate of $V1$. Arbitrarily the required resistance of R_1 might be at least 50 times the reactance of C_1 at the ripple frequency. To minimize phase shift and maximize a-c circuit gain, C_3 should be chosen so that its reactance is much less than the resistance of R_5 at both the ripple frequency and the lowest frequency at which the load current varies.

R_5 usually has a value of several hundred thousand ohms. The value of the sum of R_{16}, R_{17}, and R_{18} should be high enough to avoid excessive current drain. The choice of values for R_{16}, R_{17}, and R_{18} must permit the setting of the proper bias on $V2$. An increase in line voltage increases the g_m of $V2$ due to the increase in the cathode temperature of $V2$. The result is a decrease in the voltage drop across $V2$, which increases the bias on $V1$ and in turn causes a decrease in the regulator output voltage. However, the same increase in line voltage causes an increase in the d-c voltage at the plate of $V1$ which tends to increase the d-c output voltage. The d-c output-voltage change in volts due to the first effect, i.e., the effect of a g_m change in $V2$ causing the quiescent operating voltages across $V2$ and $V1$ to change, is approximately equal to

(a) SIMPLE SERIES REGULATOR

(b) IMPROVED SERIES REGULATOR
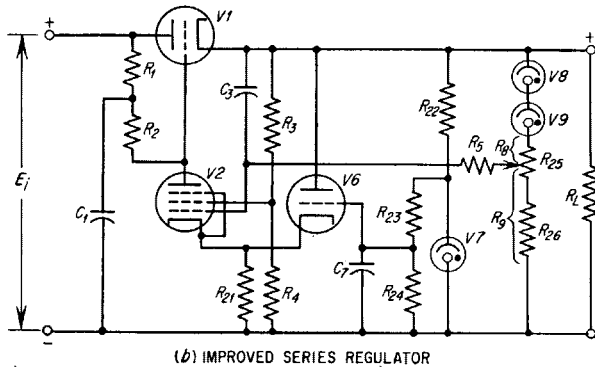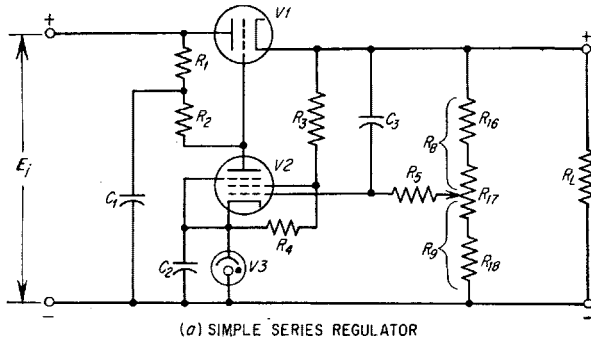
FIG. 15.30. Series types of degenerative regulators with one control tube.

the quantity $-0.01B(R_8 + R_9)/R_9$ where B is equal to the per cent change in line voltage. The d-c output-voltage change due to the second effect is of the opposite polarity and is equal to the d-c voltage change at the plate of the series tube $V1$ (caused by the line voltage change) divided by the d-c reduction factor K_r [see Eq. (15.24)]. If the amplifier gain is very high, K_r is very large and the second effect becomes very small. If this is the case, the first effect predominates and the d-c output voltage will decrease or increase as the line voltage is increased or decreased, respectively. If the amplifier gain is sufficiently low, the second effect predominates. For a regulator of the type shown in Fig. 15.30a, either effect may predominate.

The more complicated regulator shown in Fig. 15.30b provides two improvements over the regulator in Fig. 15.30a. First, because of the addition of $V8$ and $V9$, a larger percentage of the d-c output-voltage variations are coupled into the control tube $V2$. The number and type of voltage-regulating tubes used in this fashion are dependent on the required d-c voltage drop between the output and the grid of $V2$.

It is desirable that the largest portion of this voltage drop be acquired with gas tubes. The second improvement is the circuit containing $V6$ and $V7$ which permits the selection of any desired reference voltage for the cathode of $V2$. The circuit[1] involving $V6$ and $V7$ should be designed so that the cathodes of $V2$ and $V6$ are stabilized to a value of voltage at least 40 volts more negative than the minimum required voltage to ground at the grid of $V1$ (maximum bias on $V1$ which occurs at no-load and high-line voltage). This is based on a minimum voltage drop across $V2$ of 40 volts when at zero bias. An additional improvement is realized because of the fact that essentially a constant current flows through $V7$, $V8$, and $V9$, so that their dynamic resistance is of little importance. This permits the selection of $V7$, $V8$, and $V9$ solely on the basis of their ability to regulate at the same voltage each time the power supply is turned on and the degree to which they are able to regulate at a constant voltage as long as the supply is operating. Tubes such as the 5651 and 6308 are well suited for this particular use. For maximum stability, R_3, R_4, R_{23}, R_{24}, R_{25}, and R_{26} should be temperature-compensated resistors.

If a capacitor is placed across the output terminals of either the power supply shown in Fig. 15.30a or b, the value of k_r will be larger than that determined from Eq. (15.22) and the value of Z_o will be less than that determined by Eq. (15.23).

In the following equations the subscripts of the tube symbols μ, r_p, and g_m refer to the tubes with the same numbers.

For the circuits shown in Fig. 15.30a and b

$$k_r = \frac{\Delta e_i}{\Delta e_o} = \frac{r_{p1}}{R_k} + 1 + \mu_1 + \mu_1 A_2 \tag{15.22}$$

$$Z_o = \frac{r_{p1} + Z_s}{1 + \mu_1 + \mu_1 A_2} \tag{15.23}$$

$$K_r = \frac{\Delta E_i}{\Delta E_o} = \frac{\dfrac{r_{p1}}{R_k} + 1 + \mu_1 + \mu_1 K_2 A_6}{1 + \mu_1 K_1} \tag{15.24}$$

$$R_o = \frac{r_{p1} + R_s(1 + \mu_1 K_1)}{1 + \mu_1 + \mu_1 K_2 A_6} \tag{15.25}$$

where $\quad A_2 =$ a-c gain of $V2$

$$= \frac{\mu_2 R_2}{r_{p2} + R_2 + Z_1(\mu_2 + 1)} \tag{15.26}$$

$$\simeq \frac{g_{m2} R_2}{1 + g_{m2} Z_1} \quad \text{provided } r_{p2} \gg (R_2 + Z_1)$$

$A_6 =$ d-c gain of $V2$ $\tag{15.27}$

$$= \frac{\mu_2(R_1 + R_2)}{R_1 + R_2 + r_{p2} + R_{10}(\mu_2 + 1)}$$

$$\simeq \frac{g_{m2}(R_1 + R_2)}{1 + g_{m2} R_{10}} \quad \text{provided } r_{p2} \gg (R_1 + R_2 + R_{10})$$

$$K_1 = \frac{r_{p2} + R_{10}(\mu_2 + 1)}{R_1 + R_2 + r_{p2} + R_{10}(\mu_2 + 1)} \tag{15.28}$$

$$\simeq \frac{1 + g_{m2} R_{10}}{(R_1 + R_2)/r_{p2} + 1 + g_{m2} R_{10}} \quad \text{provided } \mu_2 \gg 1$$

$$K_2 = \frac{R_9}{R_8 + R_9} \tag{15.29}$$

NOTE: In the case of pentodes operating at low potentials it is frequently difficult to determine r_{p2} and g_{m2} from the published tube characteristics. This makes it necessary to estimate their values.

---

[1] This configuration in which $V2$ and $V6$ share a common cathode resistor makes the value of the voltage to ground at the plate of $V2$, and the grid of $V1$, less sensitive to changes in the g_m of $V2$ caused by filament-temperature changes resulting from line-voltage changes.

R_{10} = in Fig. 15.30a, dynamic resistance of $V3$ (usually between 80 and 200 ohms); in Fig. 15.30b, $r_{p6}/(\mu_6 + 1)$ in parallel with R_{21}.

$$R_k = \frac{\text{d-c output voltage}}{\text{total d-c current through } V1}.$$

R_s = effective internal resistance of power transformer, rectifier, and filter which supplies voltage to regulator.

Z_1 = impedance in the cathode circuit of $V2$. In Fig. 15.30a, Z_1 = impedance of $V3$ in parallel with reactance of C_2. In Fig. 15.30b, $Z_1 = r_{p6}/(\mu_6 + 1)$ in parallel with R_{21}.

Z_s = effective internal impedance of power transformer, rectifier, and filter which supplies voltage to regulator.

Example 15.13

Design a regulated power supply of the type shown in Fig. 15.30a which will deliver 0 to 100 ma at a regulated d-c output of 300 volts for operation from a 105- to 130-volt 60-cycle source. Assume the rectifier tube to be a 5R4GY.

Solution

1. Establish the upper and lower line-voltage limits between which the power supply must regulate.
   The maximum line voltage is 130 volts, and the minimum is 105 volts.
2. Determine the maximum current through the series tubes.
   If the sum of the reference tube and bleeder currents is tentatively assumed to be 25 ma, the maximum current through the series tubes is equal to 100 + 25, or 125 ma.
3. Tentatively determine the number and type of series tubes to be used and the maximum current through each tube.
   The series tubes are usually triodes or triode-connected tetrodes or pentodes. The reason for triode-connecting tetrodes and pentodes is the usual lack of a stable screen-voltage supply. The series tubes ideally should have a high μ for good regulation and low output impedance, and a low static plate resistance to minimize wattage dissipation. If it should be necessary to make a choice between tubes having a low static plate resistance or a high μ, it is usually more economical to choose the tubes with the low static plate resistance since the series tubes are ordinarily fairly large and their size, for a given current rating, is a function of their static plate resistance. A series tube with a low μ can usually be compensated by the use of one or more high-gain control tubes. A few of the most commonly used series tubes of the receiver type are the 6AS7, 6B4, 6L6, 6V6, 6Y6, 6080, 6336, 6337, and 5902. High-power applications make it desirable to consider tubes of the transmitter type such as the 304TL, 4D32, 807, etc.
   There are many possible types and combinations of series tubes which will satisfy the requirement. One choice which tentatively appears satisfactory is the use of the two sections of a 6080 operated in parallel. Each section, therefore, will have to carry a maximum of $125/2$, or 63 ma.
4. From the tube characteristics, determine the minimum voltage drop required across the series tube when operated at or near zero bias and maximum load current.
   From the 6080 tube characteristics, the required voltage drop across each section at zero bias and a current of 63 ma is approximately 20 volts. To allow for variations in tubes, assume that the required tube voltage drop is 30 volts.
   NOTE: With certain types of loads it is possible for there to be a superimposed a-c load current component on the d-c component. As an example, assume that the load current through the series tube should vary between 100 and 150 ma, the average would be 62.5 ma per section. The minimum permissible voltage drop across the series tube, however, would have to be based on the peak current value of 75 ma per section.
5. With the power supply at full load, determine the minimum instantaneous voltage required at the plates of the series tube.

$$\begin{matrix} \text{Minimum instantaneous} \\ \text{voltage required at the} \\ \text{plates of the series tube} \end{matrix} = \text{d-c output volt.} + \begin{matrix} \text{tube drop deter-} \\ \text{mined in step 4} \end{matrix}$$
$$= 300 + 30 = 330 \text{ volts}$$

6. With the power supply at full load, determine the minimum d-c voltage required at the plates of the series tube, taking into account the effect of the ripple voltage superimposed on the d-c voltage E_i.

In this example arbitrarily assume that the power supply employs a single-phase full-wave rectifier and a single-section choke-input filter with a swinging choke having a resistance of 200 ohms and an inductance which varies between 5 and 25 henrys when the load current varies between 125 and 25 ma. If the filter capacitor has a value of 8 μf, the per cent ripple voltage at full load will be approximately 2 per cent as determined from Fig. 15.5. In step 14, where the a-c ripple-reduction factor k_r is determined, the output ripple can be determined and it can be established then whether or not the assumed choke-input filter is adequate. At full load, the d-c voltage supplied to the plates of the series tube less the peak value of the superimposed ripple voltage must be equal to or greater than the minimum instantaneous voltage requirement of 330 volts established in step 5. This d-c voltage can be determined from the following equation.

$$B = \frac{D}{1 - (E/70.7)}$$

where B = minimum d-c voltage required at plates of series tube at full load
D = minimum instantaneous voltage required at plates of series tube as determined in step 5
E = per cent ripple voltage at plates of series tube at full load
Therefore,

$$B = \frac{330}{1 - (2/70.7)}$$
$$= 340 \text{ volts}$$

7. With the power supply at full load, determine the minimum transformer secondary terminal rms voltage required to provide the minimum d-c voltage established in step 6.

For a single-phase full-wave rectifier and a choke-input filter (see Table 15.1), the minimum required rms voltage of one-half the transformer secondary can be determined from the following equation.

$$\begin{pmatrix} \text{Minimum required rms} \\ \text{terminal voltage of one-} \\ \text{half the transformer sec.} \end{pmatrix} = 1.11 \begin{pmatrix} \text{sum of voltage drops across the rec-} \\ \text{tifier and choke at full load + volt-} \\ \text{age from step 6} \end{pmatrix}$$
$$= 1.11(45 + 25 + 340) = 455 \text{ volts}$$

where the d-c voltage drop across the rectifier tube at full load is 45 volts (from 5R4GY tube curves) and the drop across the choke is 200 $\times$ 0.125, or 25 volts.

If a capacitor-input filter is employed, the calculations will be rather lengthy, and it is suggested that reference be made to Sec. 15.4.

8. Determine the d-c voltage at the plates of the series tube at full load when the line voltage is at the upper specified limit.

To minimize wattage dissipation in the series tube, the minimum secondary voltage calculated in step 7 should be obtained with the minimum specified line voltage; therefore, the d-c voltage at the plates of the fully loaded series tube at the maximum line voltage is given by the following equation.

$$F \simeq \frac{KG}{1.11} - H$$

where F = d-c voltage at plates of series tube at high line voltage and full load
K = ratio of maximum line voltage to minimum line voltage
G = voltage from step 7
H = sum of rectifier and choke d-c voltage drops at full load
Therefore,

$$F \simeq \left(\frac{130}{105} \times \frac{455}{1.11}\right) - 70 = 438 \text{ volts}$$

9. Determine the maximum wattage dissipated in each section of the series tube.

$$\begin{pmatrix} \text{Maximum dissipa-} \\ \text{tion in each section} \end{pmatrix} = (\text{voltage from step 8} - \text{output voltage}) \times \text{maximum tube current}$$
$$= (438 - 300) \times (0.063)$$
$$= 8.7 \text{ watts (each section of a 6080 tube is rated at 13 watts max.)}$$

If the maximum wattage dissipation were more than the tube rating, it would be necessary either to change the type of series tube or to parallel additional tubes. If the wattage dissipation were far below the tube rating, the possibility of changing tube type should be investigated.

10. Determine the d-c voltage at the plates of the series tube at maximum line voltage and no load.

At no load the current through the series tube is the sum of the bleeder and reference-tube currents which in this example has arbitrarily been established as being equal to 25 ma. Determine whether the 25-henry inductance of the swinging choke is equal to or greater than the critical inductance.

$$L_c = \frac{\text{maximum rms voltage of } \frac{1}{2} \text{ transformer sec.}}{\text{minimum current through series tube}} \times \frac{1}{1.11 A}$$

where A is determined from Fig. 15.3.

$$L_c = \frac{^{130}\!/_{105} \times 455}{0.025} \times \frac{1}{1.11 \times 1133}$$
$$= 17.9 \text{ henrys}$$

If the value of critical inductance had been larger than the choke selected, the value of critical inductance could have been reduced by increasing the fixed load current through either the bleeder or reference tube.

In this case the inductance of the input choke is actually larger than the critical inductance, therefore the maximum d-c voltage at the plates of the series tubes can be determined from the following equation.

$$\begin{pmatrix} \text{D-C voltage at plates} \\ \text{of series tube at no} \\ \text{load and maximum} \\ \text{line voltage} \end{pmatrix} = \frac{\text{maximum rms terminal voltage of } \frac{1}{2} \text{ sec.}}{1.11} - \begin{array}{l}\text{voltage drop across} \\ \text{rectifier and choke} \\ \text{at no load}\end{array}$$

$$= \frac{^{130}\!/_{105} \times 455}{1.11} - (10 + 5)$$
$$= 492 \text{ volts}$$

NOTE: The d-c voltage drop across the rectifier tube at 25 ma is 10 volts (from 5R4GY tube curves), and the drop across the choke is 200 × 0.025, or 5 volts.

If the inductance of the input choke were smaller than the value of critical inductance, the preceding equation would not apply. It would be necessary to determine the maximum d-c voltage at the plates of the series tubes experimentally or else assume the voltage to be equal to the maximum possible value, which is 1.414 times the maximum rms voltage of one-half the transformer secondary.

If the filter is the capacitor-input type, it is necessary to determine the d-c voltage at the plates of the series tubes with a 25-ma load in accordance with the data in Sec. 15.4.

11. Determine the maximum d-c voltage across the series tube.

(Maximum d-c voltage across series tubes) = voltage from step 10 − d-c output voltage
$$= 492 - 300$$
$$= 192 \text{ volts}$$

This is below the maximum plate-voltage rating of a 6080 tube.

12. Determine if the control tube can provide the necessary bias on the series tube at maximum line voltage and no load.

At no load and maximum line voltage the per cent ripple voltage at the plates of the series tubes is equal to approximately 0.42 per cent of 492 volts, or 2.1 volts rms. This is based on $L = 25$ henrys, $C = 8 \mu f$, and Fig. 15.5. The peak voltage at the plates of the series tube will therefore be equal to 492 + (1.41 × 2.1), or 495 volts. By referring to the series tube characteristics, note that the required bias on a 6080 tube for a 195-volt drop and the no-load current of $^{25}\!/_{2}$, or 13 ma, per section is approximately −116 volts. The voltage between the grids of the series tube and ground at maximum line voltage and no load will therefore be 184 volts since the cathode voltage will be equal to 300 volts and the bias will be equal to −116 volts. Since the voltage drop across the control tube $V2$ (see Fig. 15.30a) at zero bias can be expected to be as low as 30 or 40 volts, the voltage-regulator tube $V3$ should have a nominal voltage somewhat less than 184 − 40 volts in order to ensure that the control tube can place the desired bias on the series tube at no load. Assum-

ing that $V3$ is an OB2 tube (108-volt nominal), the regulator circuit is able to place a minimum voltage of $108 + 40$ volts or less between the grids of the series tubes and ground. This is equivalent to a bias of -152 volts on the series tube. This ensures that the regulator will be able to supply the maximum required bias to the series tube.

13. Design the control section.

The regulator section has been specified to be of the type of Fig. 15.30a. $V1$ has been established as being a 6080 tube, and $V3$ is an OB2. Let $V2$ be a 6AH6. Typical values for the resistors and capacitors in the control circuit are shown in Fig. 15.31. The sum of R_3 and R_4 was determined as being that value necessary to place a continuous 25-ma load on the power supply, thereby reducing the value of critical inductance for the input choke. In accordance with accepted good practice, parasitic resistors have been placed in the grid and plate leads of the paralleled sections of the series tube. The remaining circuit values are not critical and are typical for this type of regulator.
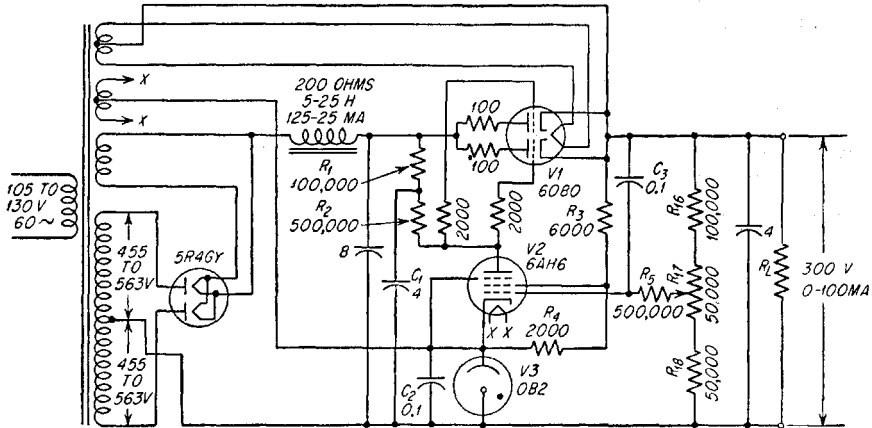
FIG. 15.31. Regulated power supply using a regulator of the type shown in Fig. 15.30a.

14. Determine k_r, K_r, Z_o, and R_o at full load.

Assume the plate voltage on the series tube to be the average of the maximum and minimum values at full load.

From Eq. (15.22)

$$k_r = \frac{190}{2,400} + 1 + 2 + (2 \times 345)$$

= 693 (Value of r_{p1} = dynamic resistance of two halves of the series tube and their parasitic resistors in parallel. In determination of A_2, constants g_{m2}, r_{p2}, and Z_1 have been assumed to be equal to 1,200 μmhos, 1 megohm, and 200 ohms, respectively.)

From Eq. (15.24)

$$K_r = \frac{190/2,400 + 1 + 2 + (2 \times 0.33 \times 391)}{1 + (2 \times 0.67)}$$

= 112 (In determination of K_1 and A_6, constants g_{m2}, r_{p2}, and R_{10} have been assumed to be equal to 1,200 μmhos, 1 megohm, and 200 ohms, respectively.)

From Eq. (15.23)

$$Z_o = \frac{190 + Z_s}{1 + 2 + (2 \times 345)}$$

$$= 0.27 + \frac{Z_s}{693} \quad \text{ohms}$$

Z_o will vary with frequency since Z_s varies with frequency. The value of Z_s is essentially equal to the reactance of the input choke in parallel with the reactance of the 8-μf capacitor and must be known at the particular frequency at which Z_o is to be determined.

From Eq. (15.25)

$$R_o = \frac{190 + R_s[1 + (2 \times 0.67)]}{1 + 2 + (2 \times 0.33 \times 391)} = \frac{190 + 2.34R_s}{261}$$

and R_s = sum of choke resistance (200 ohms), rectifier resistance (300 ohms), and effective transformer resistance as seen at secondary terminals (assume 100 ohms)

Therefore,

$$R_o = 6.1 \text{ ohms}$$

Summary

Since k_r is equal to 693, the output ripple voltage at low line voltage and full load is equal to 2 per cent of 340 volts (see step 6) divided by 693, which is equal to 9.8 mv rms. The capacitor shunting the output will further reduce this value. Neglecting changes in the g_m of $V2$ caused by line-voltage changes and considering that K_r is equal to 112, a change from low line voltage to high line voltage when the power supply is at full load would cause an increase in the d-c output of $(438 - 340)/112$, or 0.88 volt. The reduction in the d-c output voltage due to the change in the g_m of $V2$ caused by the increase in line voltage is given by $-0.01B(R_8 + R_9)/R_9$, where B is the per cent change in line voltage. In this example, the increase in line voltage from 105 to 130 volts represents an increase of 23.8 per cent. Since the potentiometer R_{17} controlling the bias on $V2$ will be set at approximately 108 volts, $(R_8 + R_9)/R_9$ will approximately be equal to 3. Therefore the incremental change in the output voltage due to the cathode temperature change of $V2$ would be equal to approximately -0.71 volt. The net effect of a 0.88-volt increase and a 0.71-volt decrease is a 0.17-volt increase.

Series-type Degenerative Regulator with Two Control Tubes. The circuit in Fig. 15.32 is a typical degenerative regulator which contains a two-stage amplifier and two
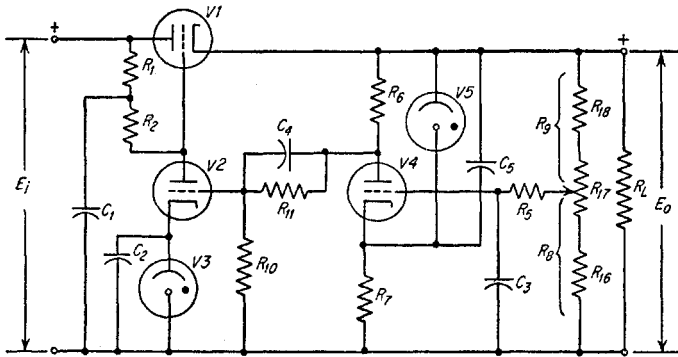
FIG. 15.32. Series type of degenerative regulator with two control tubes.

reference tubes. Operationally the circuit is the same as that shown in Fig. 15.30a with the exception of the additional amplifier stage and reference tube. Considerably more precaution must be exercised in the design of a two-stage regulator since there is the possibility of oscillation within the regulator. The design of the amplifier must be similar to the design of any other feedback amplifier (see Sec. 18). The advantage of a two-stage amplifier is that it is possible to obtain higher gain in the amplifier and hence both an increase in ripple reduction and an improvement in the regulation for load-current changes. Line-voltage variations will change the transconductances of both $V2$ and $V4$, which consequently changes their quiescent operating potentials. The resulting changes in the plate potentials across $V2$ and $V4$ tend to cause a change in the d-c output voltage. This phenomenon in $V4$ has the largest effect on the output voltage and, neglecting the effects of the changing d-c voltage at the plate of $V1$ caused by line-voltage changes, there would be an increase or decrease

in the power-supply output voltage associated with an increase or decrease, respectively, in the line voltage. The approximate change in the d-c output voltage due to this effect can be determined by the equation given for the circuit shown in Fig. 15.30a. The only difference is that the sign of the d-c output change is positive instead of negative.

The a-c and d-c characteristics of the regulator differ because of the circuit capacitors. The design considerations for C_1, C_2, and C_3 are the same as for the circuit shown in Fig. 15.30a. The reactance of C_4 should be much less than the resistance of R_{10} at both the ripple frequency and the lowest frequency at which the load current varies; however, if R_{11} is small compared to R_{10}, C_4 can be eliminated. C_5 has the same maximum size limitation as C_2. If a capacitor is placed across the output terminals of the power supply, the value of k_r will be larger than that determined from Eq. (15.30) and the value of Z_o will be less than that determined from Eq. (15.31). In the following equations, the subscripts of the tube symbols μ and r_p refer to the tubes with the same numbers.

For the circuit shown in Fig. 15.32

$$k_r = \frac{\Delta e_i}{\Delta e_o} = \frac{r_{p1}}{R_k} + 1 + \mu_1 + \mu_1 A_2 A_4 \qquad (15.30)$$

$$Z_o = \frac{r_{p1} + Z_s}{1 + \mu_1 + \mu_1 A_2 A_4} \qquad (15.31)$$

$$K_r = \frac{\Delta E_i}{\Delta E_o} = \frac{\dfrac{r_{p1}}{R_K} + 1 + \mu_1 + \mu_1 K_2 A_6 A_8}{1 + \mu_1 K_1} \qquad (15.32)$$

$$R_o = \frac{r_{p1} + R_s(1 + \mu_1 K_1)}{1 + \mu_1 + \mu_1 K_2 A_6 A_8} \qquad (15.33)$$

where $\quad A_4$ = a-c gain of $V4$

$$= \frac{(\mu_4 + 1)R_{10}R_6}{r_{p4}(R_{10} + R_6) + R_{10}R_6} \qquad (15.34)$$

A_8 = d-c gain from cathode of $V4$ to grid of $V2$

$$= \frac{(\mu_4 + 1)R_{10}R_6}{R_6(R_{10} + R_{11}) + r_{p4}(R_{10} + R_6 + R_{11})} \qquad (15.35)$$

NOTE: For definitions of A_2, A_6, K_1, K_2, R_k, and Z_s refer to Eqs. (15.26) to (15.29).

15.8c. Degenerative Voltage Regulator with A-C and D-C Compensation. In certain applications it may be desirable to obtain a specific performance characteristic which is superior to that which can be obtained from the typical degenerative-type regulator. For example, it might be that for a constant line voltage the power-supply d-c output resistance should be equal to zero, thereby ensuring a constant output voltage even though the d-c load current should change. It is possible to obtain this performance from a degenerative regulator which has been modified by the introduction of the proper amount of d-c compensating voltage into the amplifier circuit.

With the exception of R_{19}, R_{20}, C_6, and the more desirable voltage reference source for the cathode of $V2$, the circuit in Fig. 15.33 is identical with that of Fig. 15.30a. The addition of R_{19} and R_{20} transforms the degenerative regulator to one with both a-c and d-c compensation controls. C_6 has been added to minimize the a-c signals on the d-c compensation input. The modification permits the injection of portions of both the a-c and d-c input voltages, that is, e_i and E_i, respectively, into the grid of $V2$. Potentiometers R_{19} and R_{20} are the a-c and d-c controls, respectively, and their values are usually hundreds of thousands of ohms. Proper adjustment of the compensation controls will prevent those changes in the output voltage which would otherwise exist in a strictly degenerative regulator under conditions of changing input voltage and changing load. The adjustments of both the a-c and d-c compensation

controls are dependent on whether they are adjusted for input-voltage changes or load-current changes. Proper circuit operation for a-c and d-c load-current changes is dependent on the input voltage source having an internal impedance and resistance, respectively, across which a voltage can be developed because of the changing load current. These requirements are usually satisfied by the presence of the power transformer, rectifier, and filter. Without this changing voltage, it would be impossible
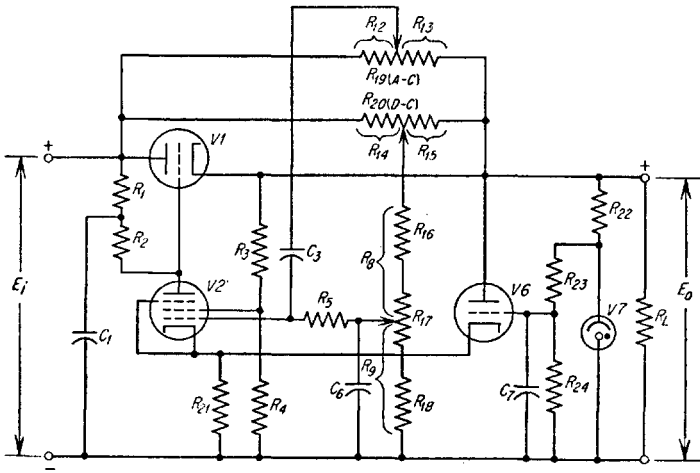
FIG. 15.33. Series type of degenerative regulator with a-c and d-c compensation controls.

to obtain the necessary compensating voltage. The operating conditions theoretically possible for the circuit of Fig. 15.33 are as follows:

1. *For a constant load current* it is possible to obtain zero a-c rectifier ripple voltage on the d-c output voltage if R_{19} is adjusted so that

$$\frac{R_{13}}{R_{12} + R_{13}} = \frac{1}{\mu_1 A_2} \tag{15.36}$$

then

$$k_r = \frac{\Delta e_i}{\Delta e_o} = \infty$$

and

$$Z_o = \frac{r_{p1}}{\mu_1(A_2 + 1)} \tag{15.37}$$

2. *For a constant line voltage,* it is possible to obtain an a-c output impedance of 0 ohms if R_{19} is adjusted so that

$$\frac{R_{13}}{R_{12} + R_{13}} = \frac{r_{p1} + Z_s}{\mu_1 A_2 Z_s} \tag{15.38}$$

then

$$k_r = \frac{\Delta e_i}{\Delta e_o} = 1 - \frac{Z_s}{R_k} - \frac{\mu_1 Z_s(A_2 + 1)}{r_{p1}} \tag{15.39}$$

and

$$Z_o = 0$$

3. *For a constant line voltage,* a d-c output resistance of 0 ohms can be obtained by adjusting R_{20} so that

$$\frac{R_{15}}{R_{14} + R_{15}} = \frac{r_{p1} + R_s(1 + \mu_1 K_1)}{\mu_1 A_6 K_2 R_s} \tag{15.40}$$

then

$$K_r = \frac{\Delta E_i}{\Delta E_o} = 1 - \frac{R_s}{R_k} - \frac{\mu_1 R_s}{r_{p1}}(1 + K_2 A_6 - K_1) \tag{15.41}$$

and

$$R_o = 0$$

4. *For a constant load current and a slowly varying line voltage,* a constant output voltage can be obtained if the filament voltage of $V2$ is stabilized[1] and provided that R_{20} is adjusted so that:

$$\frac{R_{15}}{R_{14} + R_{15}} = \frac{1 + \mu_1 K_1}{\mu_1 A_6 K_2} \tag{15.42}$$

then

$$K_r = \frac{\Delta E_i}{\Delta E_o} = \infty$$

and

$$R_o = \frac{r_{p1}}{\mu_1(1 + K_2 A_6 - K_1)} \tag{15.43}$$

Since there are separate a-c and d-c compensation controls, it is possible to obtain simultaneously both an ideal a-c and d-c operating condition, viz., 1 and 3, 1 and 4, 2 and 3, or 2 and 4. For the same reason it is impossible to realize simultaneously two ideal a-c or two ideal d-c conditions. The normal procedure is to evaluate the desirability of the obtainable conditions and then adjust the control potentiometers for the operation most applicable, which in some instances may require a compromise between two different ideal operating conditions.

15.8d. Special Circuits and Measurements for Regulated Power Supplies. There are many possible modifications in the basic degenerative series regulator circuits of Fig. 15.30. A few of those which are most commonly used are treated in the following paragraphs. Details for measuring regulation and output impedances have also been included.

Pentodes as Series Tubes. In general, a pentode requires a smaller control voltage and, for a given load current, can be operated at a lower plate-to-cathode voltage drop than a triode. In addition, a pentode will provide a high degree of attenuation to voltage changes at its plate because of its high dynamic plate resistance and its high transconductance. For these reasons, pentodes frequently make desirable series tubes. Normally the difficulty in using a pentode as a series tube is that of obtaining a stable screen-grid voltage source. One method is to provide a high-impedance filter to supply a well-filtered voltage to the screen grid. The plate voltage on the pentode can then be very poorly filtered since the pentode will provide the desired ripple reduction. In this manner the filter requirements are minimized.

A necessary precaution in the use of a pentode which has separate voltage sources for the screen grid and plate is the incorporation of a means to limit the screen-grid current in the event the plate voltage is removed. This can be accomplished by fusing the screen-grid circuit, using a screen-grid limiting resistor, or by interlocking the screen voltage with the plate voltage.

Paralleling Series Tubes with Fixed Resistors. If the load on a regulated power supply is nearly constant, frequently it is possible to replace one or more of several paralleled series tubes with fixed resistors and still obtain satisfactory regulation. The value of the resistor replacing any given tube should be equal to the nominal voltage drop across the tube divided by its plate current. Replacing series tubes with fixed resistors will destroy the ability of the power supply to regulate at no load.

Series Regulators Which Do Not Require Reference Tubes. If a stable negative voltage source is available, the cathode of the control tube can be grounded. The grid for the control tube is then connected to a potentiometer which is in a divider network connected between the output voltage and the stable negative supply (see Fig. 15.34).

[1] There is no setting of R_{20} which will permit the realization of condition 4 unless for a given line-voltage change in the uncompensated regulator, the effect of the variation in the filament voltage on the d-c output voltage is less than the effect of the associated d-c voltage change at the plate of the series tube on the d-c output voltage. The filament voltage can be completely stabilized by either applying a regulated d-c voltage to the filament or by using a special constant-voltage filament transformer.

Lightweight Power-supply Techniques. Aircraft power-supply design is influenced by the need for high efficiency, small size, light weight, and good power factor. The rectifier circuit most commonly used to obtain these features is a three-phase bridge rectifier. The use of selenium or especially silicon diode rectifiers also helps to reduce the size of the rectifier. For this type of rectifier, the rms ripple voltage is about 4 per cent of the d-c output voltage; consequently, little or no filtering is required. If pentode series regulator tubes are used, the efficiency can be improved by peak-detecting the ripple and placing the screens of the series tubes at a higher d-c voltage than the plates. (This technique is illustrated in Fig. 15.34 although there is no advantage in this particular circuit due to the very small ripple voltage.) Figure 15.34 is a typical aircraft power supply. A resistor has been placed in parallel with the series regulator tube to reduce the plate dissipation. Current through this resistor at high line and minimum load must be less than the minimum load current in order to ensure current flow through the series tube and thereby permit regulation.
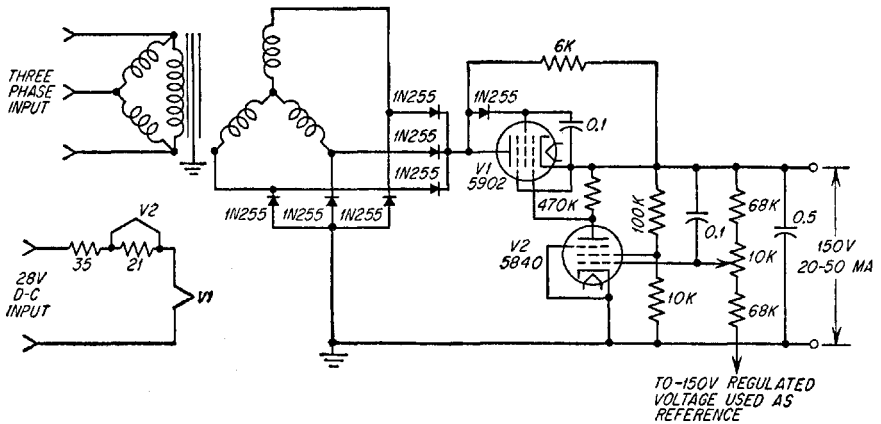
FIG. 15.34. Typical three-phase lightweight power supply.

Measurement of D-C Regulation and D-C Output Impedance. To make a precise measurement of the d-c regulation of a power supply, usually it is necessary to have a stable reference voltage such as batteries or another regulated supply of approximately the same output voltage. The procedure is to connect a meter between the reference voltage and the output voltage of the supply to be tested and note the difference in voltages as the power supply under test is loaded and unloaded. The change in the meter reading as the power supply is loaded and unloaded represents the difference between the no-load and full-load voltages. The per cent d-c regulation is given by Eq. (15.44).

$$\% \text{ d-c regulation} = \frac{\text{no-load voltage} - \text{full-load voltage}}{\text{full-load voltage}} \times 100 \qquad (15.44)$$

The d-c output impedance is given by Eq. (15.45).

$$\text{D-C output impedance} = \frac{\text{no-load voltage} - \text{full-load voltage}}{\text{full-load current}} \qquad (15.45)$$

Measurement of a Power-supply Output Impedance. The output impedance of a power supply at the frequency of any a-c load is equal to the a-c voltage (caused by

the a-c load) which is superimposed on the power-supply d-c output voltage divided by the a-c load current. The a-c voltage superimposed on the d-c output voltage can easily be measured with an oscilloscope. The a-c load current can be determined by measuring the a-c voltage across a known resistor in series with the ground side of the load and dividing the measured voltage by the value of the resistor.

An a-c load could be provided by placing a low-voltage transformer winding in series with the load resistor.

15.9. Zener Diodes[1] as Voltage-regulator Elements. The Zener characteristic of silicon junction diodes permits the use of these diodes as voltage reference elements in a manner identical to the applications utilizing gas-tube voltage-regulating elements. As discussed in Sec. 12.3a, the Zener region is the voltage region in which the back resistance of the diode is sharply reduced from several megohms to a very few ohms for a relatively small change in the back voltage. The back a-c resistance, i.e., the effective resistance for incremental changes in the back voltage, in the Zener region is typically in the order of 5 to 15 ohms for low Zener voltages and in the order of 1,000 to 2,000 ohms for high Zener voltages. Used as a regulating element, the Zener region establishes the voltage-regulating level, and the diode a-c resistance in the Zener region establishes the effectiveness of the diode as a regulator for voltage variations in the source and for variations in any load shunting the diode.

15.9a. Zener Voltage Ratings. Zener diodes, i.e., silicon junction diodes which are being utilized for their Zener characteristic, can be obtained with ratings from a few volts to approximately one hundred volts. Although the Zener region cannot be exactly controlled in production, diodes can be obtained from most manufacturers having Zener ratings in accordance with the standard RETMA values, such as is the case with resistors and capacitors, with accuracies which are within 5 or 10 per cent of the nominal values. There is no standard method of defining the Zener voltage. Ordinarily the Zener region is much more sharply defined for those diodes having large Zener voltages than for those with small Zener voltages. For this reason, the Zener voltage of a diode with a large rating, e.g., 100 volts, is ordinarily defined as the back voltage which causes the back current to increase to 200 or 300 microamperes, whereas in those diodes having a Zener voltage of only a few volts, the Zener voltage is often defined as the back voltage which causes the back current to increase to 2 or more milliamperes.

15.9b. Temperature Stability of the Zener Voltage. The Zener voltage rating of a silicon junction diode increases with an increase in temperature. Typical temperature coefficients are:

$$\simeq 0.13\,\%/\text{C}° \text{ for a Zener voltage of 100 volts}$$
$$\simeq 0.08\,\%/\text{C}° \text{ for a Zener voltage of 20 volts}$$
$$\simeq 0.04\,\%/\text{C}° \text{ for a Zener voltage of 6 volts}$$

For Zener voltages less than approximately 5 volts, the temperature coefficient may be either positive or negative.

In the forward direction the voltage drop across a silicon junction diode decreases with an increase in temperature. Since the forward temperature coefficient is in the opposite direction to the Zener temperature coefficient, increased temperature stability can be achieved by adding a silicon junction diode in series and in opposite polarity with the silicon junction diode which is being utilized in the Zener application. As an example of the temperature stability which can be achieved with this technique, the Hoffman Semi-Conductor Division of Hoffman Electronic Corporation reference diode IN429 has a temperature coefficient which is in the order of 0.007%/C°.

[1] Much of the information concerning Zener diodes was obtained from H. F. Schoemehl of the Hoffman Semi-Conductor Division of the Hoffman Electronic Corporation.

15.9c. Typical Circuit Configurations. The simplest circuit utilizing a Zener diode as a reference element is identical to that shown for the gas-tube voltage regulator except that the Zener diode is substituted for the tube (see Fig. 15.24). The value of the dropping resistor must be such as to limit the maximum diode current to a value within the diode's dissipation rating when the source voltage is the maximum and the load current is at a minimum. Low voltage Zener diodes have no characteristic analogous to the ionization potential in gas tubes. However in certain high-voltage Zener diodes, there is a requirement to exceed the Zener voltage in order to initiate the Zener action.

A Zener diode can also be substituted for any of the voltage-regulator applications shown in Figs. 15.30 through 15.33.

A voltage reference source which approaches the stability of a standard cell can be obtained by utilizing a two-stage Zener regulator in which the diodes and circuit resistances are maintained in a temperature-controlled oven. For example, a typical circuit would operate from an unregulated d-c source which is considerably larger than the desired d-c reference output. The first regulator stage would be of the type shown in Fig. 15.24 and would utilize either one or a series combination of Zener diodes which would provide an output intermediate between the unregulated source and the desired voltage reference output of the second regulator. The second regulator stage would be identical to the first stage except that it would employ a Zener diode having the desired voltage-output rating. If a calibrated output load resistance or divider is used, it should be constant and for best results should also be contained in the temperature-controlled oven.

Filters

16.1. Constant-k and m-derived Filters. Constant-k and m-derived filters represent one unique family of filters and are sometimes referred to as Zobel[1] filters.

16.1a. Filter Classifications. The four basic filter classifications are as follows:

Low-pass filter. Transmission band extending from zero frequency to the frequency of cutoff f_c and an attenuation band extending from f_c to infinite frequency

High-pass filter. Attenuation band extending from zero frequency to the frequency of cutoff f_c and a transmission band extending from f_c to infinite frequency

Bandpass filter. Transmission band extending from the lower cutoff frequency f_1 to the upper cutoff frequency f_2 and attenuation bands extending from zero frequency to f_1 and from f_2 to infinite frequency

Band-elimination filter. Transmission bands extending from zero frequency to the lower cutoff frequency f_1 and from the upper cutoff frequency f_2 to infinite frequency and an attenuation band extending from f_1 to f_2

16.1b. Attenuation and Transmission Bands. In the attenuation band the characteristic impedance of the filter is a pure reactance, and consequently, if the load terminating the filter is equal to the characteristic impedance of the filter, the generator will be working into a pure reactance and no power will be transmitted to the load. In a practical case where the filter is terminated with a pure resistance, there is an impedance mismatch between the filter and the load, causing the input impedance to the filter to be slightly resistive. In this case the filter does transmit a limited amount of power in the attenuation band.

In the transmission band the characteristic impedance of the filter is a pure resistance; consequently, if the load is equal to the characteristic impedance of the filter, the generator will be working into a pure resistance and all the power will be transmitted to the load since the ideal filter contains only nondissipative reactive elements. In a practical application the filter elements contain resistance and will dissipate a small percentage of the power delivered by the generator.

16.1c. Characteristic Impedance, Image Impedance, and Image Transfer Constant. The *characteristic impedance* Z_o of a symmetrical filter is the impedance with which the filter must be terminated in order to see the same impedance Z_o at the input terminals. Characteristic impedance is also known as the surge impedance. The characteristic impedance of any network can be determined by Eq. (16.1).

$$Z_o = \sqrt{Z_{oc}Z_{sc}} \qquad (16.1)$$

where Z_{oc} = input impedance with output terminals open-circuited

Z_{sc} = input impedance with output terminals short-circuited

A network is operating on an *image impedance* basis if the input impedance to the terminated network is equal to the driving circuit impedance and if the output impedance of the driven network is equal to the terminating impedance.

The *image transfer constant,* $\theta = \alpha + j\beta$, is defined as being equal to one-half the natural logarithm of the complex ratio of the volt-amperes entering to those leaving the network when the network is terminated in its image impedance. An example

[1] Otto J. Zobel, Theory and Design of Uniform and Composite Electric Wave Filters, *Bell Telephone System Tech. J.*, vol. 2, p. 1, January, 1923.

of a network operating on an image-impedance basis is shown in Fig. 16.1. Equations (16.2) and (16.3) are applicable to networks operating on this basis.

$$\theta = \alpha + j\beta = \frac{1}{2} \times \log_e \frac{e_1 i_1}{e_2 i_2} = \log_e \frac{i_1}{i_2} \sqrt{\frac{Z_{11}}{Z_{22}}} \tag{16.2}$$

$$e^\theta = e^{\alpha+j\beta} = \frac{i_1}{i_2} \sqrt{\frac{Z_{11}}{Z_{22}}} \tag{16.3}$$

where α = image attenuation constant, nepers (1 neper = 8.686 db)

β = image phase constant, radians

Z_{11} = image input impedance

Z_{22} = image output impedance

If the network is symmetrical, Eq. (16.4) is also applicable.

$$e^\theta = e^{\alpha+j\beta} = \frac{e_1}{e_2} = \frac{i_1}{i_2} \tag{16.4}$$

16.1d. Filter Configurations. The filter configurations to be discussed are shown
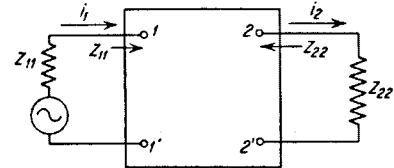
Fig. 16.1. A network terminated with its image impedances.

in Fig. 16.2. The two basic configurations are the T and the π, and they are shown in both the balanced and unbalanced form.
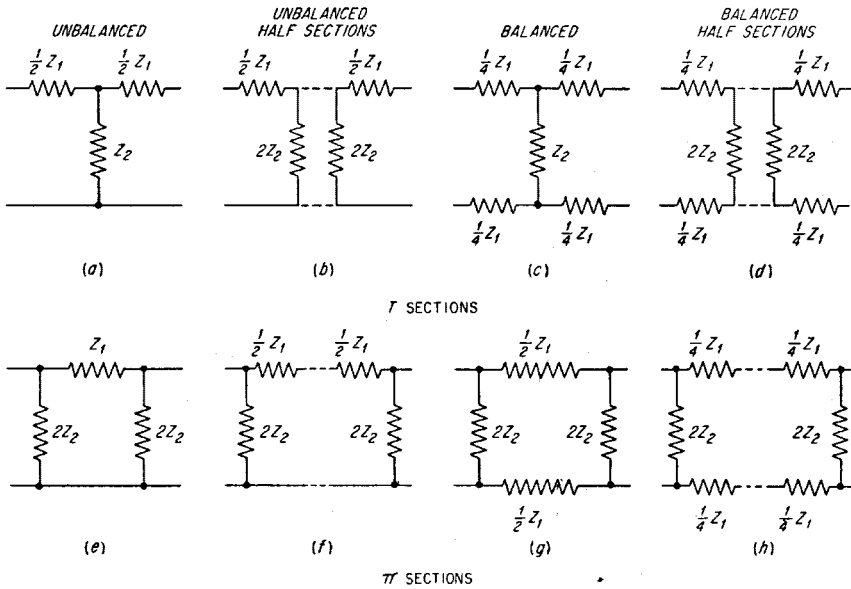
Fig. 16.2. Network configurations.

The characteristic impedances of symmetrical T and π balanced or unbalanced sections are given by Eqs. (16.5) and (16.6), respectively.

$$Z_T = \sqrt{Z_1 Z_2 \left(1 + \frac{Z_1}{4Z_2}\right)} \tag{16.5}$$

$$Z_\pi = \sqrt{\frac{Z_1 Z_2}{1 + \frac{Z_1}{4Z_2}}} \tag{16.6}$$

An examination of the two equations shows that the characteristic impedances of the two networks are a function of frequency, and hence, it is difficult to terminate either exactly with a simple network. If the terminations are assumed to be resistive and equal to $\sqrt{Z_1 Z_2}$, the terminating resistance R will be equal to $\sqrt{L/C}$, provided the shunt and series arms are inductive and capacitive, respectively, or capacitive and inductive, respectively. Equations (16.5) and (16.6) can then be rewritten as follows:

$$Z_T = R \sqrt{1 + \frac{Z_1}{4Z_2}} \qquad (16.7)$$

$$Z_\pi = \frac{R}{\sqrt{1 + \dfrac{Z_1}{4Z_2}}} \qquad (16.8)$$

16.1e. Constant-k Filters. In a constant-k filter, the product of the series and shunt impedances is a constant independent of frequency.

$$k = \sqrt{Z_1 Z_2} = R \qquad (16.9)$$

where R = terminating resistance

Each inductance in the shunt and series arms is associated with a capacitor in the series and shunt arms, respectively. Assuming nondissipative elements, Eq. (16.10) indicates the required relation between the elements.

$$k = \sqrt{\frac{L_1}{C_2}} = \sqrt{\frac{L_2}{C_1}} = \cdots = R \qquad (16.10)$$

The frequencies of maximum attenuation are those farthest removed from the transmission band; e.g., in a constant-k bandpass filter the frequencies of maximum attenuation are at zero and infinite frequency. If more attenuation is desired adjacent to the transmission band, an m-derived type of filter must be employed.

16.1f. m-derived Filters. There are two types of m-derived filter sections, *series-derived* and *shunt-derived*.

Series m-derived Sections. The characteristic impedance of a symmetrical T network is ordinarily referred to as the *mid-series image impedance* because it is that impedance which is measured if a long ladder network is opened in the middle of one of the series arms as shown in Fig. 16.3a.

If the series arm of the T network shown in Fig. 16.4a is modified by the factor m as shown in the network of Fig. 16.4b, the characteristic impedance of the two networks can be made the same if the shunt arm of the network shown in Fig. 16.4b is made equal to $(Z_2/m) + Z_1(1 - m^2)/4m$. The quantity m is a real number which varies between 0 and 1. The series m-derived section thus obtained will provide a proper impedance match for any other series m-derived section or for a constant-k section of the T type.

Shunt m-derived Sections. The characteristic impedance of a symmetrical π network is ordinarily referred to as the *mid-shunt image impedance* because it is that impedance which is measured if a long ladder network is opened in the middle of one of the shunt arms as shown in Fig. 16.3b.

If the shunt arms of the π network shown in Fig. 16.5a are modified by the factor $1/m$ as shown in the network of Fig. 16.5b, the characteristic impedances of the two networks can be made the same if the series arm of the network in Fig. 16.5b is made equal to mZ_1 in parallel with $Z_2 4m/(1 - m^2)$. The shunt m-derived section thus obtained will provide a proper impedance match for any other shunt m-derived section or for a constant-k section of the π type.
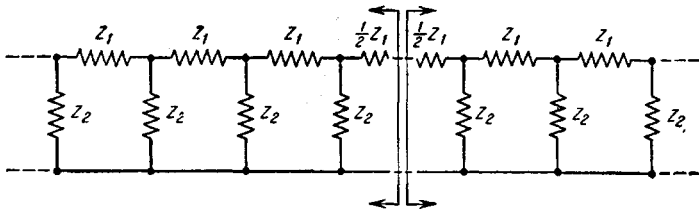
(a) POINT OF MID SERIES IMAGE IMPEDANCE MEASUREMENT
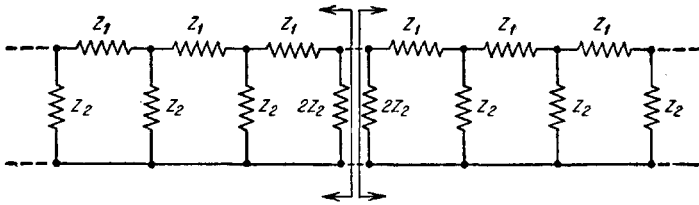
(b) POINT OF MID SHUNT IMAGE IMPEDANCE MEASUREMENT

FIG. 16.3. Infinitely long ladder networks.
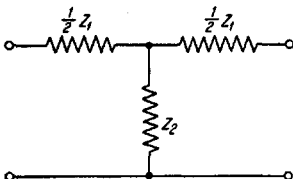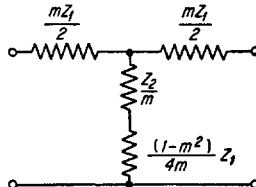
(a) T SECTION   (b) SERIES m-DERIVED T SECTION

FIG. 16.4. T networks which have the same characteristic impedances but different mid-shunt impedances.
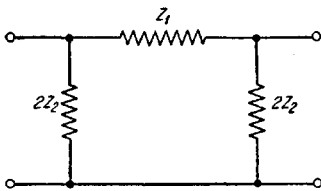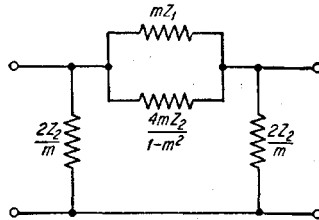
(a) π SECTION   (b) SHUNT m-DERIVED π SECTION

FIG. 16.5. π networks which have the same characteristic impedances but different mid-series impedances.

Characteristics of m-derived Sections. In either a low- or high-pass filter, an m-derived section will theoretically introduce infinite attenuation at some frequency f_∞ in the attenuation band. If the filter is of the bandpass type, an m-derived section will theoretically introduce infinite attenuation at two frequencies $f_{1\infty}$ and $f_{2\infty}$ in the lower and upper attenuation bands, respectively. In a band-reject filter, an m-derived section will theoretically introduce infinite attenuation at two frequencies, $f_{1\infty}$ and $f_{2\infty}$, in the attenuation band. The smaller the value of m, the closer the frequencies of

maximum attenuation will be to the transmission bands. The attenuation at f_∞, $f_{1\infty}$, and $f_{2\infty}$ would be infinite with nondissipative elements; however, because of resistance in the elements, the attenuation has a finite value and becomes less and less as m is reduced and as the frequencies of maximum attenuation approach the transmission band.

16.1g. Filter Terminations. The characteristic impedances of full T and π sections vary considerably with frequency because of the change in values of Z_1 and Z_2 as shown by Figs. 16.6 and 16.7 which are plots of Eqs. (16.7) and (16.8), respectively. The fact that the characteristic impedances are not independent of frequency presents no problem in combining any number of filter sections as long as the characteristic impedances of any two sections connected together are identical at all frequencies.

Since the characteristic impedances of a T section of the constant-k type and of a T section of the series m-derived type are equal, these two types of filters can be connected together. Also, since the characteristic impedances of a π section of the
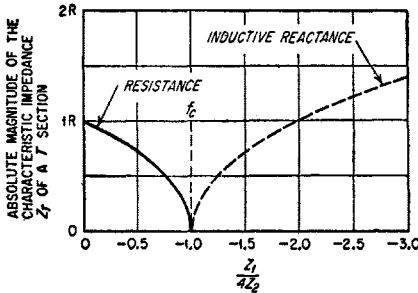
FIG. 16.6. Characteristic impedance Z_T of a T section versus $Z_1/4Z_2$.
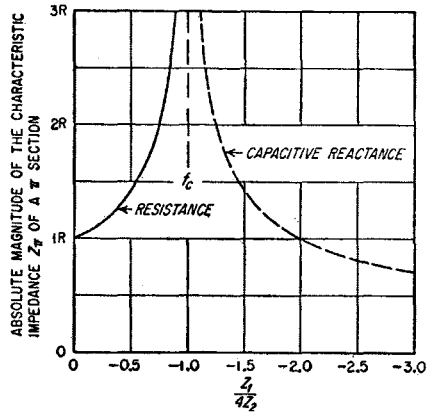
FIG. 16.7. Characteristic impedance Z_π of a π section versus $Z_1/4Z_2$.

constant-k type and of a π section of the shunt m-derived type are equal, these two types can be connected together.

It is possible to split T and π sections into half sections as shown in Figs. 16.2b and 16.2f. In the case of the split T section, the characteristic impedance looking from the mid-section in either direction, assuming proper terminations, is termed the mid-shunt impedance of the T section. Although the characteristic impedances of constant-k and m-derived T sections are identical, the mid-shunt impedances are not. Also the mid-series impedances of constant k and m-derived π sections are different even though the input impedances to the full sections are identical. These particular characteristics of filter sections are of great value since the mid-shunt impedance of an m-derived T section and the mid-series impedance of an m-derived π section can be made to be very nearly equal to a constant resistance through the transmission band, and consequently these half sections can be used as satisfactory matching sections between full filter sections and resistive loads.

In general, *series m-derived T sections and shunt m-derived π sections are never divided into half sections except when used as terminating half sections.* The mid-shunt impedance Z'_T of the series m-derived T section and the mid-series impedance Z'_π of the shunt m-derived π section can be made to approximate the terminating resistance R over most of the passband.

The mid-shunt impedance Z_T' of a series m-derived T section can be determined from Eq. (16.11).

$$Z_T' = Z_\pi \left[1 + (1 - m^2) \frac{Z_1}{4Z_2} \right] \qquad (16.11)$$

The ratio R/Z_T' indicates the mismatch between R and Z_T' and is given by Eq. (16.12).

$$\frac{R}{Z_T'} = \sqrt{1 + \frac{Z_1}{4Z_2}} \; \frac{1}{1 + (1 - m^2) \dfrac{Z_1}{4Z_2}}$$

$$(16.12)$$

The mid-series impedance Z_π' of a shunt m-derived π section can be determined from Eq. (16.13).

$$Z_\pi' = Z_T \frac{1}{1 + (1 - m^2) \dfrac{Z_1}{4Z_2}} \qquad (16.13)$$

The ratio Z_π'/R indicates the mismatch between Z_π' and R and is given by Eq. (16.14).

$$\frac{Z_\pi'}{R} = \sqrt{1 + \frac{Z_1}{4Z_2}} \; \frac{1}{1 + (1 - m^2) \dfrac{Z_1}{4Z_2}}$$

$$(16.14)$$

Fig. 16.8. Z_T' and Z_π' as a function of $Z_1/4Z_2$ for different values of m.

Equations (16.12) and (16.14) have been plotted in Fig. 16.8 and show that when $m \simeq 0.6$, the best impedance match to a resistive load is obtained.

16.1h. *Attenuation and Phase Shift in T and π Networks.* The attenuation and phase shift is the same in the properly terminated symmetrical T and symmetrical π sections shown in Fig. 16.2 and can be expressed by Eq. (16.15).

$$\frac{e_1}{e_2} = 1 + 2Y + \sqrt{4Y(1 + Y)} \qquad (16.15)$$

where $Y = \dfrac{Z_1}{4Z_2} = U + jV$

From Eq. (16.4)

$$e^\theta = e^{\alpha + j\beta} = 1 + 2Y + \sqrt{4Y(1 + Y)} \qquad (16.16)$$

from which the following can be written,

$$\cosh \theta = \cosh (\alpha + j\beta) = \frac{e^\theta + e^{-\theta}}{2} = 1 + 2Y = (1 + 2U) + j2V \quad (16.17)$$

and by making use of the identity,

$$\cosh (\alpha + j\beta) = \cosh \alpha \cos \beta + j \sinh \alpha \sin \beta \qquad (16.18)$$

Equations (16.19) and (16.20) can be obtained by equating the real and imaginary terms of Eqs. (16.17) and (16.18).

$$1 + 2U = \cosh \alpha \cos \beta \qquad (16.19)$$
$$2V = \sinh \alpha \sin \beta \qquad (16.20)$$

and then from the identity

$$\cos^2 \beta + \sin^2 \beta = 1 \qquad (16.21)$$

the following equation can be written

$$\left(\frac{1 + 2U}{\cosh \alpha}\right)^2 + \left(\frac{2V}{\sinh \alpha}\right)^2 = 1 \qquad (16.22)$$

Equation (16.22) relates the real and imaginary terms of $Z_1/4Z_2$ with the attenuation constant α.   Figure 16.9 is a plot of the attenuation constant α versus the absolute

FIG. 16.9. Relation between the attenuation constant α per section of any ladder type of structure and one-fourth the ratio of its series to its shunt impedance.

value of $Z_1/4Z_2$ for different absolute values of the angle ϕ whose tangent is equal to V/U.   The angle ϕ can also be expressed as being equal to the absolute value of the angle obtained by dividing the complex quantity Z_1 by the complex quantity Z_2. The attenuation in symmetrical constant-k or m-derived filter sections having any degree of dissipation in the elements can be determined from Fig. 16.9.   To determine the attenuation constant as a function of frequency, it is first necessary to determine the absolute value and phase angle of $Z_1/4Z_2$ as a function of frequency. The terms Z_1 and Z_2 refer to the lumped impedances of the series and shunt arms, respectively.

The phase shift β is the same in the properly terminated symmetrical T and π sections shown in Fig. 16.2 and can be determined from Eqs. (16.19) and (16.20) and the identity

$$\cosh^2 \alpha - \sinh^2 \alpha = 1 \qquad (16.23)$$

Therefore,

$$\left(\frac{1 + 2U}{\cos \beta}\right)^2 - \left(\frac{2V}{\sin \beta}\right)^2 = 1 \qquad (16.24)$$

Equation (16.24) relates the real and imaginary terms of $Z_1/4Z_2$ with the phase-shift constant β. Figure 16.10 is a plot of the phase-shift constant versus the absolute value of $Z_1/4Z_2$ for different absolute values of the angle ϕ whose tangent is equal to V/U. The phase shift in symmetrical constant-k or m-derived filter sections having any degree of dissipation in the elements can be determined from Fig. 16.10. To
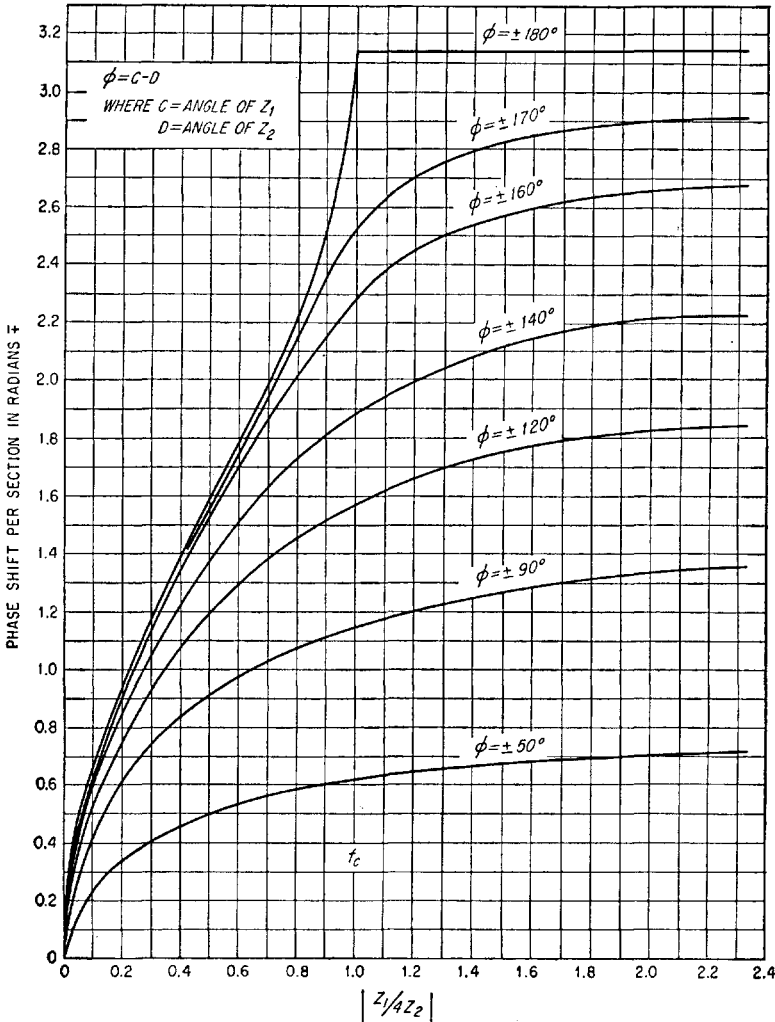
FIG. 16.10. Phase shift in properly terminated filter sections.

determine the phase-shift constant β as a function of frequency, it is first necessary to determine the absolute value and phase angle of $Z_1/4Z_2$ as a function of frequency.

16.1i. Practical Design Considerations. Figures 16.11 through 16.14 inclusive show the configurations for various types of filters employing constant-k and m-derived symmetrical sections. If the two halves of an m-derived section are employed as end sections, precautions should be taken to ensure that the configurations are as shown in the figures to avoid impedance mismatches within the filter.

The *dissipation factor d* is normally applied to filters instead of the figure of merit Q. The term d is defined as the reciprocal of Q, that is, $d = 1/Q$. At audio frequencies d is essentially dependent on the filter inductors alone since the Q's of the circuit capacitors are relatively much greater. At much higher frequencies the Q's of both the capacitors and inductors may be comparable; consequently d can be considered as being equal to the sum of the dissipation factors for the capacitors and inductors.

$$m = \sqrt{1 - \left(\frac{f_c}{f_\infty}\right)^2}$$
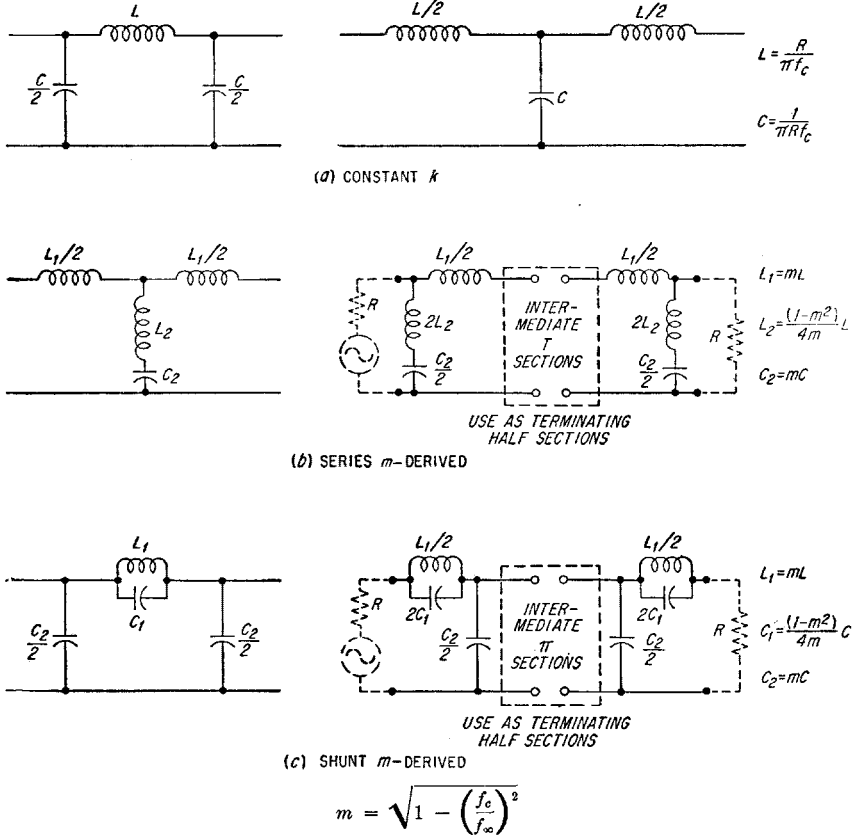
where f_c = frequency of cutoff
f_∞ = frequency of maximum attenuation

Fig. 16.11. Low-pass filter sections.

The performance of symmetrical filter sections having dissipation factors of 0.01 ($Q = 100$) and 0.03 ($Q = 33$) are plotted in Figs. 16.15 to 16.36. Performance characteristics of filter sections having different Q's can usually be determined by interpolation or extrapolation of these data.

To obtain the total attenuation through a filter, it is necessary to determine the sum of the attenuations in decibels of each individual filter section.

Identical performance will be obtained from either T or π filter sections. The choice is normally dependent on which is the more economical to manufacture.

$$L = \frac{R}{4\pi f_c}$$

$$C = \frac{1}{4\pi f_c R}$$

(a) CONSTANT k

$$C_1 = \frac{C}{m}$$

$$L_2 = \frac{L}{m}$$

$$C_2 = \frac{4mC}{1-m^2}$$

USE AS TERMINATING
HALF SECTIONS

(b) SERIES m-DERIVED

$$L_1 = \frac{4mL}{1-m^2}$$

$$L_2 = \frac{L}{m}$$

$$C_1 = \frac{C}{m}$$

USE AS TERMINATING
HALF SECTIONS
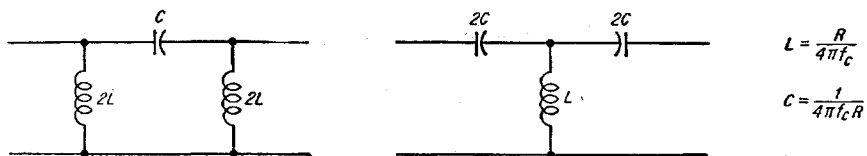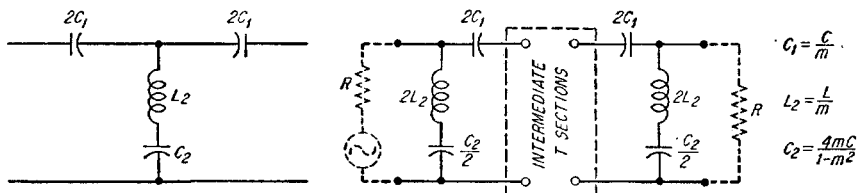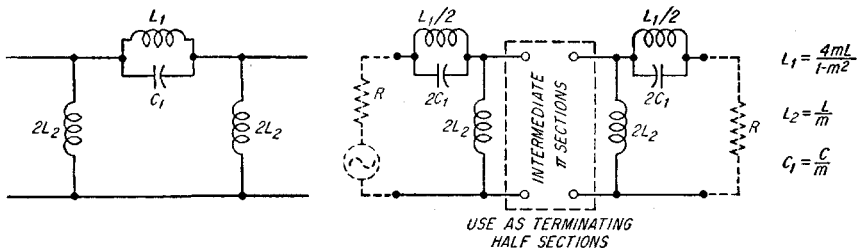
(c) SHUNT m-DERIVED

$$m = \sqrt{1 - \left(\frac{f_\infty}{f_c}\right)^2}$$

where f_∞ = frequency of maximum attenuation
f_c = frequency of cutoff

FIG. 16.12. High-pass filter sections.

Example 16.1

Design a 15- to 20-kc bandpass filter with a characteristic impedance of 600 ohms and whose attenuation at 15 and 20 kc is within 1.5 db of the mean frequency attenuation. Assume that the attenuation at 25 kc must be at least 30 db greater than the attenuation at the mean frequency. Assume inductor Q's of 100.

Solution

Normally, a filter will be selected which provides the sharpest cutoff just outside the desired passband. This requires that the filter be designed to have the maximum allowable attenuation at the frequency limits of the passband.

1. Determine f_2/f_m and sections to be used.

$$f_m = \sqrt{15 \times 20} = 17.32 \text{ kc}$$

At the frequency where the attenuation is to be 1.5 db greater than that at the mean frequency,

$$\frac{f}{f_m} = \frac{20}{17.32} = 1.155$$

and at the frequency where the attenuation is to be at least 30 db greater than that at the mean frequency,

$$\frac{f}{f_m} = \frac{25}{17.32} = 1.444$$

From Fig. 16.22, where $f_2/f_m = 1.2$ and if $f/f_m = 1.155$, the attenuation in a typical section will be approximately 0.6 db greater than at the mean frequency. In addition, the total attenuation for the combination of a single $m = 1$ section and a single $m = 0.6$ section at $1.444 f_m$ is approximately 37 db. Therefore, a bandpass filter with $f_2/f_m = 1.2$ and

(a) CONSTANT k

$L_1 = \dfrac{R}{\pi(f_2-f_1)}$ $C_1 = \dfrac{f_2-f_1}{4\pi f_1 f_2 R}$

$L_2 = \dfrac{(f_2-f_1)R}{4\pi f_1 f_2}$ $C_2 = \dfrac{1}{\pi(f_2-f_1)R}$

(b) SERIES m–DERIVED   USE AS TERMINATING HALF SECTIONS

$L_3 = mL_1$ $C_3 = C_1/m$

$L_4 = L_1 A\left(1+1/N^2\right)$ $C_4 = \left(\dfrac{C_1}{1+N^2}\right)\dfrac{1}{A}$

$C_5 = \left(\dfrac{C_1}{1+1/N^2}\right)\dfrac{1}{A}$ $L_5 = L_1 A\left(1+N^2\right)$

(c) SHUNT m–DERIVED   USE AS TERMINATING HALF SECTIONS

$L_6 = mL_1\left[\dfrac{(N-1/N)^2}{1+N^2}\right]$ $C_6 = \dfrac{C_1}{m}\left[\dfrac{1+1/N^2}{(N-1/N)^2}\right]$

$L_7 = mL_1\left[\dfrac{(N-1/N)^2}{1+1/N^2}\right]$ $C_7 = \dfrac{C_1}{m}\left[\dfrac{1+N^2}{(N-1/N)^2}\right]$

$L_8 = \dfrac{L_2}{m}$ $C_8 = mC_2$

$A = (1-m^2)/4m$

$m = \sqrt{1-\dfrac{\left(f_2/f_m - f_m/f_2\right)^2}{\left(f_{2\infty}/f_m - f_m/f_{2\infty}\right)^2}}$

$N = \dfrac{f_{2\infty}}{\sqrt{f_1 f_2}} = \dfrac{f_{2\infty}}{f_m}$

WHERE $f_1 = $ LOWER CUTOFF FREQUENCY

$f_2 = $ UPPER CUTOFF FREQUENCY

$f_m = \sqrt{f_1 f_2} = \sqrt{f_{1\infty} f_{2\infty}}$

$f_{2\infty} = $ UPPER FREQUENCY OF MAXIMUM ATTENUATION

$f_{1\infty} = $ LOWER FREQUENCY OF MAXIMUM ATTENUATION

FIG. 16.13. Bandpass filter sections.

composed of an $m = 1$ section and two $m = 0.6$ half sections, for proper terminations, will satisfy the requirements.

NOTE: If additional attenuation at some specific frequency is desired, add an additional filter section. This, however, complicates the problem since the total attenuation at $f/f_m = 1.155$ will be more than the desired 1.5 db. This necessitates interpolation between Figs. 16.22 and 16.24 for the proper value of f_2/f_m for there to be no more than a total of 1.5 db attenuation at $f/f_m = 1.155$.

2. Determine f_1 and f_2.

From step 1, f_2/f_m must equal 1.2; therefore,

$$f_2 = 1.2 f_m = 1.2 \times 17.32 = 20.8 \text{ kc}$$

$$f_1 = \frac{f_m}{1.2} = \frac{17.32}{1.2} = 14.43 \text{ kc}$$
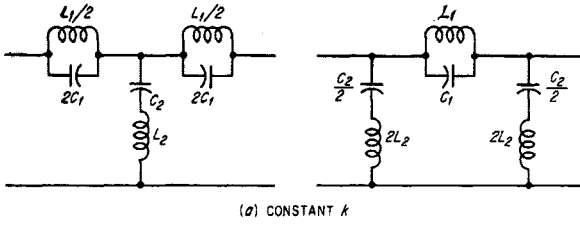
$$L_1 = \frac{(f_2 - f_1)R}{\pi f_1 f_2}$$

$$C_1 = \frac{1}{4\pi(f_2 - f_1)R}$$

$$L_2 = \frac{R}{4\pi(f_2 - f_1)}$$

$$C_2 = \frac{f_2 - f_1}{\pi f_1 f_2 R}$$

(a) CONSTANT k

USE AS TERMINATING
HALF SECTIONS
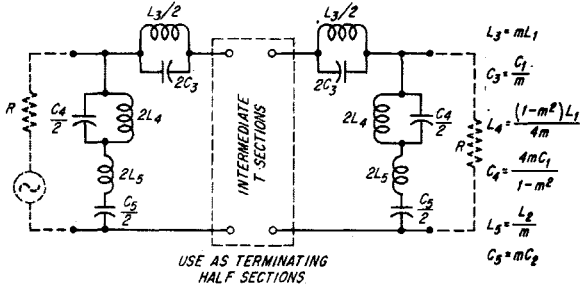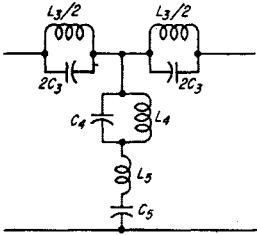
$$L_3 = mL_1$$

$$C_3 = \frac{C_1}{m}$$

$$L_4 = \frac{(1 - m^2)L_1}{4m}$$

$$C_4 = \frac{4mC_1}{1 - m^2}$$

$$L_5 = \frac{L_2}{m}$$

$$C_5 = mC_2$$

(b) SERIES m–DERIVED

USE AS TERMINATING
HALF SECTIONS
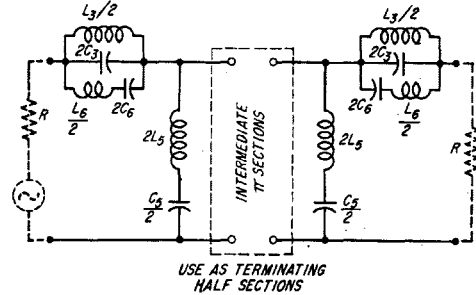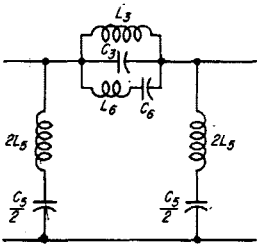
$$L_6 = \left[\frac{4m}{1 - m^2}\right]L_2$$

$$C_6 = \left[\frac{1 - m^2}{4m}\right]C_2$$

(c) SHUN m–DERIVED

$$m = \sqrt{1 - \frac{\left(\dfrac{f_{2\infty}}{f_m} - \dfrac{f_m}{f_{2\infty}}\right)^2}{\left(\dfrac{f_2}{f_m} - \dfrac{f_m}{f_2}\right)^2}}$$

$$f_m = \sqrt{f_1 f_2} = \sqrt{f_{1\infty}f_{2\infty}}$$

where f_1 = lower cutoff frequency
f_2 = upper cutoff frequency
$f_{1\infty}$ = lower frequency of maximum attenuation
$f_{2\infty}$ = upper frequency of maximum attenuation

FIG. 16.14. Band-elimination filter sections.
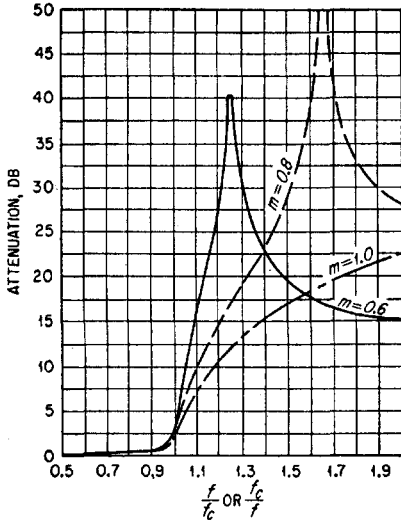
FIG. 16.15. High-pass or low-pass filter char-
acteristics; $d = 0.03$; $m = 1.0$; $m = 0.8$;
$m = 0.6$. (*Reprinted from "Simplified Fil-
ter Design" by J. Ernest Smith through the
courtesy of RCA Institutes, Inc.*)
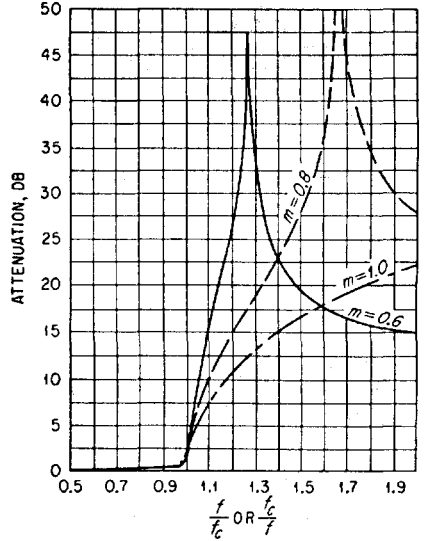
FIG. 16.16. High-pass or low-pass filter char-
acteristics; $d = 0.01$; $m = 1.0$; $m = 0.8$;
$m = 0.6$. (*Reprinted from "Simplified Fil-
ter Design" by J. Ernest Smith through the
courtesy of RCA Institutes, Inc.*)
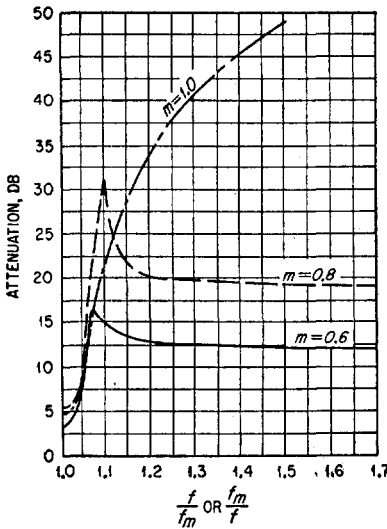
FIG. 16.17. Bandpass filter characteristics;
$f_m/f_1 = f_2/f_m = 1.05$; $d = 0.03$; $m = 1.0$;
$m = 0.8$; $m = 0.6$. (*Reprinted from "Sim-
plified Filter Design" by J. Ernest Smith
through the courtesy of RCA Institutes, Inc.*)
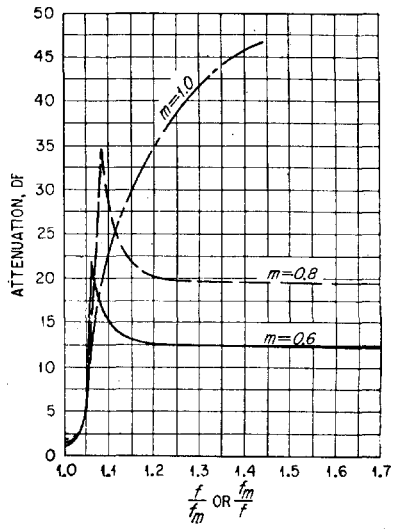
FIG. 16.18. Bandpass filter characteristics;
$f_m/f_1 = f_2/f_m = 1.05$; $d = 0.01$; $m = 1.0$;
$m = 0.8$; $m = 0.6$. (*Reprinted from "Sim-
plified Filter Design" by J. Ernest Smith
through the courtesy of RCA Institutes, Inc.*)
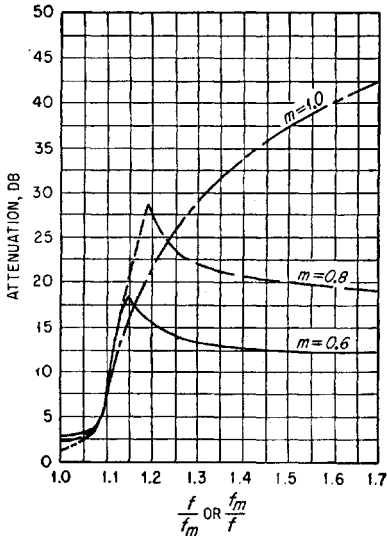
FIG. 16.19. Bandpass filter characteristics; $f_m/f_1 = f_2/f_m = 1.1$; $d = 0.03$; $m = 1.0$; $m = 0.8$; $m = 0.6$. (Reprinted from "Simplified Filter Design" by J. Ernest Smith through the courtesy of RCA Institutes, Inc.)
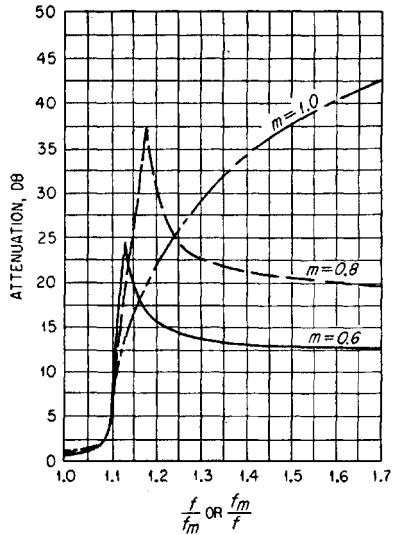
FIG. 16.20. Bandpass filter characteristics; $f_m/f_1 = f_2/f_m = 1.1$; $d = 0.01$; $m = 1.0$; $m = 0.8$; $m = 0.6$. (Reprinted from "Simplified Filter Design" by J. Ernest Smith through the courtesy of RCA Institutes, Inc.)
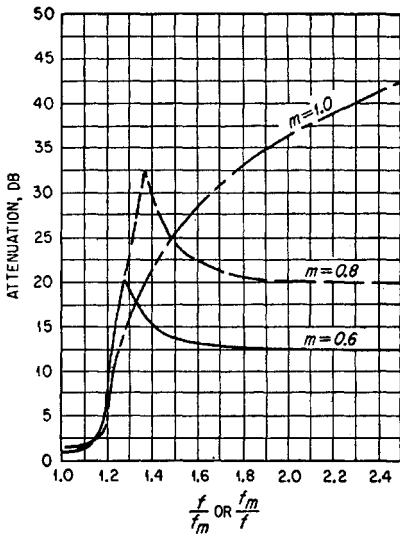
FIG. 16.21. Bandpass filter characteristics; $f_m/f_1 = f_2/f_m = 1.2$; $d = 0.03$; $m = 1.0$; $m = 0.8$; $m = 0.6$. (Reprinted from "Simplified Filter Design" by J. Ernest Smith through the courtesy of RCA Institutes, Inc.)
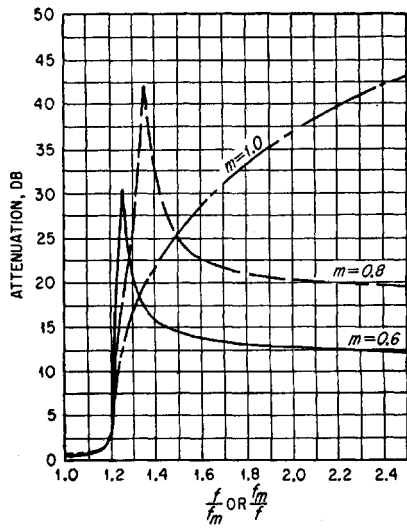
FIG. 16.22. Bandpass filter characteristics; $f_m/f_1 = f_2/f_m = 1.2$; $d = 0.01$; $m = 1.0$; $m = 0.8$; $m = 0.6$. (Reprinted from "Simplified Filter Design" by J. Ernest Smith through the courtesy of RCA Institutes, Inc.)
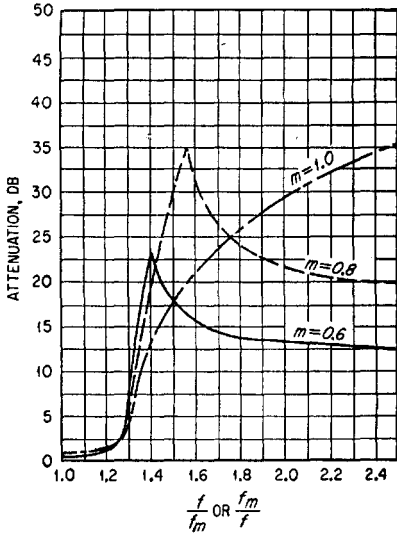
FIG. 16.23. Bandpass filter characteristics; $f_m/f_1 = f_2/f_m = 1.3$; $d = 0.03$; $m = 1.0$; $m = 0.8$; $m = 0.6$. (*Reprinted from "Simplified Filter Design" by J. Ernest Smith through the courtesy of RCA Institutes, Inc.*)
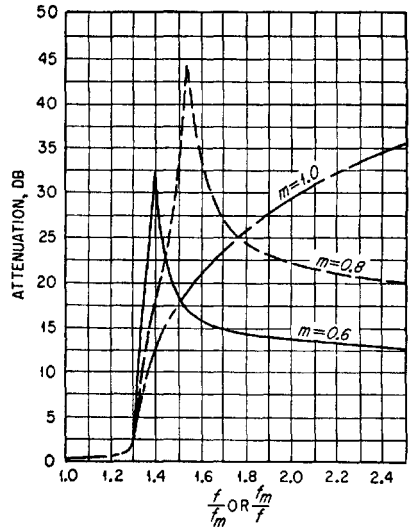
FIG. 16.24. Bandpass filter characteristics; $f_m/f_1 = f_2/f_m = 1.3$; $d = 0.01$; $m = 1.0$; $m = 0.8$; $m = 0.6$. (*Reprinted from "Simplified Filter Design" by J. Ernest Smith through the courtesy of RCA Institutes, Inc.*)
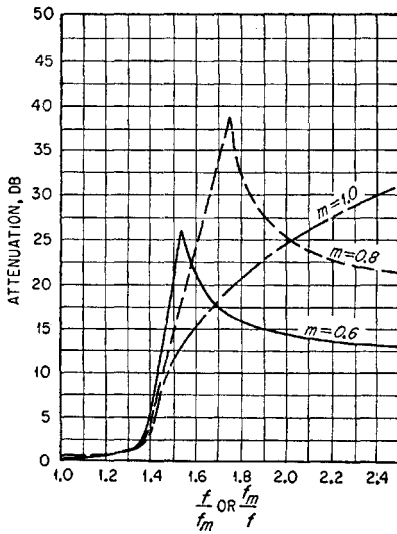
FIG. 16.25. Bandpass filter characteristics; $f_m/f_1 = f_2/f_m = 1.4$; $d = 0.08$; $m = 1.0$; $m = 0.8$; $m = 0.6$. (*Reprinted from "Simplified Filter Design" by J. Ernest Smith through the courtesy of RCA Institutes, Inc.*)
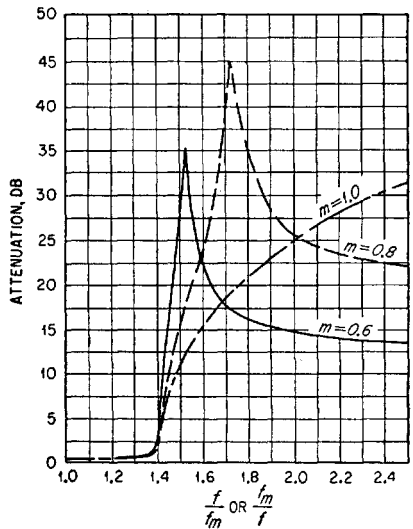
FIG. 16.26. Bandpass filter characteristics; $f_m/f_1 = f_2/f_m = 1.4$; $d = 0.01$; $m = 1.0$; $m = 0.8$; $m = 0.6$. (*Reprinted from "Simplified Filter Design" by J. Ernest Smith through the courtesy of RCA Institutes, Inc.*)
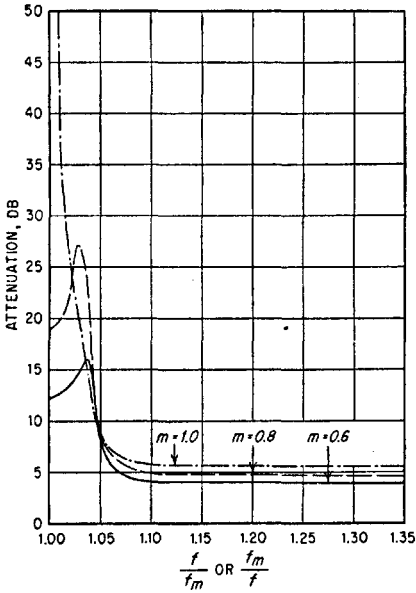
FIG. 16.27. Band-elimination filter charac-
teristics; $f_m/f_1 = f_2/f_m = 1.05$; $d = 0.03$;
$m = 1.0$; $m = 0.8$; $m = 0.6$.
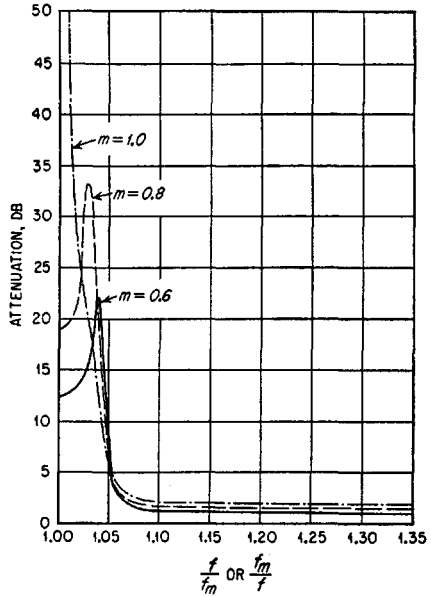
FIG. 16.28. Band-elimination filter charac-
teristics; $f_m/f_1 = f_2/f_m = 1.05$; $d = 0.01$;
$m = 1.0$; $m = 0.8$; $m = 0.6$.
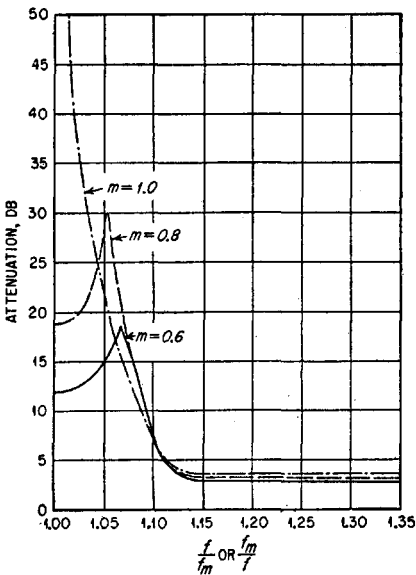
FIG. 16.29. Band-elimination filter charac-
teristic; $f_m/f_1 = f_2/f_m = 1.10$; $d = 0.03$;
$m = 1.0$; $m = 0.8$; $m = 0.6$.
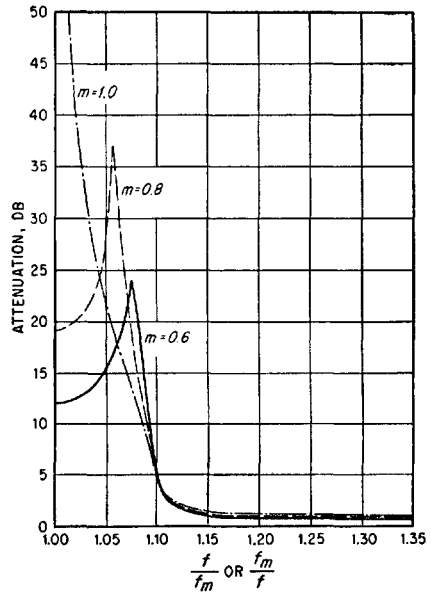
FIG. 16.30. Band-elimination filter charac-
teristics; $f_m/f_1 = f_2/f_m = 1.10$; $d = 0.01$;
$m = 1.0$; $m = 0.8$; $m = 0.6$.
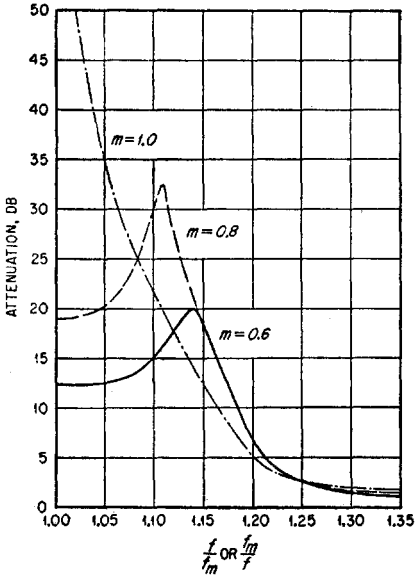
FIG. 16.31. Band-elimination filter characteristics; $f_m/f_1 = f_2/f_m = 1.2$; $d = 0.03$; $m = 1.0$; $m = 0.8$; $m = 0.6$.
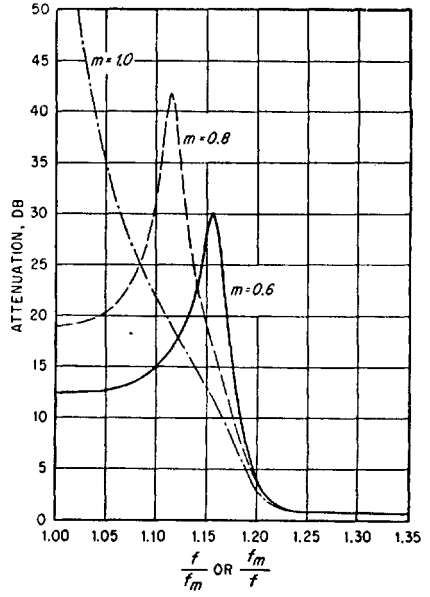
FIG. 16.32. Band-elimination filter characteristics; $f_m/f_1 = f_2/f_m = 1.2$; $d = 0.01$; $m = 1.0$; $m = 0.8$; $m = 0.6$.
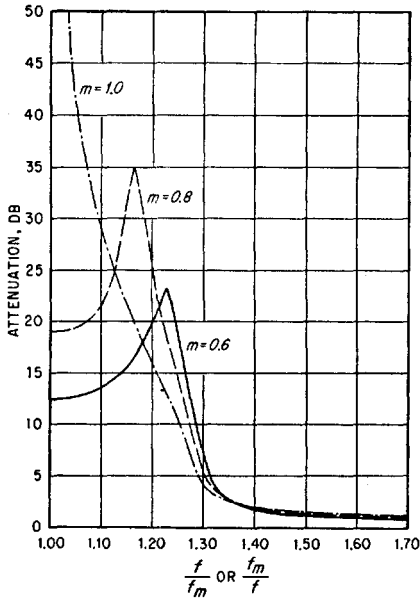
FIG. 16.33. Band-elimination filter characteristics; $f_m/f_1 = f_2/f_m = 1.3$; $d = 0.03$; $m = 1.0$; $m = 0.8$; $m = 0.6$.
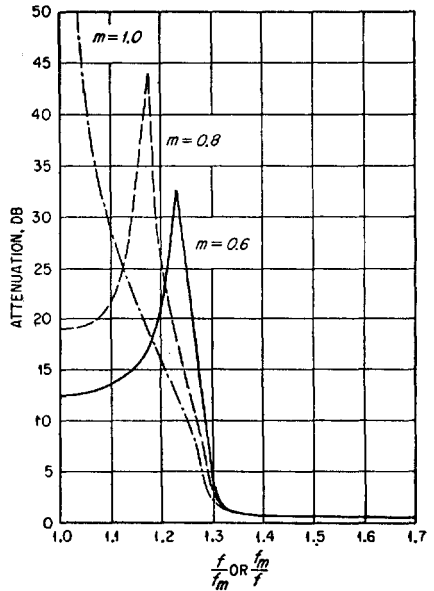
FIG. 16.34. Band-elimination filter characteristics; $f_m/f_1 = f_2/f_m = 1.3$; $d = 0.01$; $m = 1.0$; $m = 0.8$; $m = 0.6$.
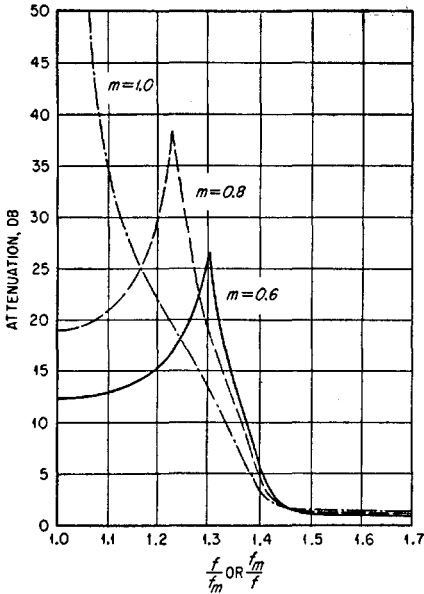
FIG. 16.35. Band-elimination filter charac-
teristics; $f_m/f_1 = f_2/f_m = 1.4$; $d = 0.03$;
$m = 1.0$; $m = 0.8$; $m = 0.6$.
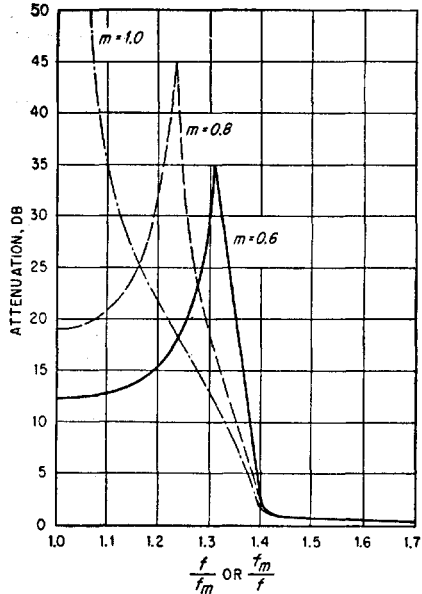
FIG. 16.36. Band-elimination filter charac-
teristics; $f_m/f_1 = f_2/f_m = 1.4$; $d = 0.01$;
$m = 1.0$; $m = 0.8$; $m = 0.6$.
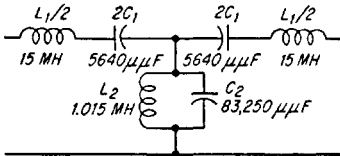
(a) CONSTANT k SECTION
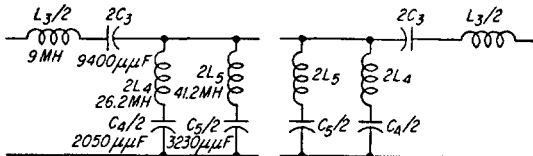
(b) TWO HALVES OF AN m =0.6 SECTION
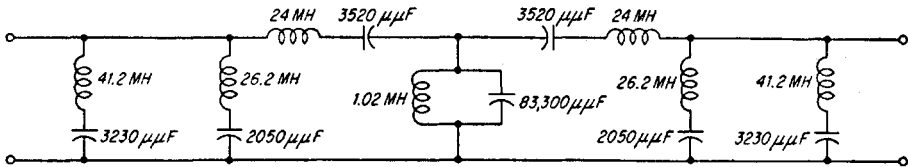
(c) COMBINED SECTIONS

FIG. 16.37. 15- to 20-kc bandpass filter sections.

3. Design the $m = 1$ or constant-k section (Fig. 16.13).
Assuming that T sections are to be employed,

$$L_1 = \frac{600}{\pi(20,800 - 14,430)} = 0.030 \text{ henry, or 30 mh}$$

$$C_1 = \frac{20,800 - 14,430}{4\pi \times 14,430 \times 20,800 \times 600} = 0.00282 \text{ farad, or 2,820 } \mu\mu f$$

$$L_2 = \frac{(20,800 - 14,430)600}{4\pi \times 14,430 \times 20,800} = 0.001015 \text{ henry, or 1.015 mh}$$

$$C_2 = \frac{1}{\pi(20,800 - 14,430)600} = 0.08325 \text{ farad, or 83,250 } \mu\mu f$$

The $m = 1$ section is shown in Fig. 16.37a.

4. Design the $m = 0.6$ half sections (Fig. 16.13).

The 0.6 half sections must be halves of a T section to match the constant-k section designed in step 3.

$$L_3 = 0.6 \times 30 = 18 \text{ mh}$$

$$C_3 = \frac{2,820}{0.6} = 4,700 \ \mu\mu f$$

The values of A and N must be calculated before the following equations can be written.

$$L_4 = 30 \times 0.2668 \left(1 + \frac{1}{1.255^2} \right) = 13.08 \text{ mh}$$

$$C_4 = \frac{2,820}{1 + 1.255^2} \times \frac{1}{0.2668} = 4,100 \ \mu\mu f$$

$$L_5 = 30 \times 0.2668(1 + 1.255^2) = 20.6 \text{ mh}$$

$$C_5 = \frac{2,820}{1 + \dfrac{1}{1.255^2}} \times \frac{1}{0.2668} = 6,460 \ \mu\mu f$$

The $m = 0.6$ half sections are shown in Fig. 16.37b.

5. Combine the half sections with the constant-k section. See Fig. 16.37c.

NOTE: The response of the completed filter is shown in Fig. 16.38.
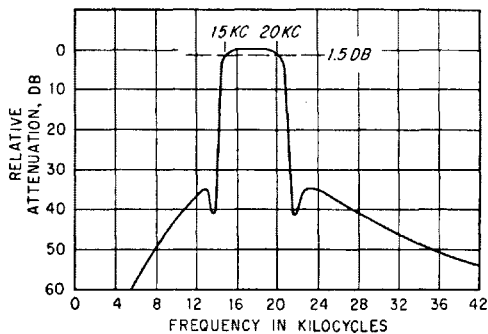
FIG. 16.38. Characteristics of filter shown in Fig. 16.37.

16.2. Bridged-T and Parallel-T Networks. Bridged-T and parallel-T networks have three general uses: as band-elimination filters, as feedback networks in frequency-selective amplifiers and oscillators, and as calibrated bridges in measuring devices.

Unloaded bridged-T and parallel-T networks of the types shown in Figs. 16.39 to 16.41 introduce no attenuation to a signal at either zero or infinite frequency and theoretically can be designed to have any desired amount of attenuation at any single frequency. Although it is possible to design the circuits in Figs. 16.39 and 16.40 to have infinite attenuation at the center frequency, the attenuation will be limited by the unbalance of the circuit elements, inadequate shielding between the input and output circuits, and the stray capacitance shunting the circuit elements. In general, however, a signal attenuation of 50 to 60 db can be obtained without difficulty.

The phase shift in bridged-T and parallel-T networks is 0° at zero frequency, and with increased frequency the phase lags and reaches a maximum on the low-frequency side of the frequency of maximum attenuation. The phase shift reverses to a leading phase shift of equal amplitude on the high-frequency side of the center frequency, and as the frequency is further increased, the phase shift reduces and approaches zero

as a limit at infinite frequency. Depending on the design of the network, the maximum phase shift can be as much as 180°.

The *RLC* bridged-T network shown in Fig. 16.39 is unique in that there is no theoretical limit as to the minimum width of the attenuation band. In an actual application, however, the minimum attenuation bandwidth is limited by the *Q* of the inductor. For the parallel-T and *RC* bridged-T networks shown in Figs. 16.40 and 16.41, respectively, there are theoretical limits as to the minimum widths of the attenuation bands. Narrower attenuation bandwidths can be obtained with parallel-T networks than with *RC* bridged-T networks.

It is possible to design *RLC* bridged-T networks (see Fig. 16.39) and parallel-T networks (see Fig. 16.40) to have negative transfer functions.[1]
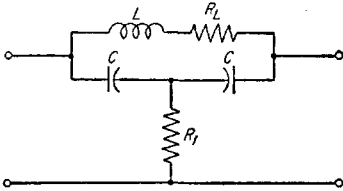
FIG. 16.39. Bridged-T network.   FIG. 16.40. Parallel-T network.
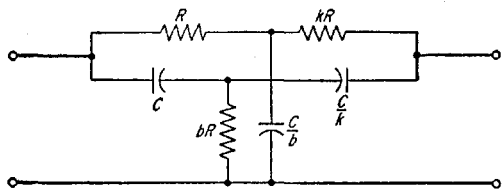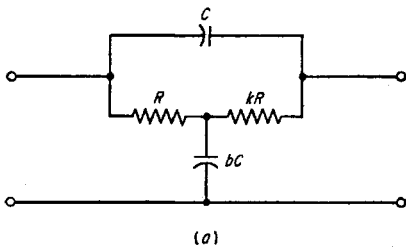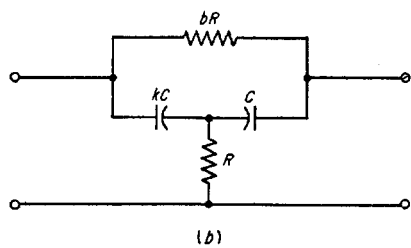
(a)   (b)

FIG. 16.41. Bridged-T attenuating networks.

The *RC* bridged-T networks shown in Fig. 16.41 have the disadvantage of requiring circuit values with extremely large ratios whenever very large amounts of attenuation are desired.

16.2a. RLC Bridged-T Network. The transfer function for the unloaded *RLC* bridged-T network shown in Fig. 16.39 is given by Eq. (16.25).

For Fig. 16.39

$$\frac{E_o}{E_i} = \frac{(1 - \omega^2 R_1 R_L C^2) + j(\omega 2 R_1 C - \omega^3 R_1 L C^2)}{(1 - \omega^2 R_1 R_L C^2 - \omega^2 LC) + j(\omega 2 R_1 C + \omega R_L C - \omega^3 R_1 L C^2)} \quad (16.25)$$

If the network is to be a true null network, i.e., have infinite attenuation at some specific frequency ω_o, Eqs. (16.26) and (16.27) must be satisfied.

For a complete null at ω_o

$$\omega_o = \sqrt{\frac{2}{LC}} \quad (16.26)$$

and $$R_1 R_L = \frac{L}{2C} \quad (16.27)$$

where ω_o = frequency at which null occurs

[1] A system has a negative transfer function when the output voltage is 180° out of phase with respect to the input voltage. A network of this type is often used as the feedback network in a single-stage oscillator.

For the case of the true null network, Eq. (16.25) can be simplified to the form given by Eq. (16.28).

For Fig. 16.39 and a complete null

$$\frac{E_o}{E_i} = \frac{1}{1 + j\dfrac{2X}{Q_L(1 - X^2)}}$$ (16.28)

where $Q_L = \dfrac{\omega_o L}{R_L}$

$X = \dfrac{\omega}{\omega_o}$

In the case of a complete null, the attenuation bandwidth can be expressed by the network figure of merit Q_n.

$$Q_n = \frac{\omega_o}{\omega_2 - \omega_1}$$ (16.29)

where ω_o = null frequency

ω_1, ω_2 = frequencies at which network attenuates input signal by 3 db

If the RLC bridged-T network shown in Fig. 16.39 is adjusted for a complete null, the value of Q_n can be determined from Eq. (16.30).

$$Q_n = \frac{Q_L}{2}$$ (16.30)

The center frequency ω_o, i.e., the frequency of maximum attenuation, for the circuit in Fig. 16.39 can be determined exactly from Eq. (16.26) only in the case where Eq. (16.27) is satisfied. For other than the complete null case, however, the error in determining the frequency of maximum attenuation from Eq. (16.26) is very small provided that the product $R_1 R_L$ is no more than approximately four times as large as the value specified by Eq. (16.27) and provided $R_1 \gg R_L$. Under these conditions, the transfer function for the network at the frequency of maximum attenuation can very nearly be determined from Eq. (16.31).

For Fig. 16.39 and $\omega = \omega_o$

$$\frac{E_o}{E_i} = \frac{\omega_o^2 R_1 R_L C^2 - 1}{1 + \omega_o^2 R_1 R_L C^2}$$ (16.31)

If the product $R_1 R_L$ is less than the value specified by Eq. (16.27), the transfer function at ω_o as given by Eq. (16.31) will be negative. In an actual circuit it is therefore possible to make R_1 adjustable and vary the degree of attenuation over wide ranges and obtain either a

FIG. 16.42. Single-stage oscillator using a feedback network having a negative transfer function. (R_c must be large enough that it does not load the feedback network.)

positive or negative transfer function without appreciably changing the center frequency of the network.

A network with a negative transfer function has value in the design of a single-stage oscillator. The oscillator circuit can be quite simple as shown in Fig. 16.42. It should be noted that the amplifier gain without feedback must be sufficient to compensate for the loss in the feedback network. Maximum frequency stability is achieved if $R_1 R_L$ is no smaller than absolutely necessary to cause the network to have a negative transfer function. This condition ensures a maximum rate of phase shift

versus frequency through the network in the region of the center frequency and consequently reduces the range of frequencies at which the phase shift in the network approaches $\pm 180°$.
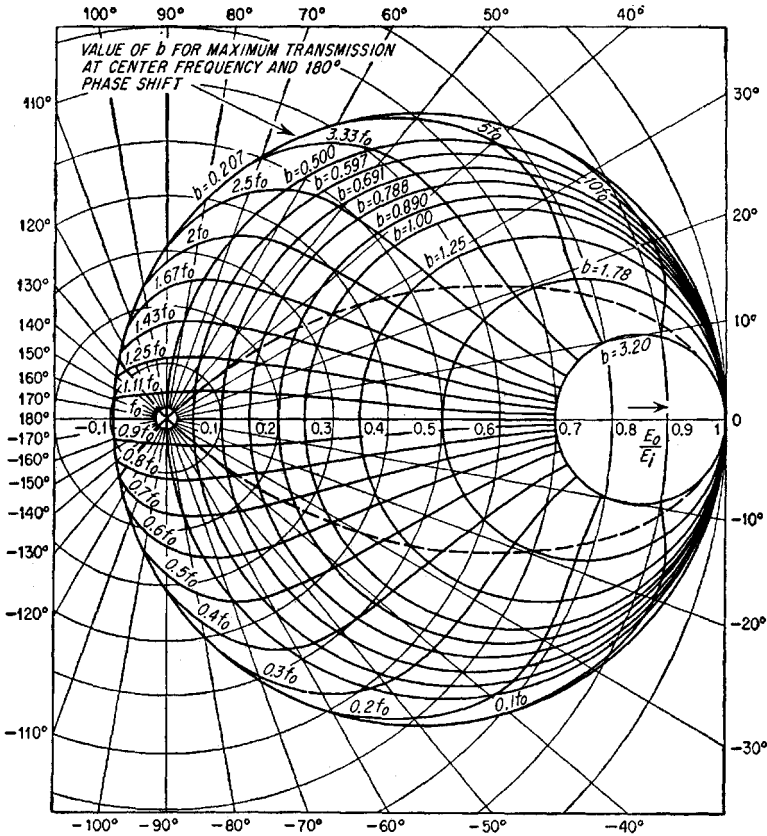
FIG. 16.43. Amplitude and phase characteristics of parallel-T networks of the type shown in Fig. 16.40 for $k = 1$. The frequencies above and below f_o at which the transmission through the network is 3 db greater than the transmission at the center frequency f_o can be determined by the intersection of the dashed lines and the applicable circle.

16.2b. RC Parallel-T Network. The transfer function for the unloaded RC parallel-T network shown in Fig. 16.40 is given by Eq. (16.32).

$$\frac{E_o}{E_i} = \frac{\left(\dfrac{\omega_o}{\omega} - \dfrac{\omega}{\omega_o}\right) + j\left(b + \dfrac{b}{k} - 1\right)}{\left(\dfrac{\omega_o}{\omega} - \dfrac{\omega}{\omega_o}\right) + j\left(b + \dfrac{b}{k} + \dfrac{1}{k} + \dfrac{1}{b}\right)} \qquad (16.32)$$

where $\omega_o = 1/RC$

At the center frequency ω_o, i.e., the frequency of maximum attenuation, Eq. (16.32) reduces to Eq. (16.33).

For Fig. 16.40 at $\omega = \omega_o$ and $RC = 1/\omega_o$

$$\frac{E_o}{E_i} = \frac{b^2(k+1) - bk}{b^2(k+1) + b + k} \qquad (16.33)$$

From Eq. (16.32) it is easily shown that a complete null is obtained with the network shown in Fig. 16.40 provided Eqs. (16.34) and (16.35) are satisfied.

For Fig. 16.40 and a complete null at $\omega = \omega_o$

$$b = \frac{k}{k+1} \tag{16.34}$$

and

$$RC = \frac{1}{\omega_o} \tag{16.35}$$

In the case of the complete null networks for $k = 100$ and $k = 1$, the values for Q_n [see Eq. (16.29)] are approximately $\frac{1}{2}$ and $\frac{1}{4}$, respectively.
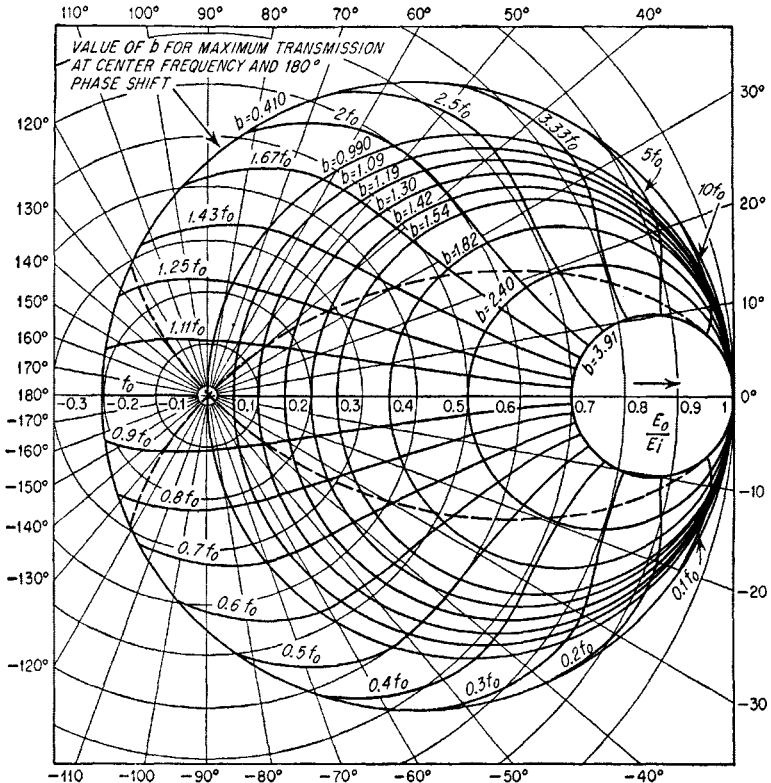
FIG. 16.44. Amplitude and phase characteristics of parallel-T networks of the type shown in Fig. 16.40 for $k = 100$. The frequencies above and below f_o at which the transmission through the network is 3 db greater than the transmission at the center frequency f_o can be determined by the intersection of the dashed lines and the applicable circle.

Equation (16.33) indicates that it is possible to obtain the same values of attenuation at ω_o with an infinite number of combinations of values for b and k. Figures 16.43 and 16.44 are plots of the amplitude and phase characteristics as determined from Eq. (16.32) for $k = 1$ and $k = 100$, respectively. As can be seen, the attenuation bandwidths are much narrower for $k = 100$ than for $k = 1$. Values of k larger than 100, however, make no appreciable reduction in the attenuation bandwidth. If the circle which defines the locus of the network transfer function encompasses the origin, the transfer function for the network will be negative at ω_o. The largest circles shown in Figs. 16.43 and 16.44 describe the loci of parallel-T transfer functions

which have the largest possible negative amplitudes at the center frequency ω_o for $k = 1$ and $k = 100$, respectively. The value of b which will cause an RC parallel-T network to have the largest possible negative transfer function at ω_o can be determined from Eq. (16.36).

$$b = \frac{0.414k}{k + 1} \tag{16.36}$$

The amplitude of the transfer function at ω_o can then be determined from Eq. (16.33).
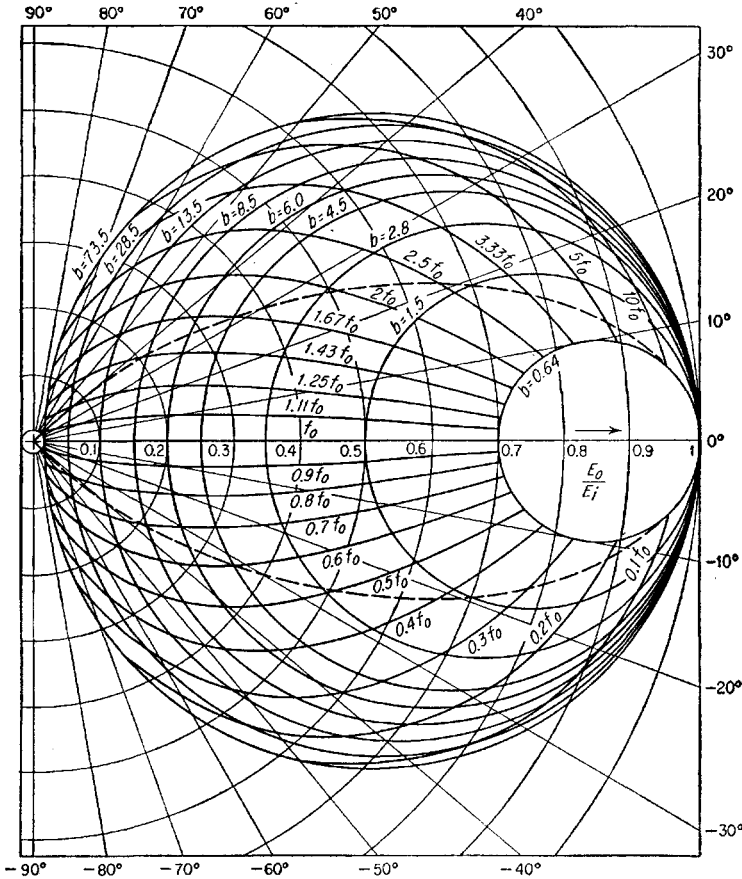
FIG. 16.45. Amplitude and phase characteristics of RC bridged-T networks of the types shown in Fig. 16.41 for $k = 0.5$. The frequencies above and below f_o at which the transmission through the network is 3 db greater than the transmission at the center frequency f_o can be determined by the intersection of the dashed lines and the applicable circle.

16.2c. RC Bridged-T Network. The transfer function for the unloaded RC bridged-T networks shown in Fig. 16.41 is given by Eq. (16.37).

$$\frac{E_o}{E_i} = \frac{\sqrt{bk}\left(\dfrac{\omega_o}{\omega} - \dfrac{\omega}{\omega_o}\right) + j(1 + k)}{\sqrt{bk}\left(\dfrac{\omega_o}{\omega} - \dfrac{\omega}{\omega_o}\right) + j(1 + k + b)} \tag{16.37}$$

where $\omega_o = \dfrac{1}{RC\sqrt{bk}}$

At the center frequency ω_o, Eq. (16.37) reduces to Eq. (16.38).

For Fig. 16.41a and b at $\omega = \omega_o$ and $RC = \dfrac{1}{\omega_o \sqrt{bk}}$

$$\frac{E_o}{E_i} = \frac{1 + k}{1 + k + b} \qquad (16.38)$$

Equation (16.38) shows that neither a complete null nor a negative transfer function is possible. Very large values of attenuation are obtained only for large values of b.
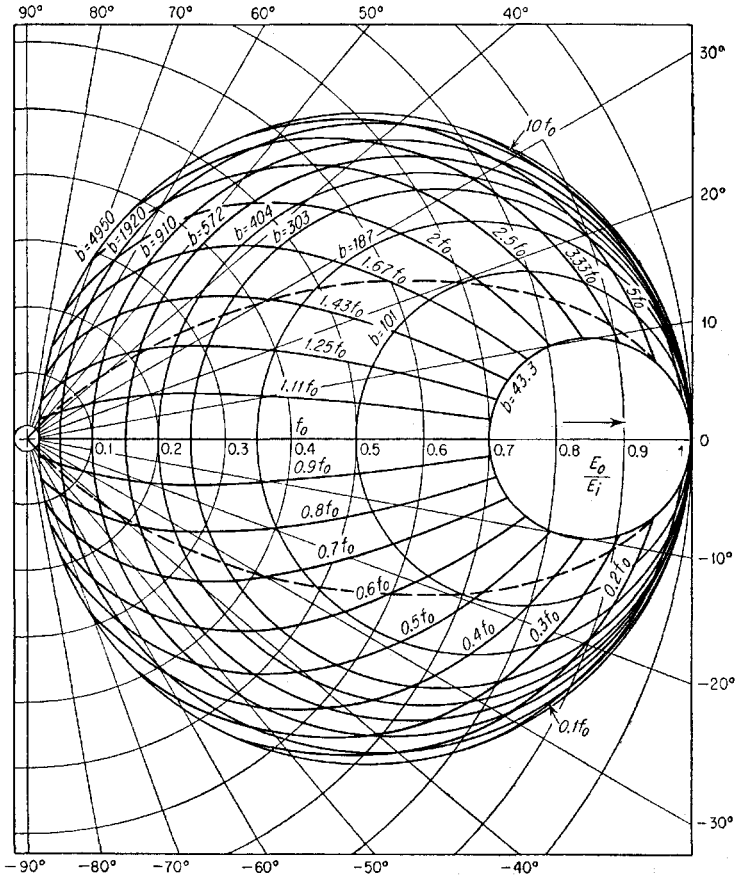
FIG. 16.46. Amplitude and phase characteristics of RC bridged-T networks of the types shown in Fig. 16.41 for $k = 100$. The frequencies above and below f_o at which the transmission through the network is 3 db greater than the transmission at the center frequency f_o can be determined by the intersection of the dashed lines and the applicable circle.

Families of transfer functions have been determined from Eq. (16.37) for $k = 0.5$ and $k = 100$ and are plotted in Figs. 16.45 and 16.46. As in the case of the parallel-T networks, narrower attenuation bandwidths are obtained with larger values of k. Values of k greater than 100, however, make no appreciable reduction in the attenuation bandwidth.

16.3. Special Applications of Bridged-T and Parallel-T Networks. In addition to their use as feedback networks in oscillators, bridged-T and parallel-T networks

are frequently used as feedback networks in rejection amplifiers and frequency-selective amplifiers. To simplify the analysis of rejection and frequency-selective amplifiers, the feedback network has been assumed to be a complete null network in the material which follows. As will be shown later, however, certain precautions must be exercised in the use of null net-

works in feedback amplifiers to achieve stability.

FIG. 16.47. Frequency-selective amplifier.

16.3a. *Frequency-selective Amplifiers.* In a negative-feedback amplifier having a null network in the feedback loop as shown in Fig. 16.47, the amount of feedback at the null frequency theoretically is zero, and the gain of the amplifier at the null frequency theoretically is the same as the gain of the amplifier without feedback. On either side of the null frequency the feedback network has less attenuation and the amplifier gain will be reduced because of the negative feedback. The gain of a frequency-selective amplifier employing a null network in the feedback loop can be determined from Eq. (16.39).

$$A_s = \frac{A}{1 + A\beta} \qquad (16.39)$$

where A_s = gain with feedback

A = gain without feedback

β = transfer function of the feedback network

From Eq. (16.39) it can be seen that A_s will equal A if $\beta = 0$ and will equal $A/(A + 1)$ when $\beta = 1$. It follows that the maximum possible difference in signal level that can exist between the center frequency and any other frequency is $A + 1$. At the -3-db frequencies, because of the manner in which the Q of the null network is defined, the transmission band of the frequency-selective amplifier is narrower than the rejection
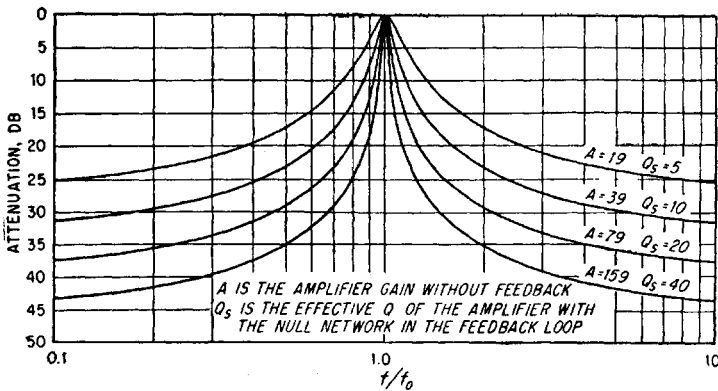
FIG. 16.48. Selectivity curves of a frequency-selective amplifier having a null network with a Q of $\frac{1}{4}$ in the feedback loop. (Symmetrical parallel-T null network where $k = 1$ and $b = 0.5$.)

band of the null circuit by itself; however, the ratio of the center frequency signal level to the level of signals far removed from the center frequency may be much greater in the null network by itself than in the frequency-selective amplifier utilizing the null network.

A family of selectivity curves for frequency-selective amplifiers utilizing symmetrical parallel-T null networks has been plotted in Fig. 16.48.

Stability Considerations. At frequencies near the center of the passband, i.e., the null frequency of the feedback network, the amplifier should have a minimum amount of phase shift. This requirement exists because the negative feedback inherently introduces $\pm 180°$ phase shift and the null network adds an additional $\pm 90°$ phase shift at those frequencies adjacent to the null frequency. Consequently, if the phase shift in the amplifier near the center frequency approaches or becomes equal to $\pm 90°$, the amplifier with feedback will either have pronounced distortion in the selectivity curve or will oscillate.

Care must be exercised to prevent the feedback network from being adjusted to have a negative transfer function since the system would then be regenerative at the null frequency and consequently would oscillate if the amplifier gain were sufficiently high. In addition, excessive phase shift in the amplifier can cause the center frequency of the frequency-selective amplifier to be slightly different from the center frequency of the null network.

Effective Q of a Frequency-selective Amplifier. The effective figure of merit Q_s for the frequency-selective amplifier, i.e., the center frequency divided by the bandwidth at the -3-db frequencies, can be determined from Eq. (16.40).

$$Q_s \simeq (A + 1)Q_n \qquad (16.40)$$

In the case of an RLC bridged-T null network, Q_n is given by Eq. (16.30). For the parallel-T null network, Q_n is approximately equal to $\frac{1}{2}$ when k is equal to 100, and $\frac{1}{4}$ when k is equal to 1.
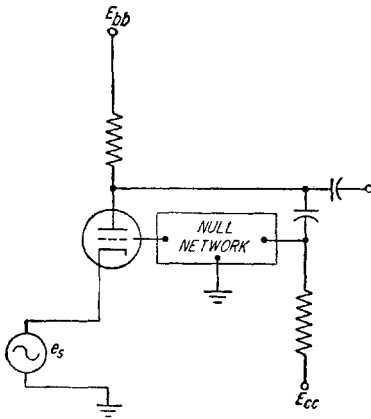
FIG. 16.49. One-stage frequency-selective amplifier.
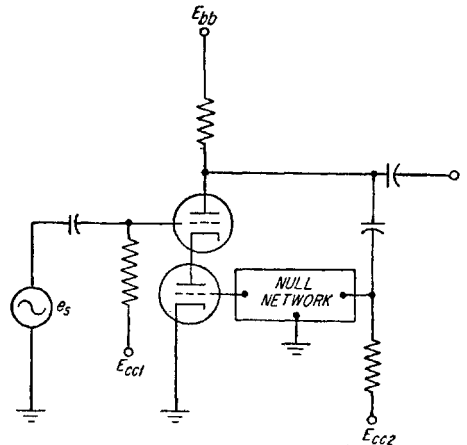
FIG. 16.50. Two-stage frequency-selective amplifier.

Sample Circuits. Ideally a null network should not be loaded. Consequently, in circuit applications the output of the null network is usually connected to the grid of a tube. Bias for this particular tube must therefore be supplied through the network. Typical feedback configurations for frequency-selective amplifiers are shown in Figs. 16.49 and 16.50.

In certain applications it may be advantageous to drive the null networks from cathode followers so that the amplifier circuits are not loaded by the null networks. The result is an increase in amplifier gain and a higher-Q selective amplifier.

Bandpass Selective Amplifiers. Frequently there is the requirement that the passband of the amplifier be flat instead of having a single peak as is typical of a single resonant circuit. The flat or nearly flat top can be acquired by cascading several

independent feedback stages which are staggered in frequency. The frequency to which each feedback stage should be tuned can be determined in the same manner that the frequencies of staggered i-f stages in receivers[1] are determined.

16.3b. Rejection Amplifiers. A rejection amplifier can be realized from the circuit of Fig. 16.47 provided the output is taken at point 2 in the circuit. Preferably, the output should be taken through an isolating stage such as a cathode follower so as not to load the null network. The equation for the gain A_r of this particular rejection amplifier is given by Eq. (16.41).

$$A_r = \frac{A\beta}{1 + A\beta} \tag{16.41}$$

where β = transfer function of null network

The effective figure of merit Q_r for rejection amplifiers using RLC bridged-T and parallel-T null networks is given by Eq. (16.42).

$$Q_r = (A + 1)Q_n \tag{16.42}$$

where Q_r defines the response curve for the rejection amplifier in the same manner that Q_n defines the response curve for the null network.

(a) NEGATIVE FEEDBACK AMPLIFIER USED AS
BOTH A FREQUENCY SELECTIVE AMPLIFIER
AND A REJECTION AMPLIFIER

*WHERE β' IS THE VALUE OF β AT
THE NULL FREQUENCY*

(b) TYPICAL RESPONSE CHARACTERISTICS
OF THE COMBINATION AMPLIFIER
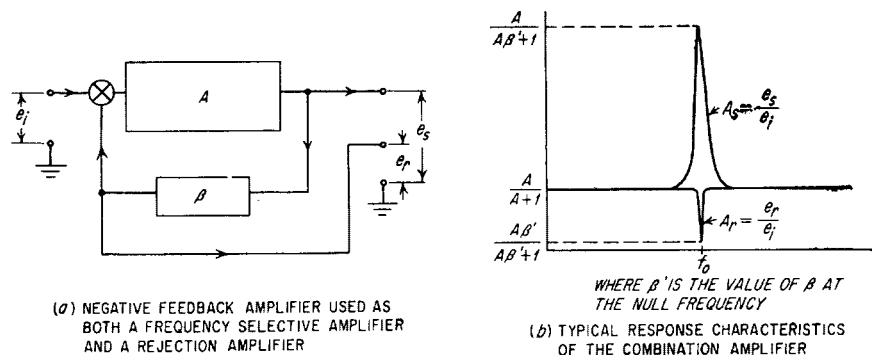
FIG. 16.51. Combination frequency-selective and rejection amplifier.

It should be noted that $Q_r \simeq Q_s$ for either type of null network. Figure 16.51 shows the relative frequency response of A_r and A_s for an amplifier which might serve as both a frequency-selective amplifier and a rejection amplifier. If the feedback network by itself does not have a complete null, the depth of the null for the rejection amplifier will be less than the depth of the null for the null network taken by itself. However, the reduction in attenuation is accompanied by a narrowing of the attenuation band. The actual depth of the null for the rejection amplifier which has a feedback network with an incomplete null is given by Eq. (16.43).

$$\frac{\text{Flat}}{\text{Null}} = \frac{1 + A\beta'}{\beta'(1 + A)} \tag{16.43}$$

where β' = value of β at null frequency

Example 16.2

Design a parallel-T null network that has its null at 30 cps and which has the narrowest possible attenuation band (see Fig. 16.40).

[1] See Sec. 7.4f.

Solution

For the narrowest attenuation band, $k = 100$ and $b = 0.99$ (see Fig. 16.44).
From Eq. (16.35)

$$RC = \frac{1}{2\pi \times 30} = 0.00531$$

R can arbitrarily be made equal to any value, and the preceding equation can then be solved for C. For example, let $R = 10,000$ ohms.

$$C = \frac{0.00531}{10,000}$$
$$= 0.531 \times 10^{-6} = 0.531 \; \mu f$$

With reference to **Fig.** 16.40, the other circuit values are as follows:

$$\frac{C}{k} = 5,310 \; \mu\mu f \qquad kR = 1 \text{ megohm}$$

$$\frac{C}{b} = 0.536 \; \mu f \qquad bR = 9,900 \text{ ohms}$$

NOTE: The effective Q of this particular null network is approximately one-half; therefore the bandwidth between -3-db frequencies is approximately 60 cycles. With reference to Fig. 16.44, the transmission is equal to 0.707 at approximately $0.4f_o$ and $2.5f_o$. The phase shift at $0.4f_o$ is $-44°$, and the phase shift at $2.5f_o$ is $+44°$.

Attenuators and Equalizers

17.1. Fixed Attenuators. Networks which introduce a fixed amount of attenuation independent of frequency have extensive use. They can be designed to have
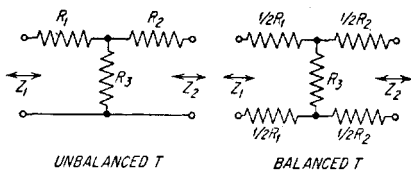
FIG. 17.1. T and H attenuator networks.
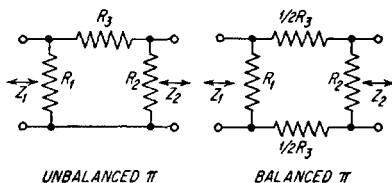
FIG. 17.2. π and O attenuator networks.

equal or unequal input and output impedances and to provide different amounts of attenuation.

Unbalanced and balanced T and π networks are shown in Figs. 17.1 and 17.2. For every ratio Z_1/Z_2 of the values of terminating impedances there is an associated minimum value of the ratio of input power to the attenuator to output power from the attenuator which can be realized.
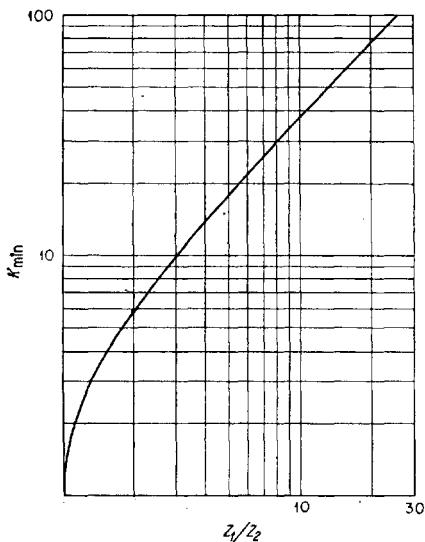
FIG. 17.3. Plot of the minimum possible values of K as a function of Z_1/Z_2 (refer to Figs. 17.1 and 17.2). If $K = K_{min}$, $R_1 = 0$ in Fig. 17.1 and $R_1 = \infty$ in Fig. 17.2.

$$K = \frac{\text{power into network}}{\text{power out of network}}$$

$$K_{min} = \frac{2Z_1}{Z_2} - 1 + 2\sqrt{\frac{Z_1}{Z_2}\left(\frac{Z_1}{Z_2} - 1\right)} \tag{17.1}$$

where K_{min} = minimum possible value of K for particular ratio of Z_1 to Z_2. The minimum possible values of the power ratio K as a function of Z_1/Z_2 are given in Fig. 17.3. There is no maximum value for the power ratio. It should be noted that Z_1 is always taken as the larger impedance and can be either the input or output impedance.

Since these networks can be made to have unequal input and output impedances, they are frequently used for impedance matching even though there is an associated power loss.

For the balanced and unbalanced T networks shown in Fig. 17.1 where $Z_1 \geq Z_2$, the values of R_1, R_2, and R_3 can be determined from Eqs. (17.2) to (17.6).

$$R_1 = \frac{Z_1(K + 1) - 2\sqrt{KZ_1Z_2}}{K - 1} \tag{17.2}$$

$$R_2 = \frac{Z_2(K + 1) - 2\sqrt{KZ_1Z_2}}{K - 1} \tag{17.3}$$

$$R_3 = \frac{2\sqrt{KZ_1Z_2}}{K - 1} \tag{17.4}$$

17-2

and if $Z_1 = Z_2$,

$$R_1 = R_2 = Z_1 \left(\frac{\sqrt{K} - 1}{\sqrt{K} + 1} \right) \tag{17.5}$$

$$R_3 = \frac{2Z_1 \sqrt{K}}{K - 1} \tag{17.6}$$

For the balanced and unbalanced π networks shown in Fig. 17.2 where $Z_1 \geq Z_2$, the values of R_1, R_2, and R_3 can be determined from Eqs. (17.7) to (17.11).

$$R_1 = \frac{(K - 1)Z_1 \sqrt{Z_2}}{(K + 1) \sqrt{Z_2} - 2 \sqrt{KZ_1}} \tag{17.7}$$

$$R_2 = \frac{(K - 1)Z_2 \sqrt{Z_1}}{(K + 1) \sqrt{Z_1} - 2 \sqrt{KZ_2}} \tag{17.8}$$

$$R_3 = \frac{K - 1}{2} \sqrt{\frac{Z_1 Z_2}{K}} \tag{17.9}$$

and if $Z_1 = Z_2$,

$$R_1 = R_2 = Z_1 \left(\frac{\sqrt{K} + 1}{\sqrt{K} - 1} \right) \tag{17.10}$$

$$R_3 = \frac{Z_1(K - 1)}{2 \sqrt{K}} \tag{17.11}$$

Example 17.1

Design a network to match a 500-ohm generator to a 200-ohm load with the minimum possible power loss.

Solution

1. Determine the ratio of Z_1 to Z_2 and the minimum possible value of the ratio of network input power to network output power.

$$\frac{Z_1}{Z_2} = \frac{500}{200} = 2.50$$

From Eq. (17.1)

$$K_{\min} = 2 \times 2.50 - 1 + 2 \sqrt{2.50(2.50 - 1)}$$
$$= 7.87$$

2. Determine the type of network to be used and the network values.

Since the type of network was not specified, an arbitrary choice might be an unbalanced T. It was stated that the network loss should be a minimum, therefore K must equal 7.87. In Fig. 17.3 it is stated that $R_2 = 0$ ohms when K is equal to its minimum value; therefore, it is only necessary to determine R_1 and R_3.

From Eqs. (17.2) and (17.3)

$$R_1 = \frac{500(7.87 + 1) - 2\sqrt{7.87 \times 500 \times 200}}{7.87 - 1}$$
$$= 387 \text{ ohms}$$

$$R_3 = \frac{2 \sqrt{7.87 \times 500 \times 200}}{7.87 - 1}$$
$$= 258 \text{ ohms}$$
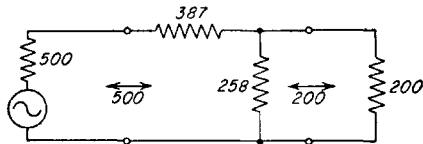
FIG. 17.4. Circuit for Example 17.1.

(Refer to Fig. 17.4.)

3. Determine the loss in the network.

$$\text{Loss} = 10 \log_{10} K$$
$$= 8.96 \text{ db}$$

Example 17.2

Design an unbalanced π attenuator with a loss of 20 db ($K = 100$) to operate between a 200-ohm line and a 500-ohm line.

Solution

1. Determine the ratio of Z_1 to Z_2 and the minimum possible value of K.

$$\frac{Z_1}{Z_2} = \frac{500}{200} = 2.50$$

From Example 17.1, K must be equal to or greater than 7.87.

2. Determine the network values.

Network was specified as an unbalanced π (see Fig. 17.2), and K is equal to 100. From Eqs. (17.7) to (17.9)
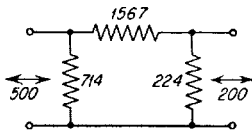
Fig. 17.5. Circuit for Example 17.2.

$$R_1 = \frac{(100 - 1)500 \sqrt{200}}{(100 + 1) \sqrt{200} - 2 \sqrt{100 \times 500}}$$
$$= 714 \text{ ohms}$$

$$R_2 = \frac{(100 - 1)200 \sqrt{500}}{(100 + 1) \sqrt{500} - 2 \sqrt{100 \times 200}}$$
$$= 224 \text{ ohms}$$

$$R_3 = \frac{(100 - 1)}{2} \sqrt{\frac{200 \times 500}{100}}$$
$$= 1{,}567 \text{ ohms}$$

(See Fig. 17.5.)

17.2. Amplitude Equalizers. Amplitude equalizers have an insertion loss which varies as some desired function of frequency, and consequently they are employed in electronic circuitry as a means of establishing or correcting the circuit gain characteristics.

In "Motion Picture Sound Engineering," Kimball[1] has provided an excellent treatment of amplitude equalizers, and the material in this section is based on his work.

On the assumption that an amplitude equalizer operates from a source impedance R_o and into a load impedance R_o, it is possible to design several equalizers having different configurations but which provide exactly the same attenuation characteristics as a function of frequency. The seven specific configurations for which design information is presented are:

1. Series-impedance type
2. Shunt-impedance type
3. Full-series type
4. Full-shunt type
5. T type
6. Bridged-T type
7. Lattice type

Shown in Fig. 17.6 are the required variations in these seven basic configurations for obtaining the type of attenuation characteristics indicated by the insertion loss curves in each column. The configurations in the last three rows have constant input and output impedances as a function of frequency, and the types in rows 3 and 4 have a constant input impedance.
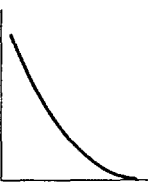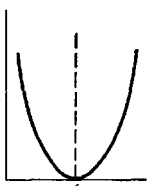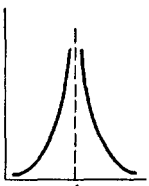
Although two examples are given to aid in the use of the design data, the following suggestions are included to further help in the design of the different types of equalizers.

[1] Harry Kimball, "Motion Picture Sound Engineering," Chap. 16. D. Van Nostrand Company, Inc., Princeton, N.J., 1938.
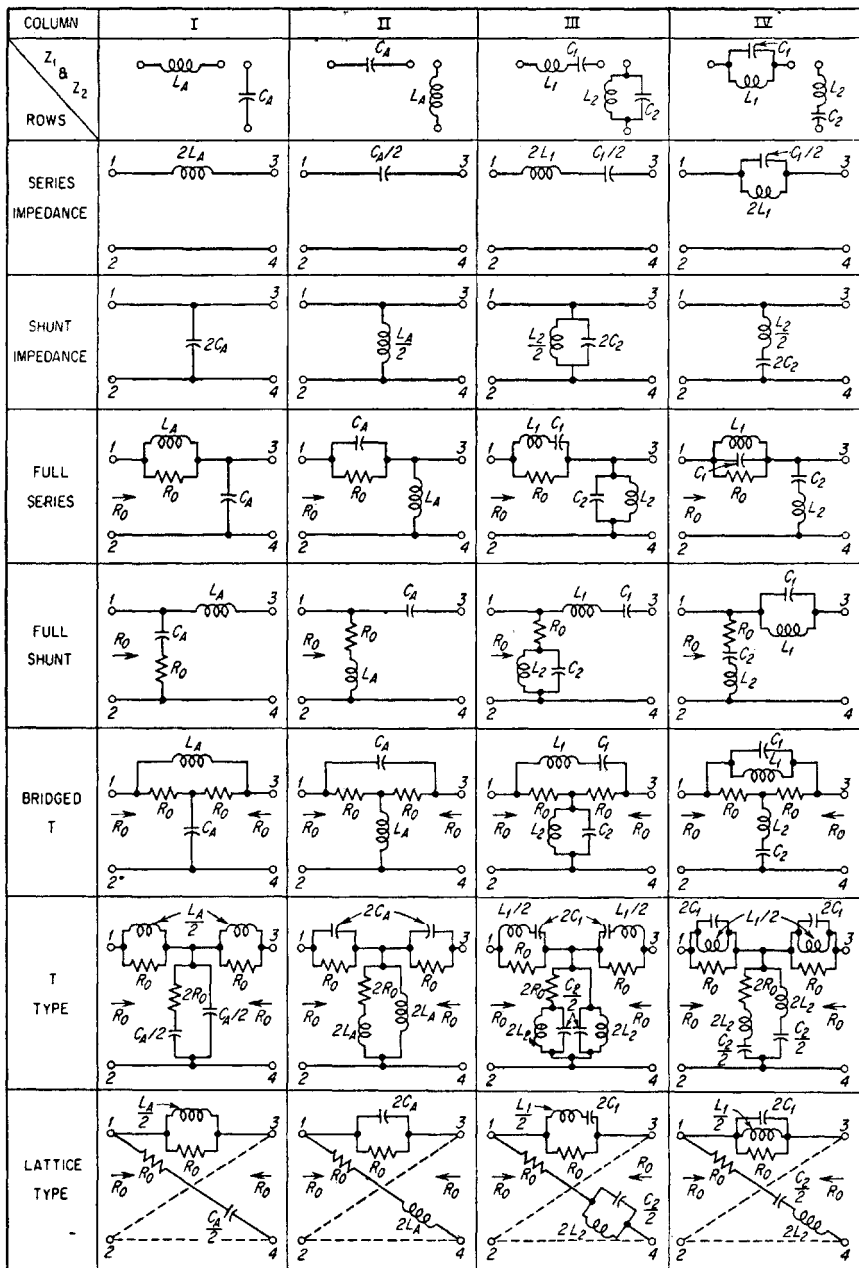
Fig. 17.6. Network configurations and formulas for attenuation equalizers. (*From "Motion Picture Sound Engineering," by Research Council of the Academy of Motion Picture Arts and Sciences, copyright 1938, D. Van Nostrand Company, Inc.*)

COLUMN	I	II	III	IV
INSERTION LOSS CHARACTERISTIC				
REFER TO	FIG. 17.7	FIG. 17.8	FIG. 17.9	FIG. 17.10
CURRENT RATIO $\left\|\left[\frac{I_L}{I_L}\right]^2\right\|$	$1+\left[\dfrac{f}{f_0}\right]^2$	$1+\left[\dfrac{f_0}{f}\right]^2$	$1+\left[\dfrac{\frac{f}{f_R}-\frac{f_R}{f}}{a-\frac{1}{a}}\right]^2$	$1+\left[\dfrac{a-\frac{1}{a}}{\frac{f}{f_R}-\frac{f_R}{f}}\right]^2$
DESIGN FORMULAE	$L_A=\dfrac{R_0}{2\pi f_0}=\dfrac{R_0}{\omega_0}$ $C_A=\dfrac{1}{2\pi f_0 R_0}=\dfrac{1}{\omega_0 R_0}$ $f_0=\dfrac{1}{2\pi\sqrt{L_A C_A}}$ $R_0=\sqrt{\dfrac{L_A}{C_A}}$		$L_1=L_A\dfrac{1}{a^2-1}$ $L_2=L_A\dfrac{a^2-1}{a^2}$ $C_1=C_A\dfrac{a^2-1}{a^2}$ $C_2=C_A\dfrac{1}{a^2-1}$	$L_1=L_A\dfrac{a^2-1}{a^2}$ $L_2=L_A\dfrac{1}{a^2-1}$ $C_1=C_A\dfrac{1}{a^2-1}$ $C_2=C_A\dfrac{a^2-1}{a^2}$
NOTES	$f_R=$ RESONANT FREQUENCY OF Z_1 AND Z_2 ARMS $f_0=$ FREQUENCY OF 3 DB INSERTION LOSS $f\ =$ ANY FREQUENCY $a=\dfrac{f_R}{f_0}=$ DEFINED AS GREATER THAN UNITY $R_0=$ EQUALIZER RESISTANCE INSERTION LOSS $=10$ LOG $\left\|\left[\dfrac{I_L}{I_L}\right]^2\right\|$ $L\ =$ INDUCTANCE IN HENRIES $C\ =$ CAPACITANCE IN FARADS			
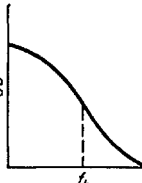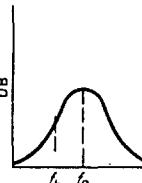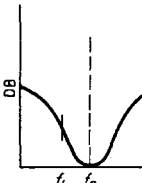
(b)

FIG. 17.6. (Continued)

(c)

FIG. 17.6. (Continued)

COLUMN	V	VI	VII	VIII
INSERTION LOSS CHARACTER-ISTIC				
REFER TO	FIG. 17.11	FIG. 17.12	FIG. 17.13	FIG. 17.14
CURRENT RATIO $\left\| \left[\dfrac{I_L}{I'_L} \right]^2 \right\|$	$1 + \dfrac{K^2-1}{1+K\left(\dfrac{f_b}{f}\right)^2}$	$1 + \dfrac{K^2-1}{1+K\left(\dfrac{f}{f_b}\right)^2}$	$1 + \dfrac{K^2-1}{1+K\left[\dfrac{\frac{f}{f_R}-\frac{f_R}{f}}{b-\frac{1}{b}}\right]^2}$	$1 + \dfrac{K^2-1}{1+K\left[\dfrac{b-\frac{1}{b}}{\frac{f}{f_R}-\frac{f_R}{f}}\right]^2}$
DESIGN FORMULAE	$L_1 = L_B \dfrac{K-1}{\sqrt{K}}$ $L_2 = L_B \dfrac{\sqrt{K}}{K-1}$ $C_1 = C_B \dfrac{\sqrt{K}}{K-1}$ $C_2 = C_B \dfrac{K-1}{\sqrt{K}}$		$L_1 = L_B \dfrac{K-1}{\sqrt{K}} \dfrac{b^2-1}{b^2}$ $L_2 = L_B \dfrac{\sqrt{K}}{K-1} \dfrac{1}{b^2-1}$ $C_1 = C_B \dfrac{\sqrt{K}}{K-1} \dfrac{1}{b^2-1}$ $C_2 = C_B \dfrac{K-1}{\sqrt{K}} \dfrac{b^2-1}{b^2}$	$L_1 = L_B \dfrac{K-1}{\sqrt{K}} \dfrac{1}{b^2-1}$ $L_2 = L_B \dfrac{\sqrt{K}}{K-1} \dfrac{b^2-1}{b^2}$ $C_1 = C_B \dfrac{\sqrt{K}}{K-1} \dfrac{b^2-1}{b^2}$ $C_2 = C_B \dfrac{K-1}{\sqrt{K}} \dfrac{1}{b^2-1}$

FOR ALL NETWORKS

$$R_0 = \sqrt{\frac{L_B}{C_B}} \qquad R_1 = R_0(K-1) \qquad R_2 = R_0 \frac{1}{K-1} \qquad R_6 = R_0 \frac{K}{K-1}$$

$$R_3 = R_0 \frac{K-1}{K+1} \qquad R_4 = R_0 \frac{K+1}{K-1} \qquad R_5 = R_0 \frac{K-1}{K}$$

$$L_B = \frac{R_0}{2\pi f_b} = \frac{R_0}{\omega_b} \qquad C_B = \frac{1}{2\pi f_b R_0} = \frac{1}{\omega_b R_0} \qquad f_b = \frac{1}{2\pi \sqrt{L_B C_B}}$$

NOTES

$f_R =$ RESONANT FREQUENCY OF Z_1 AND Z_2 ARMS

$f =$ ANY FREQUENCY

INSERTION LOSS $= 10$ LOG $\left\| \left[\dfrac{I_L}{I'_L} \right]^2 \right\|$

PAD LOSS $=$ MAXIMUM LOSS $= 20$ LOG K

$b = \dfrac{f_R}{f_b} =$ DEFINED AS GREATER THAN UNITY

$L =$ INDUCTANCE IN HENRIES

$C =$ CAPACITANCE IN FARADS

$f_b =$ FREQUENCY OF ONE-HALF PAD LOSS

$R_0 =$ EQUALIZER RESISTANCE

(d)

FIG. 17.6. (Continued)

For the equalizers treated in columns I and II of Fig. 17.6, the desired insertion loss at some frequency f must be specified. From either Fig. 17.7 or 17.8, whichever is applicable, this insertion loss can then be associated with a specific value of f/f_a.

$$LOSS = 10 \, LOG \left[1 + \left(\frac{f}{f_a} \right)^2 \right]$$
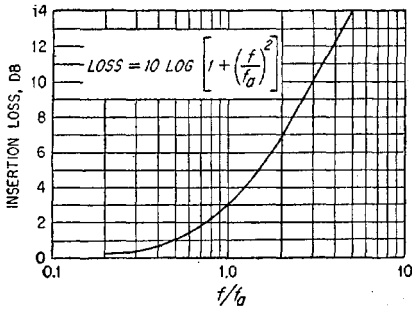
Fig. 17.7. Attenuation characteristics of networks shown in column I of Fig. 17.6. (*From "Motion Picture Sound Engineering," by Research Council of the Academy of Motion Picture Arts and Sciences, copyright 1938, D. Van Nostrand Company, Inc.*)

$$LOSS = 10 \, LOG \left[1 + \left(\frac{f_a}{f} \right)^2 \right]$$
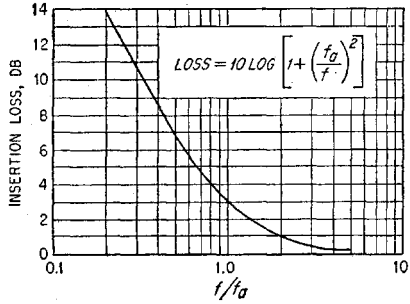
Fig. 17.8. Attenuation characteristics of networks shown in column II of Fig. 17.6. (*From "Motion Picture Sound Engineering," by Research Council of the Academy of Motion Picture Arts and Sciences, copyright 1938, D. Van Nostrand Company, Inc.*)

$$LOSS = 10 \, LOG \left[1 + \frac{\left(\frac{f}{f_R} - \frac{f_R}{f} \right)^2}{\left(\frac{a^2 - 1}{a} \right)^2} \right]$$
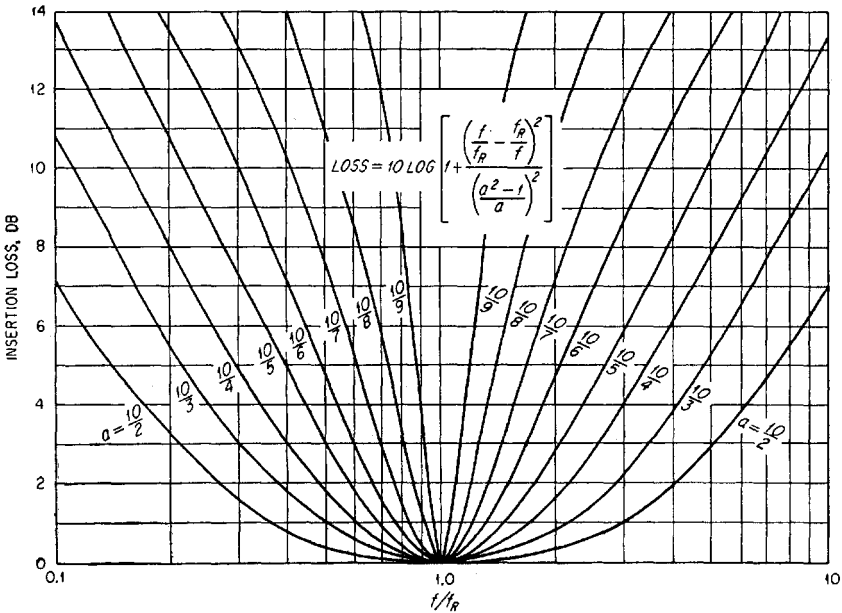
Fig. 17.9. Attenuation characteristics of networks shown in column III of Fig. 17.6. (*From "Motion Picture Sound Engineering," by Research Council of the Academy of Motion Picture Arts and Sciences, copyright 1938, D. Van Nostrand Company, Inc.*)

The value of f_a, which is required for the calculation of the equalizer circuit values, can then be determined since the values of f and f/f_a are known.

When working with equalizers of the types shown in columns III and IV of Fig. 17.6, the frequency of resonance f_R within the equalizer (associated with zero and infinite insertion losses, respectively) and the desired insertion loss at some specific value of f/f_R must be specified so that the proper attenuation curve and the asso-

$$LOSS = 10\,LOG \left[1 + \frac{\left(\frac{a^2-1}{a}\right)^2}{\left(\frac{f}{f_R} - \frac{f_R}{f}\right)^2} \right]$$
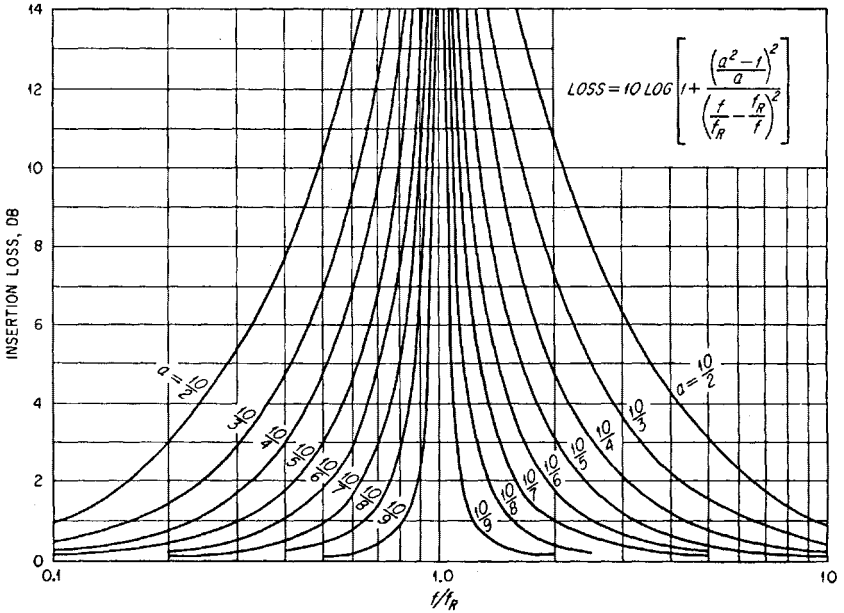
FIG. 17.10. Attenuation characteristics of networks shown in column IV of Fig. 17.6. *(From "Motion Picture Sound Engineering," by Research Council of the Academy of Motion Picture Arts and Sciences, copyright 1938, D. Van Nostrand Company, Inc.)*

$$LOSS = 10\,LOG \left[1 + \frac{K^2-1}{1+K(f_b/f)^2} \right]$$
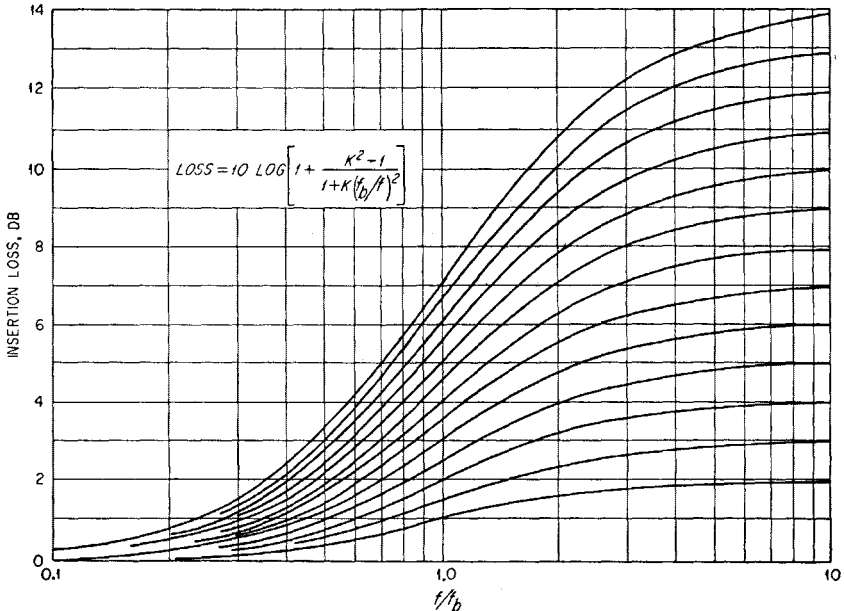
FIG. 17.11. Attenuation characteristics of networks shown in column V of Fig. 17.6. *(From "Motion Picture Sound Engineering," by Research Council of the Academy of Motion Picture Arts and Sciences, copyright 1938, D. Van Nostrand Company, Inc.)*

$$LOSS = 10\,LOG\left[1 + \frac{K^2-1}{1+K\left(f/f_b\right)^2}\right]$$
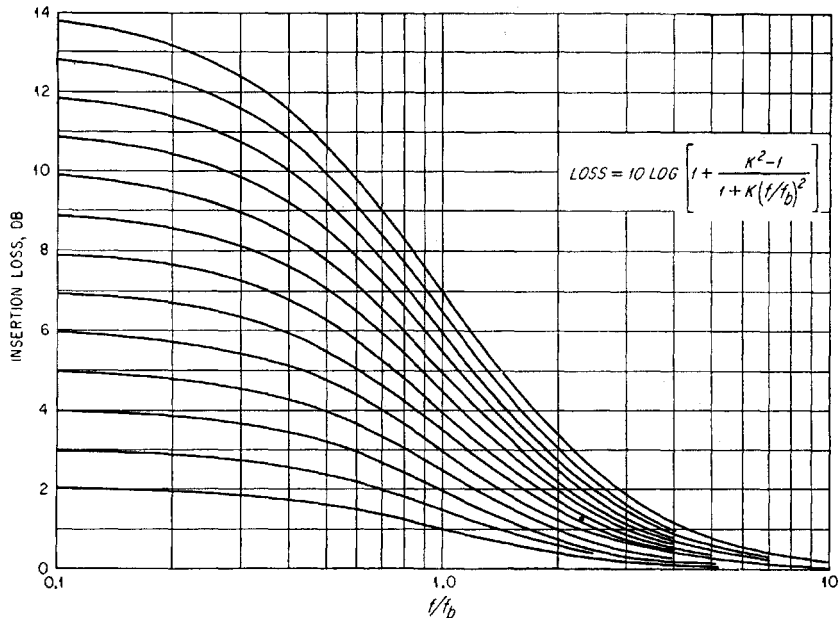
FIG. 17.12. Attenuation characteristics of networks shown in column VI of Fig. 17.6.
(*Figs. 17.12, 17.13, and 17.14 from "Motion Picture Sound Engineering," by Research Council of the Academy of Motion Picture Arts and Sciences, copyright 1938, D. Van Nostrand Company, Inc.*)
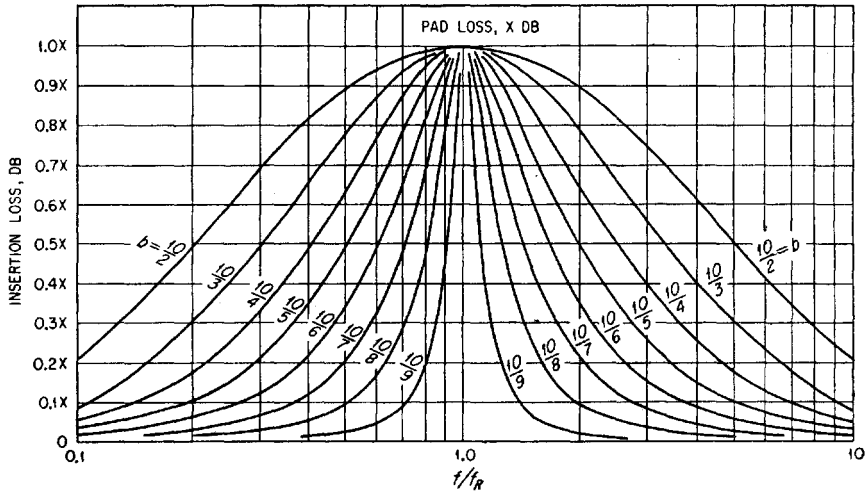
FIG. 17.13. Attenuation characteristics of networks shown in column VII of Fig. 17.6.

ciated value of a can be determined from either Fig. 17.9 or 17.10, whichever is applicable. The value of f_a, which is also required for the calculation of the equalizer circuit values, can be determined by dividing f_R by a.

The first step in the design of equalizers in columns V and VI of Fig. 17.6 is to specify the maximum desired loss and the loss at some specific frequency f. The loss at f can then be associated with a specific value of f/f_b on the curve having the desired maximum loss in either Fig. 17.11 or 17.12, whichever is applicable. The value of f_b,

which is required for the calculation of the equalizer circuit values, can then be determined since the values of f and f/f_b are known.

To design equalizers shown in columns VII and VIII of Fig. 17.6, the maximum insertion loss, the frequency of resonance f_R within the equalizer (associated with the
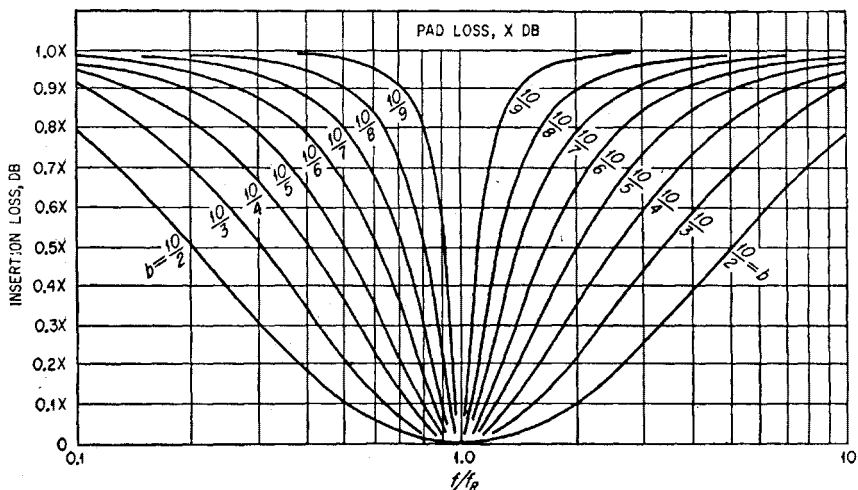
FIG. 17.14. Attenuation characteristics of networks shown in column VIII of Fig. 17.6.

maximum and zero insertion losses, respectively), and the desired insertion loss at some value of f/f_R must be specified.   It is then possible to establish the proper curve and the associated value of b from either Fig. 17.13 or 17.14, whichever is applicable.

FIG. 17.15. Full shunt equalizer for Example 17.3.

The value of f_b which is also required for the calculation of the equalizer circuit values can be determined by dividing f_R by b.

When working in columns V to VIII, the frequency f_b is the frequency at which the pad loss in decibels is one-half the maximum loss in decibels, e.g., if the maximum loss is 8 db, f_b is the frequency at which the loss is 4 db.

Example 17.3

Design a full-shunt equalizer of the type shown in column II of Fig. 17.6 which has an insertion loss of 4 db at 1 kc and an input impedance of 500 ohms.

Solution

From Fig. 17.8 $f/f_a = 0.8$ for an insertion loss of 4 db.   Therefore,

$$f_a = \frac{f}{0.8}$$

$$= \frac{1,000}{0.8} = 1,250 \text{ cycles}$$

$R_o = 500$ ohms (from statement of problem)

$$L_a = \frac{500}{2 \times 3.14 \times 1,250} = 63.7 \times 10^{-3} \text{ henry, or } 63.7 \text{ mh}$$

$$C_A = \frac{1}{2 \times 3.14 \times 1,250 \times 500} = 0.255 \times 10^{-6} \text{ farad, or } 0.255 \text{ } \mu f$$

(See Figs. 17.15 and 17.16.)
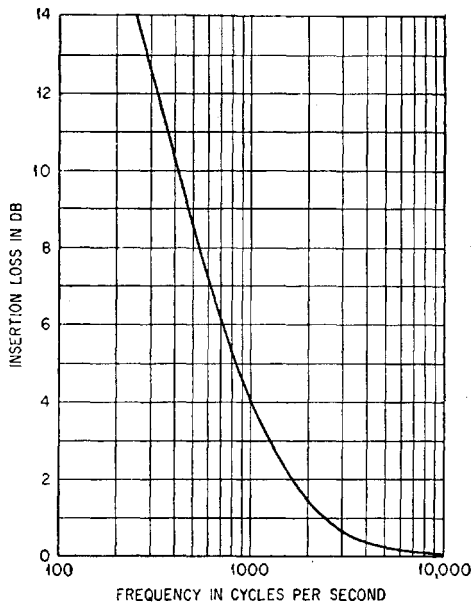
FIG. 17.16. Insertion-loss characteristics of network shown in Fig. 17.15.

Example 17.4

Design an amplitude equalizer with attenuation characteristics as indicated in column VIII of Fig. 17.6 and of the bridged-T type which will introduce zero attenuation at 5 kc (frequency of equalizer resonance), 14-db attenuation at 3.5 kc, and 20-db attenuation at frequencies far above and far below 5 kc. The equalizer should have a characteristic impedance of 200 ohms.

Solution

1. Determine f_R, b, f_b, and K.

$$f_R = 5,000 \text{ cycles (frequency of resonance and no attenuation)}$$

At 3,500 cps, $f/f_R = 3,500/5,000 = 0.7$. The maximum attenuation has been specified as being equal to 20 db; therefore, $X = 20$ db in Fig. 17.14, and the desired attenuation of 14 db at 3,500 cps is equal to $0.7X$. The curve for $b = 1\%$ satisfies these conditions. The values of f_b and K are determined as follows:

$$b = \frac{f_R}{f_b}$$

$$f_b = \frac{5,000}{1.25} = 4,000 \text{ cycles}$$

$$20 \log_{10} K = \text{maximum pad loss} = 20$$
$$K = 10$$

2. Determine the values of the elements in the bridged T.
   Refer to Fig. 17.6.

$$R_o = 200 \text{ ohms} \quad \text{(specified in statement of problem)}$$

$$L_B = \frac{200}{2 \times 3.14 \times 4,000} = 7.96 \times 10^{-3} \text{ henry, or } 7.96 \text{ mh}$$

$$C_B = \frac{1}{2 \times 3.14 \times 4,000 \times 200} = 0.199 \times 10^{-6} \text{ farad, or } 0.199 \ \mu\text{f}$$
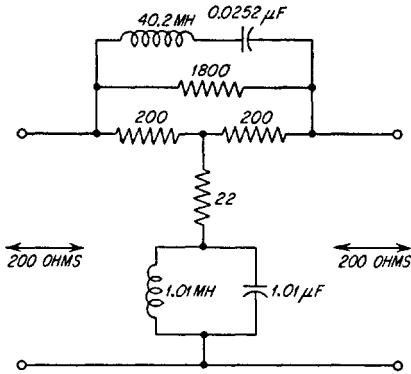
FIG. 17.17. Bridged-T equalizer for Example 17.4.

$$L_1 = 7.96 \times \frac{10 - 1}{\sqrt{10}} \times \frac{1}{1.25^2 - 1}$$
$$= 40.2 \text{ mh}$$

$$L_2 = 7.96 \times \frac{\sqrt{10}}{10 - 1} \times \frac{1.25^2 - 1}{1.25^2}$$
$$= 1.01 \text{ mh}$$

$$C_1 = 0.199 \times \frac{\sqrt{10}}{10 - 1} \times \frac{1.25^2 - 1}{1.25^2}$$
$$= 0.0252 \ \mu\text{f}$$

$$C_2 = 0.199 \times \frac{10 - 1}{\sqrt{10}} \times \frac{1}{1.25^2 - 1}$$
$$= 1.01 \ \mu\text{f}$$

$$R_1 = 200(10 - 1) = 1{,}800 \text{ ohms}$$

$$R_2 = 200 \times \frac{1}{10 - 1} = 22.2 \text{ ohms}$$

(Refer to Figs. 17.17 and 17.18.)
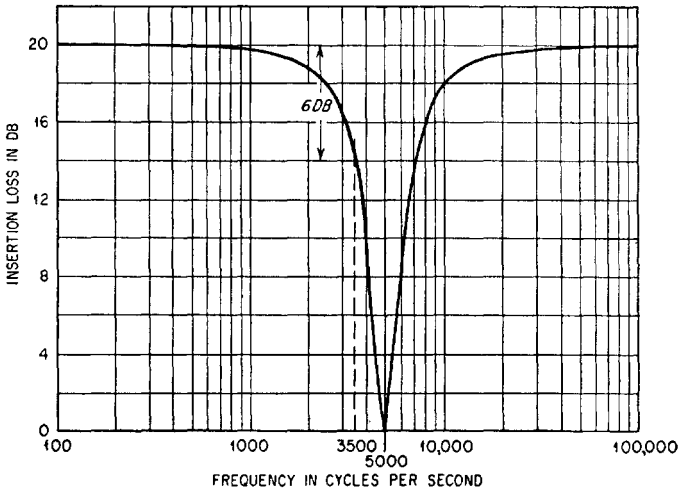
FIG. 17.18. Insertion-loss characteristics of network shown in Fig. 17.17.

17.3. Phase Equalizers. The types of phase equalizers treated are those which theoretically introduce either zero or a fixed amount of attenuation at all frequencies. They can therefore be added to existing circuits for phase correction without distorting the gain characteristics.

The shape of electrical impulses which contain many frequency components can be distorted in passing through an electrical circuit even though the circuit has the same gain for the different frequency components. If such is the case, the distortion is due to unequal transmission delays for the different frequency components. This type of distortion is called *phase distortion* and can be corrected by adding a network which will cause the total transmission period for all frequencies to be identical. The added network must, therefore, be a network in which the phase characteristics can be controlled.

Equal transmission periods for different frequency components through a circuit stipulates that the circuit must introduce either no phase shift or an amount of phase shift which is directly proportional to frequency. This is identical to stating that the transmission period must be either zero or a constant amount at all frequencies.

Four different configurations of phase equalizers[1] are shown in Figs. 17.19 to 17.21. It should be noted that the four-terminal networks can be used in only those applications in which the input and output circuits are either both balanced or are in no way connected to each other.

The circuit in Fig. 17.19 introduces an insertion loss of 6 db and does not have constant input and output impedances as a function of frequency. In addition, the
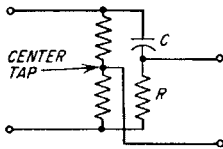
$$f_0 = \frac{1}{2\pi RC}$$
WHERE f_0 IS THE FREQUENCY AT WHICH
THE PHASE SHIFT β IS EQUAL TO -90°

FIG. 17.19. Phase equalizer with a fixed insertion loss of 6 db. The phase characteristics are exactly the same as for the lattice network shown in Fig. 17.20, provided the output is not loaded.
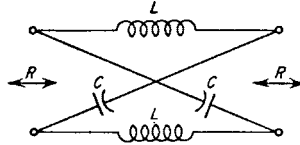
$$L = \frac{aR}{2\pi}$$
$$C = \frac{L}{R^2}$$
$$TAN \frac{\beta}{2} = af$$

$a = 1/f_0$
WHERE f_0 IS THE FREQUENCY AT
WHICH NETWORK PHASE SHIFT
IS -90°

FIG. 17.20. Phase-shift network with zero attenuation. Refer to Fig. 17.22 for phase characteristics.
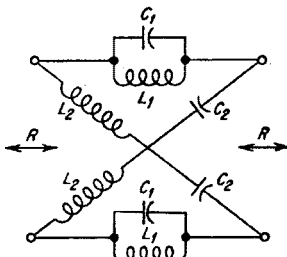
(a) LATTICE NETWORK

$$L_1 = \frac{aR}{2\pi} \qquad C_1 = \frac{b}{2\pi aR}$$
$$L_2 = R^2 C_1 \qquad C_2 = \frac{L_1}{R^2}$$
$$TAN \frac{\beta}{2} = \frac{af}{1-bf^2}$$
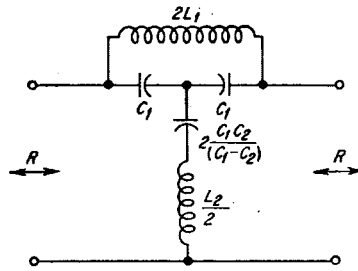
(b) BRIDGED-T EQUIVALENT TO LATTICE IF $C_1 > C_2$

$b = 1/f_0^2$ WHERE f_0 IS THE FREQUENCY
AT WHICH THE NETWORK
PHASE SHIFT IS -180°

FIG. 17.21. Phase-shift network with zero attenuation. Refer to Fig. 17.23 for phase characteristics.

phase-shift curve for the circuit, Fig. 17.22, is based on there being no terminating impedance.

A center-tapped transformer secondary winding could be substituted for the resistor in Fig. 17.19, provided the amplitude and phase characteristics of the transformer were acceptable.

The networks shown in Figs. 17.20 and 17.21 have constant input and output characteristic impedances as a function of frequency and provide phase shift without attenuation. Figures 17.22 and 17.23 indicate the phase characteristics which can be obtained.

The phase equalizers shown in Figs. 17.19 to 17.21 introduce a lagging phase shift.

[1] These networks are also referred to as *all-pass* filters.
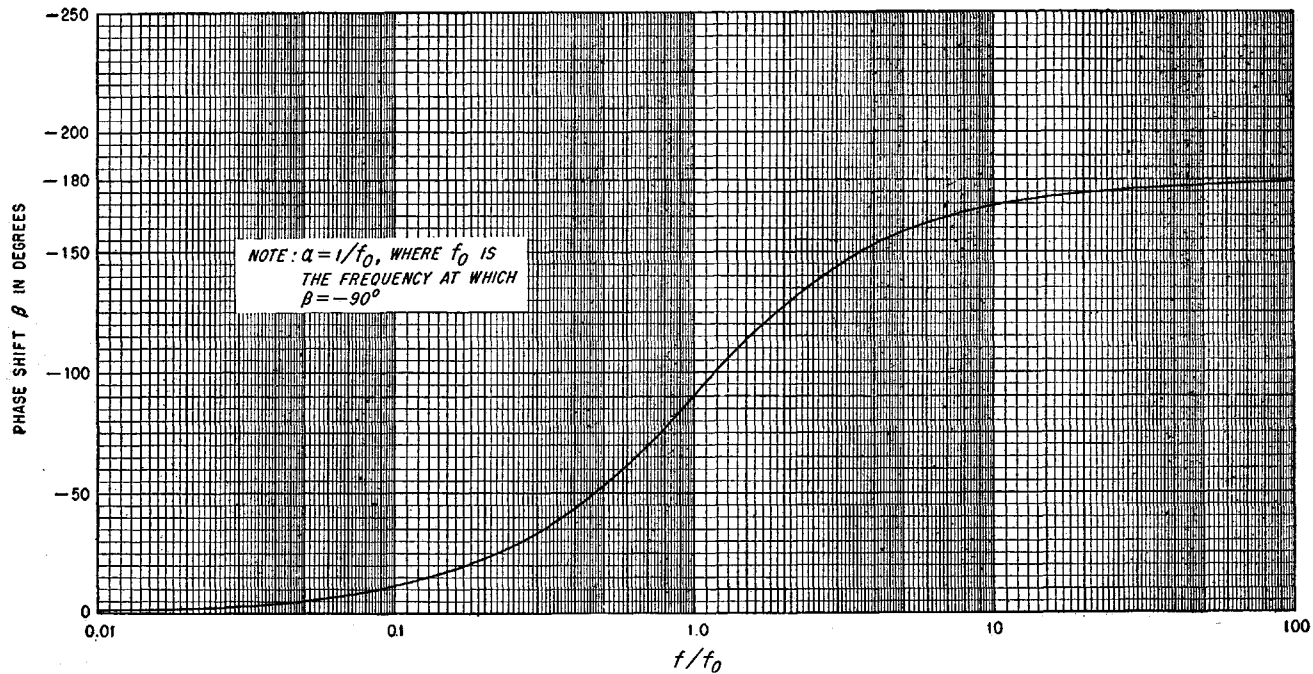
FIG. 17.22. Phase shift in an all-pass filter of the type shown in Fig. 17.20.

NOTE: $b=1/f_0^2$, WHERE f_0 IS THE FREQUENCY AT WHICH $\beta=180°$. IF $af_0 \leq 1$, THE THREE TERMINAL NETWORK SHOWN IN FIG. 17.21 CAN BE USED
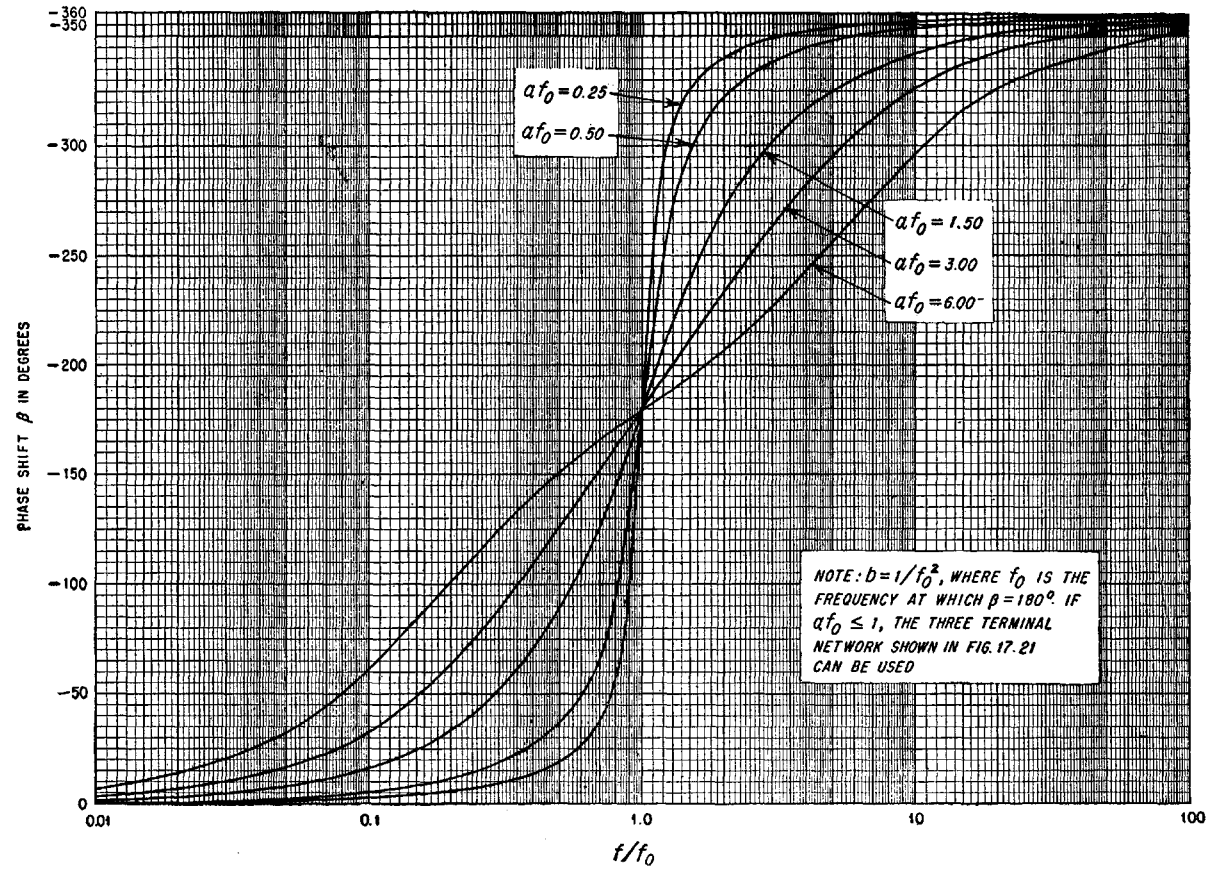
FIG. 17.23. Phase shift in an all-pass filter of the type shown in Fig. 17.21.

Example 17.5

Assume that intelligence must be transmitted in the 10- to 20-kc frequency band and that the circuit employed introduces phase shift in accordance with the following tabulation.

Phase Shift	Frequency
−27°	10 kc
−43.5°	15 kc
−63°	20 kc

Design a phase equalizer of the lattice type with a characteristic impedance of 1,000 ohms for use with this circuit.

Solution

1. Determine the required phase characteristics of the phase equalizer.

The departure from linear phase shift as a function of frequency for the existing circuit must first be determined. Since the phase shift at 20 kc is −63°, the phase shift at 10 kc should be $\frac{1}{2} \times -63$, or −31.5°, and the phase shift at 15 kc should be $\frac{3}{4} \times -63$, or −47.25°. The existing network therefore introduces a phase error of +4.5° at 10 kc and +3.75° at 15 kc.

Phase Error	Frequency
+4.5°	10 kc
+3.75°	15 kc
0°	20 kc

The phase equalizer must therefore exhibit the inverse characteristics, i.e.,

Phase Error	Frequency
−4.5°	10 kc
−3.75°	15 kc
0°	20 kc

2. Determine from Figs. 17.22 and 17.23 if the required conditions tabulated in step 1 can be satisfied with either of the networks shown in Figs. 17.20 or 17.21.

Since the network shown in Fig. 17.20 is simpler, the curve shown in Fig. 17.22 should first be examined.

The procedure is to determine if the phase shift in the equalizer at any three values of f/f_o, which are related in the same proportions as are 10, 15, and 20 kc, will depart from linear phase shift as a function of frequency by the desired amount. A few experimental groups of values of f/f_o reveal that the phase shifts for f/f_o equal to 0.4, 0.6, and 0.8 are equal to −43, −61.5, and −77°, respectively, and satisfy the specified requirements. This is true since a phase shift of −77° at $f/f_o = 0.8$ requires that the phase shift be −38.5 and −57.75° at f/f_o equal to 0.4 and 0.6, respectively, for linear phase characteristics. The phase equalizer therefore introduces phase errors of −4.5 and −3.75° when f/f_o is equal to 0.4 and 0.6, respectively. It should be apparent that the three values of f/f_o, that is, 0.4, 0.6, and 0.8, correspond to f being equal to 10, 15, and 20 kc, respectively.

3. Determine f_o and the values for the lattice elements.

$$\frac{f}{f_o} = 0.8 \text{ (at } f = 20 \text{ kc)}$$

$$f_o = 25 \text{ kc}$$

From Fig. 17.22

$$a = \frac{1}{25,000} = 4 \times 10^{-5}$$

From Fig. 17.20

$$L = \frac{4 \times 10^{-5} \times 10^3}{2 \times 3.14}$$

$$= 6.37 \times 10^{-3} \text{ henry, or } 6.37 \text{ mh}$$

$$C = \frac{6.37 \times 10^{-3}}{10^6}$$

$$= 6,370 \times 10^{-12} \text{ farad, or } 6,370 \ \mu\mu f$$

The lattice network is shown in Fig. 17.24.

FIG. 17.24. Lattice network for Example 17.5.

Principles of Feedback

18.1. Introduction. A *feedback system* is one in which some function of the output of some part of the system is fed back as a secondary input to the system so as to affect its own value (see Fig. 18.1).

In a feedback system it is possible to employ either positive or negative feedback. A *positive* feedback system is one in which the phases or polarities of the primary input and the signal fed back are such that they add in the system input mixer[1] (see Fig. 18.1) and increase the effective gain between the primary input and the output. A *negative* feedback system is one in which the phases or polarities of the primary input and the signal fed back are such that the net system input is less than without feedback, which consequently results in a reduction of the effective gain between the primary input and the output.

Fig. 18.1. Block diagram of a simple feedback system.

18.2. System Characteristics Affected by Feedback. The introduction of feedback modifies the characteristics of a system in addition to changing its gain. In general, negative feedback is employed to improve the behavior of a system. For example, with negative feedback it is possible to increase the bandwidth of an amplifier, improve its linearity and consequently decrease the amount of intermodulation and distortion, reduce its output impedance, reduce the output noise which has been introduced within the amplifier, and obtain improved gain stabilization for circuit value changes. The use of positive feedback is usually limited to either oscillators or to special types of feedback amplifiers in which there is a positive feedback system completely enclosed within a negative feedback system so as to provide an improvement in linearity (see Sec. 18.2b).

Considerable improvement can be realized in certain system characteristics with a properly designed negative feedback system. The actual characteristics of any system employing feedback are unique to that particular system. Consequently, the application of negative feedback does not ensure an improvement in the characteristics unless certain conditions are satisfied. For example, in most systems it is possible to employ enough negative feedback at midband to cause the system to become regenerative at certain frequencies associated with the skirts of the amplifier passband. This regeneration or positive feedback is due to excessive phase shift of the signal through the amplifier and the feedback path and usually occurs at those frequencies where the slopes of attenuation versus frequency are greatest. The positive feedback at these frequencies may cause the system to oscillate, or it may only adversely affect certain system characteristics which cause the system to compare unfavorably with the system in the absence of feedback. If the amount of feedback were reduced or if special equalizing networks were employed in either the amplifier or feedback path so as to sufficiently reduce the phase shift at these critical frequencies, the amount of positive feedback would be reduced and the system char-

[1] Throughout this material the mixer has been considered as performing the addition of the input signal and the signal fed back. This is indicated by the plus signs.

acteristics would be much better than without feedback. It is therefore necessary to analyze individually any given system having any amount or polarity of feedback.

18.2a. Gain and Bandwidth. The gain of a system with feedback is given by Eq. (18.1).

$$A_f = \frac{A}{1 - A\beta} \tag{18.1}$$

where A_f = system gain with feedback
A = system gain without feedback
β = gain from system output to secondary system input

NOTE: As the terms A_f, A, and β contain frequency-variant terms, they must be written as functions of $j\omega$ to permit the proper evaluation of Eq. (18.1). The sign of $A\beta$ is positive for positive feedback and negative for negative feedback.

In a negative feedback amplifier in which $|A\beta| \gg 1$, it can be seen from Eq. (18.1) that the value of A_f is very nearly equal to $1/\beta$ and is relatively independent of the amplifier characteristics. In such a system, the attenuation versus frequency characteristics of the feedback path β should be the inverse of the desired system characteristics. As an example, the use of a null network in the negative feedback path of a high-gain amplifier will cause the feedback amplifier to have the approximate inverse characteristic, i.e., a narrow passband.
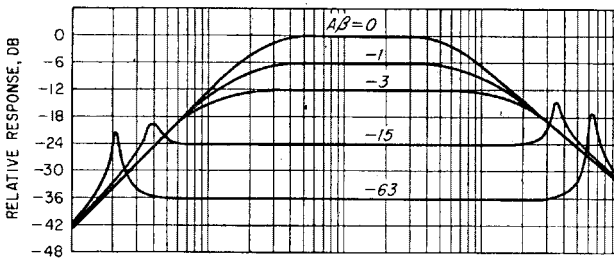
FIG. 18.2. Typical response curves of a negative feedback amplifier for several values of $A\beta$. The values of $A\beta$ are the midband values.

There are no simple rules which relate the exact bandwidth with and without feedback. In each application of feedback it is necessary to make a plot of A_f versus frequency as determined from Eq. (18.1) to determine the bandwidth of the system with feedback.

Considerable care must be exercised in the application of negative feedback since it is possible to have negative feedback at midband and positive feedback at frequencies somewhat removed from midband. This is due to the excessive phase shift in the amplifier and feedback path which causes the sign of $A\beta$ to change. The application of negative feedback usually causes the amplifier gain characteristics to vary in a manner similar to that illustrated by the curves shown in Fig. 18.2. The peaks at either end of the passband occur when $A\beta$ has values of unity or less and phase angles equal to or approaching 180°. As can be seen from the figure, the system becomes regenerative at both extremes of the passband for large amounts of feedback. If the amount of regeneration is insufficient to cause oscillation, the result is merely the severe humps in the amplitude response.

18.2b. Distortion. The linearity of an amplifier can be improved appreciably by the application of negative feedback as indicated by Eq. (18.2).

$$D_f = \frac{DA_f}{A} \tag{18.2}$$

where D_f = per cent distortion with feedback

D = per cent distortion without feedback

A reduction in the amplitude distortion also causes a reduction in the intermodulation which takes place in a nonlinear amplifier whenever two or more frequency components are amplified simultaneously.

Frequently a combination of both positive and negative feedback is employed to achieve an even further reduction in distortion. The positive and negative feedback paths cannot be around the same stages, or the effect is as though only a single feedback loop were present. The positive feedback loop is usually placed around one stage, and the negative feedback loop is designed to include the stage with the positive feedback in addition to one or two additional stages. An example of both positive and negative feedback is shown in Fig. 18.3a. The positive feedback is used to increase the system gain, and the negative feedback is employed so that the system gain is reduced to the desired value. The resulting linearity is better than would

(a) POSITIVE FEEDBACK AROUND ONE STAGE AND NEGATIVE FEEDBACK AROUND TWO STAGES

(b) NEGATIVE FEEDBACK AROUND TWO STAGES
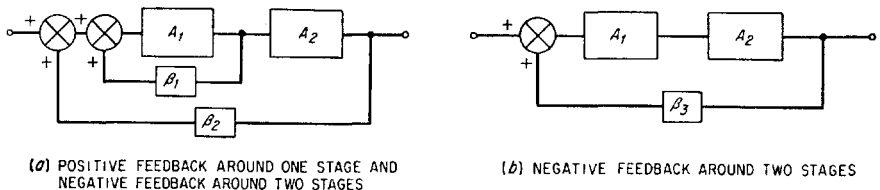
FIG. 18.3. Feedback amplifier configurations in which the sign of $A_1\beta_1$ is positive and the signs of $A_1A_2\beta_2$ and $A_1A_2\beta_3$ are negative.

have been realized if the same system gain had been achieved by negative feedback only as shown in Fig. 18.3b.

For Fig. 18.3a

$$D_1' = \frac{D_1 A_f}{A_1 A_2} \tag{18.3}$$

$$D_2' = \frac{D_2(1 - A_1\beta_1)A_f}{A_1 A_2} \tag{18.4}$$

For Fig. 18.3b

$$D_1' = \frac{D_1 A_f}{A_1 A_2} \tag{18.5}$$

$$D_2' = \frac{D_2 A_f}{A_1 A_2} \tag{18.6}$$

where D_1 = per cent distortion in A_1 without feedback

D_2 = per cent distortion in A_2 without feedback

D_1' = per cent distortion in over-all feedback system due to D_1

D_2' = per cent distortion in over-all feedback system due to D_2

It can be seen that if the amounts of feedback in the two systems shown in Fig. 18.3 are adjusted so that the over-all values of gain A_f are identical, the distortion in the output D_1' due to the distortion in A_1 is identical in each case. The distortion in the output D_2' due to the distortion in A_2 can be much less in the system employing both positive and negative feedback, however, as can be seen by comparing Eqs. (18.4) and (18.6).

18.2c. Noise. The amplitude of the noise appearing at the output of a feedback amplifier due to noise introduced at some point in the amplifier is given by Eq. (18.7).

$$N_o = N_i \frac{A_n}{1 - A\beta} \tag{18.7}$$

where N_o = amplitude of noise at output

$\quad$ N_i = amplitude of noise at point it is introduced in amplifier

$\quad$ A = over-all gain without feedback

$\quad$ A_n = gain without feedback between points at which noise is injected and output of amplifier

NOTE: Since A and β contain frequency-variant terms, they must be written as functions of $j\omega$ to permit the evaluation of Eq. (18.7).

18.2d. Output Impedance. The output impedance of a system can be changed considerably by the addition of feedback. If negative feedback is employed and if the signal fed back is a portion of the output voltage, the amplifier output impedance will be lower than without feedback. If negative feedback is employed and if the signal fed back is proportional to the output current, the output impedance will be greater than without feedback. The signal fed back in the case of current feedback is a voltage which is developed across a resistor in series with the load.

The output impedance of a system employing current and/or voltage feedback can be determined from Eqs. (18.8) to (18.10).

For voltage feedback only

$$Z_f = \frac{Z}{1 - K\beta_1} \tag{18.8}$$

For current feedback only

$$Z_f = Z - K\beta_2 R_f \tag{18.9}$$

For both voltage and current feedback

$$Z_f = \frac{Z - K\beta_2 R_f}{1 - K\beta_1} \tag{18.10}$$

where Z = output impedance of amplifier without feedback.

$\quad$ Z_f = output impedance of amplifier with feedback.

$\quad$ K = gain of unloaded amplifier without feedback. If loaded amplifier gain A without feedback is known, value of K can be determined from relationship $K = A + AZ/R_L$ where R_L = value of load resistance.

$\quad$ β_1 = ratio of voltage fed back to secondary input of amplifier to output voltage.

$\quad$ β_2 = ratio of voltage fed back to secondary input of amplifier to voltage developed across the resistor R_f in series with the load for providing current feedback.

NOTE: The signs of $K\beta_1$ and $K\beta_2$ are positive for positive feedback and negative for negative feedback.

18.2e. Stability. In feedback systems employed as amplifiers, the term *stability* indicates whether or not the system will oscillate. While a system having no feedback paths will always be stable, the application of feedback may cause the system to oscillate.

The amplitude and phase characteristics of an amplifier and its feedback path vary as a function of frequency, and for this reason the feedback may be positive at some frequencies and negative at others. Therefore, a system which has negative feedback at midband may have positive feedback at frequencies somewhat removed from midband. Also, if certain criteria are satisfied, the system will oscillate and is said to be unstable. Methods of performing system-stability analyses are discussed in Sec. 18.4.

18.2f. Transient Response. For different steady-state input functions, a system will have associated steady-state output functions. However, if an already existent steady-state input function to a physically realizable system is changed to some other steady-state input function, the output function will not instantaneously assume the

final steady-state value associated with the new steady-state input function. The *transient response* of the system refers to the system output response during the transitional period between the instant the input function departs from the initial steady-state function and some time later when the output function has essentially approached its final steady-state condition.

For optimum system-design analysis, the various transient responses should be determined for the particular transient input functions which are to be realized in the physical system; however, this is usually not feasible because of the complex and varied nature of these transient functions. In lieu of performing such an exhaustive analysis, the normal procedure is to determine the system response for one or more specific transient input functions of which a step-function input is usually included. The transient response to a step-function input is usually referred to as being either critically damped, overdamped, or underdamped. A system having a transfer function containing a d-c term is critically damped if, for a step-function input, the output has the minimum possible rise time without overshooting the final or steady-state value. For the same step-function input, an overdamped system will have a longer rise time but, as in the critically damped case, will have no overshoots. An underdamped system will have a shorter rise time than the critically damped system and will overshoot the steady-state value. The overshoot is followed by a decaying oscillation about the final value. Typical waveforms illustrating overdamping, critical damping, and underdamping are shown in Fig. 18.4. In summation, the transient response of an underdamped system to a step-function input will contain a damped frequency component as shown in both Figs. 18.4 and 18.14. If the system damping or loading is increased, a degree of damping will be found which will cause the damped frequency term to disappear. For this

FIG. 18.4. Typical waveforms illustrating different degrees of damping.

particular value of damping, the system is described as being critically damped. If the system loading is further increased, the system will be overdamped and the transient response will indicate additional damping and, as in the case with critical damping, will have no damped frequency term.

The transient response of any system employing R, L, and C components, with or without feedback, can fall in any one of the three damping categories. The transient response of a feedback amplifier employing only R and L components or R and C components to a step-function input can likewise fall in any one of the three categories. Methods of analyzing the transient response of a system to a step-function input are given in Sec. 18.4.

18.3. Transfer Functions. The *transfer function* or gain for a given network or system is the ratio of the system output to the system input provided that the system input, output, and all currents and voltages within the network are initially equal to zero.

18.3a. Transfer Functions Expressed as Functions of Frequency $(j\omega)$ and the Complex Variable (s). The transfer function is usually written as a function of either frequency or the complex variable. For a system having sine-wave inputs, it is convenient to work with a transfer function which has been written as a function of frequency. If the input to the system is nonsinusoidal, the transfer function is written as a function of the complex variable.

Instead of performing an analysis in either the frequency or complex-variable

domain, it is also possible to make the analysis in the time domain. Equations
in the time domain, however, are in integro-
differential form, and their solutions can be very
time-consuming. For this reason, it is recom-
mended that the equations be written as functions
of either frequency or the complex variable.

Frequency Domain. If the system input is
sinusoidal, its transfer function is usually written
as a function of $j\omega$. The resulting transfer func-
tion establishes both the gain and phase charac-
teristics of the system as a function of frequency.
For example, the transfer function of the active
network shown in Fig. 18.5 is given by Eq. (18.11).
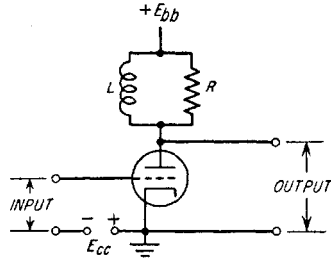
Fig. 18.5. Single-stage amplifier.

$$\frac{E_o(j\omega)}{E_i(j\omega)} = -\mu \frac{j\omega(L/r_p)}{1 + j\omega L \left(\dfrac{1}{R} + \dfrac{1}{r_p}\right)}$$

$$= -K \frac{j\omega}{\left(j\omega + \dfrac{1}{T}\right)} \tag{18.11}$$

where

$$K = \frac{\mu R}{r_p + R}$$

$$T = L\left(\frac{1}{R} + \frac{1}{r_p}\right)$$

$$\omega = 2\pi f$$

The amplitude and phase characteristics of the system transfer function can easily
be determined from Eq. (18.11). Note that the system transfer function has been
written as the product of a constant and a frequency-variant term. In Eq. (18.11) the
constant is $\mu R/(r_p + R)$ and the frequency-variant term is $j\omega/(j\omega + 1/T)$.

Time Domain. If the system input has a complex waveform, e.g., a step function,
it is often difficult to predict accurately the output waveform even if the system gain
and phase characteristics are known. The exact output for any complex input wave-
form can be determined, however, if the system equations are written as integro-
differential equations of time. In such an analysis, it is first necessary to write an
expression which equates the system input function $e_i(t)$ to some function of either
the instantaneous load current or, in the absence of a load, a function of the instan-
taneous current through one of the system elements shunting the output terminals.
The second equation which must be written is one which equates the output $e_o(t)$ to
some function of the instantaneous current i appearing in the first equation. The
solution of the first equation for i permits the solution of the second equation for
$e_o(t)$. With reference to Fig. 18.5, the differential equation which relates the input
voltage to the instantaneous current i through the inductance L is Eq. (18.12), and
the equation which relates the current i to $e_o(t)$ is Eq. (18.13).

$$e_i(t) = -\frac{r_p}{\mu}\left[i + L\left(\frac{1}{R} + \frac{1}{r_p}\right)\frac{di}{dt}\right] \tag{18.12}$$

$$e_o(t) = L\frac{di}{dt} \tag{18.13}$$

The expression for the instantaneous current i, obtained from Eq. (18.12), can be
substituted in Eq. (18.13) to determine $e_o(t)$.

Since it is necessary to solve for $e_o(t)$ as described in the preceding paragraph, it is
apparent that the transfer function $e_o(t)/e_i(t)$ does not appear in the solution for

$e_o(t)$. Consequently, the term "transfer function" is not normally associated with systems in which the input and output functions are written as functions of time.

Complex Variable Domain. To realize a simpler and quicker solution for the output of systems having nonsinusoidal inputs, the applicable equations are usually written as functions of the complex variable[1] s. An equation which has been written as a function of time t can be written as a function of s provided s is substituted directly for d/dt and $1/s$ is substituted[2] for $\int dt$. As a result of these substitutions, the equation will no longer be in the time domain and the capital letters E and I must be substituted for e and i. An equation in which the substitutions have been made is said to be written as a function of the complex variable and is represented mathematically as

$$\mathcal{L}[f(t)] = KG(s)$$

and is read, "the Laplace transform of $f(t)$ is equal to $KG(s)$." Having written both the system input function and the system transfer function in the complex-variable domain, the output function $E_o(s)$, also in the complex-variable domain, can be determined by solving Eq. (18.14) by simple algebraic methods.

$$\frac{E_o(s)}{E_i(s)} = KG(s) \tag{18.14}$$

where $KG(s) = \mathcal{L}[f(t)]$

K = invariant term of system transfer function

$E_i(s) = \mathcal{L}[e_i(t)]$

$E_o(s) = \mathcal{L}[e_o(t)]$

$G(s)$ = Laplace transform of variant portion of system transfer function

To illustrate the use of the Laplace transformation, i.e., the conversion of the system equations from the time domain to the complex domain, assume that a 2-volt step function is to be applied to the input of the circuit shown in Fig. 18.5. Item 2 of Table 23.4 shows the Laplace transform of a 2-volt step function to be $2/s$. The system transfer function of the circuit shown in Fig. 18.5 when written in the time domain is given by the ratio of Eq. (18.13) to (18.12). The Laplace transform of this expression is given by Eq. (18.15).

$$\frac{E_o(s)}{E_i(s)} = -K \frac{s}{s + 1/T} \tag{18.15}$$

where $K = \dfrac{\mu R}{r_p + R}$

$T = L\left(\dfrac{1}{R} + \dfrac{1}{r_p}\right)$

The substitution of $2/s$ for $E_i(s)$ and the simple algebraic solution establishes the value for $E_o(s)$.

$$E_o(s) = -2K \frac{1}{s + 1/T} \tag{18.16}$$

The solution for $E_o(s)$ is written as a function of the complex variable and consequently must be converted to the time domain. This process is referred to as taking the inverse transform $\mathcal{L}^{-1}$ and is written

$$\mathcal{L}^{-1}[E_o(s)] = e_o(t) \tag{18.17}$$

---

[1] See Sec. 23.6c.

[2] These substitutions can be made only if the initial conditions are equal to zero as described in Sec. 18.3.

With reference to item 3 of Table 23.4, it can be seen that the inverse transform of Eq. (18.16) is

$$\mathcal{L}^{-1}[E_o(s)] = e_o(t) = -2Ke^{-t/T} \tag{18.18}$$

which is plotted in Fig. 18.6.

18.3b. System Transfer Function for a Feedback Amplifier. The system transfer function for Fig. 18.7 is given by Eq. (18.19).

$$\frac{E_o(s)}{E_i(s)} = \frac{KG(s)}{1 - KG(s)K_1G_1(s)} \tag{18.19}$$

18.4. Methods of Analyzing Stability and Transient Response. Several methods for determining whether or not a system is stable are discussed in Secs. 18.4a to 18.4e. In most of these methods the transient characteristics can be determined with a minimum of additional effort.
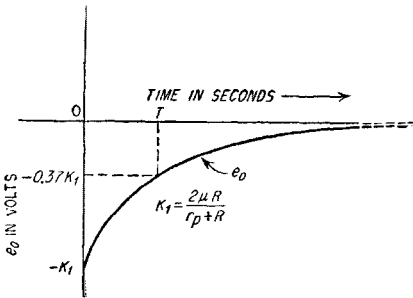
FIG. 18.6. Response of the circuit shown in Fig. 18.5 to a 2-volt step function.
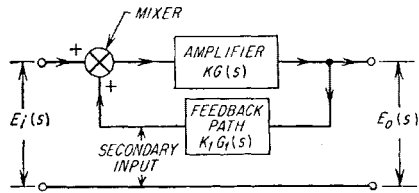
FIG. 18.7. Block diagram of a feedback amplifier.

18.4a. Stability and Transient Analyses by Means of Differential Equations. When both the input function e_i and the system transfer function are written as functions of time t, they are written in integrodifferential-equation form (see Sec. 18.3a). The determination of the system output $e_o(t)$ is consequently dependent on the solution of integrodifferential equations. A stability analysis and the response to various transients can, therefore, be determined by calculating the various output functions for the associated input functions. If the feedback system is to be used as an amplifier, i.e., not as an integrator or oscillator, it is considered to be unstable if the characteristics of the calculated output function for a step-function input are such that the output appears as a steady-state oscillation, a forever increasing oscillation, or some other forever increasing function.

For the reasons stated in Sec. 18.3a, stability and transient analyses are not usually performed using differential equations. Instead, one of the methods discussed in Secs. 18.4b to 18.4e is used.

18.4b. Stability and Transient Analyses Based on Laplace Transforms. The primary advantage in using Laplace transformations for the analysis of systems having non-sinusoidal input functions is the relative simplicity of the mathematical operations as compared to the lengthy solutions which are often associated with integrodifferential equations which have been written as functions of time. The use of Laplace transformations requires that both direct and inverse transforms be taken of certain functions in the process of a particular solution. Many transform pairs have been tabulated and often are of considerable value in simplifying a solution (see Table 23.4). Frequently, however, it may be necessary to perform the mathematical operations for obtaining both direct and inverse transforms. Methods of calculating transforms are given in Sec. 23.6c.

An output function which has been determined by the use of direct and inverse Laplace transforms is written as a function of time and is in the same form as though the solution had been obtained by the solution of integrodifferential equations of

time. The stability and transient analyses of a system can, therefore, be determined by examining the calculated system output function as discussed in Sec. 18.4a.

Satisfactory system analysis can usually be accomplished by a simplified method in which only the Laplace transformation of the system transfer function is examined. This method permits an exact determination as to whether or not the system is stable and in most instances provides a fairly accurate means of determining the system response to a step-function input. The remaining paragraphs in this section contain the necessary background material and the details of such an analysis.

Complex Planes. Two different complex planes are frequently involved in the

FIG. 18.8. Complex-variable plane or *s* plane for plotting values of *s*.

analyses of feedback systems, and it is important that their differences be clearly understood. The complex variable *s* is equal to $\alpha + j\omega$, and the complex plane required to plot values of *s* is known as the complex-variable plane or *s* plane. As an example, three values of *s* have been plotted in Fig. 18.8. The other complex plane

FIG. 18.9. Complex plane for plotting functions of *jω*. For this particular example, the function $KG(j\omega)$ is equal to $j2\omega/(1 + j\omega)$.

most frequently encountered is the one required for the plotting of functions of *jω*. As an example, a transfer function which has been written as a function of *jω* is plotted in Fig. 18.9. Since the complex variable *s* is equal to $\alpha + j\omega$, this is seen to be a special case where $s = j\omega$. The complex plot of $KG(j\omega)$ is therefore a special plot of $KG(s)$ for all values of *s* which fall on the *jω* axis of the complex-variable plane or *s* plane.

Zeros and Poles. If a transfer function has been written as a function of the complex variable s, the *zeros* of the transfer function are those values of the complex variable which cause the transfer function to have a value of zero. It follows that the roots of the numerator of the expression are equal to the zeros. In Eq. (18.20), the single zero is given by $s = -\frac{2}{3}$.

$$\frac{E_o(s)}{E_i(s)} = 21 \frac{s + \frac{2}{3}}{s^2 + 4s + 3} \tag{18.20}$$

The *poles* of the transfer function are those values of the complex variable s which cause the transfer function to have a value of infinity. As an example, assume that the transfer function for a given system is given by Eq. (18.20). The values of s which cause the equation to have a value of infinity are those values which cause the denominator to have a value of zero, and these values are the roots of $s^2 + 4s + 3$, or $s = -1$ and $s = -3$. The poles of Eq. (18.20) are therefore given by $s = -1$ and $s = -3$.

Zeros are usually plotted as dots on the s plane, and the poles are plotted as small crosses.

Stability and Transient Characteristics Based on the s-plane Location of the Poles of the System Transfer Function. The stability of a system can be established simply by determining the poles of the system transfer function and referring to Table 23.7. An examination of Table 23.7 indicates that an amplifier will be stable provided that all the poles of the system transfer function are located in the left half of the s plane. These poles may either lie on the negative real axis, or they may have complex values as shown in Fig. 18.8.

Certain characteristics of the system transient response for various transient input functions can be recognized provided the poles of the system transfer function are known. As has been stated, stability in an amplifier is realized only if all the poles lie in the left half of the s plane. The exact positions of the poles in the left half of the s plane, however, determine the transient response of the system. As an example, if the input to a system is a step function, a pole on the negative real axis will cause the system to introduce attenuation in the form of a decaying exponential (see the overdamped and critically damped curves shown in Fig. 18.4). A pair of complex-conjugate poles in the left half plane will introduce a frequency component having an amplitude which decreases exponentially (see the underdamped curve shown in Fig. 18.4). The frequency of the decaying oscillation is determined by the vertical displacement of the complex-conjugate poles from the real axis on the s plane. For the poles on the negative real axis and the complex-conjugate poles, the time constants of the decaying exponentials are determined by the reciprocals of the horizontal displacements of the poles from the imaginary axis on the s plane. A system which introduces damped-frequency components to a step-function input is said to be underdamped, and the system transfer function will, therefore, have at least one pair of complex-conjugate poles in the left-hand plane. By increasing the damping in the system, the complex-conjugate poles can be made to fall on the negative real axis of the s plane and consequently are then no longer complex. Under this condition, the system will introduce no damped-frequency components to a step-function input. Instead, the step-function input will be attenuated by two decaying exponentials.

If all the system-transfer-function poles fall on the negative real axis, the system is referred to as being either overdamped or critically damped. A system which can be made to be either under- or overdamped by changing the loading within the system is said to be critically damped when the loading has been increased only to the extent that the complex-conjugate poles disappear and appear as a second-order pole on the negative real axis of the s plane.

To illustrate several terms which are used to describe certain system character-

istics, consider the network shown in Fig. 18.10. The transfer function of the network[1] is given by Eq. (18.21).

$$\frac{E_o(s)}{E_i(s)} = \frac{1}{LC}\frac{1}{s^2 + Rs/L + 1/LC} \quad (18.21)$$
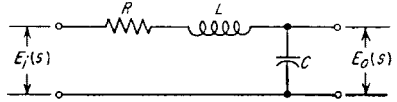
FIG. 18.10. Simple quadratic system.

The network is referred to as a quadratic or a second-order system since the system transfer function has two poles. Equation (18.21) is frequently written in the form given by Eq. (18.22) or by (18.23).

$$\frac{E_o(s)}{E_i(s)} = \omega_n^2 \frac{1}{s^2 + 2\zeta\omega_n s + \omega_n^2} \quad (18.22)$$

$$\frac{E_o(s)}{E_i(s)} = \omega_n^2 \frac{1}{s^2 + 2\alpha s + \omega_n^2} \quad (18.23)$$

where $\omega_n = 1/\sqrt{LC}$ = undamped natural frequency, i.e., frequency in radians per second at which system would oscillate if R were equal to 0

$\alpha = R/2L$ = attenuation factor

$\zeta = R\sqrt{C/L}/2 = \alpha/\omega_n$ = damping ratio

The poles s_1 and s_2 of the transfer function are given by the roots of the denominator.

$$s_1 = -\alpha + j\sqrt{\omega_n^2 - \alpha^2} = -\alpha + j\omega_o$$
$$s_2 = -\alpha - j\sqrt{\omega_n^2 - \alpha^2} = -\alpha - j\omega_o$$

where $\omega_o = \sqrt{\omega_n^2 - \alpha^2}$ = damped resonant frequency, i.e., frequency of oscillation in radians per second of response overshoots

The locus of the values of the complex-conjugate poles, s_1 and s_2, for all the values of R which satisfy the relationship $\zeta < 1$, is a semicircle on the s plane as shown in
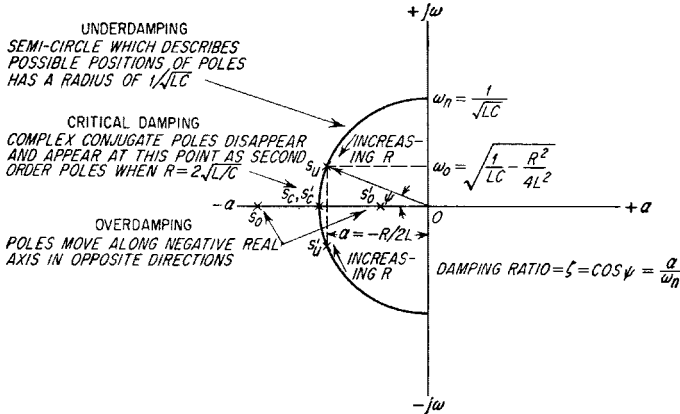
FIG. 18.11. Locus of the transfer-function poles for the network shown in Fig. 18.10 if L and C are both constant and R is varied.

Fig. 18.11. Since the complex-conjugate poles indicate that the system is underdamped, the pair of conjugate poles which have been plotted in Fig. 18.11 are designated as s_u and s_u'. Similarly, at critical damping the poles are designated as s_c

---

[1] It is of no consequence in this analysis whether the system is a feedback system or a simple network since the response of a system is dependent only on the over-all system transfer function. For this reason, the transfer function of a simple network can be used to illustrate how the transfer-function poles affect the transient response.

and s'_c, and with overdamping s_o and s'_o.   Critical damping is obtained when $\zeta = 1$, and overdamping is obtained when $\zeta > 1$.
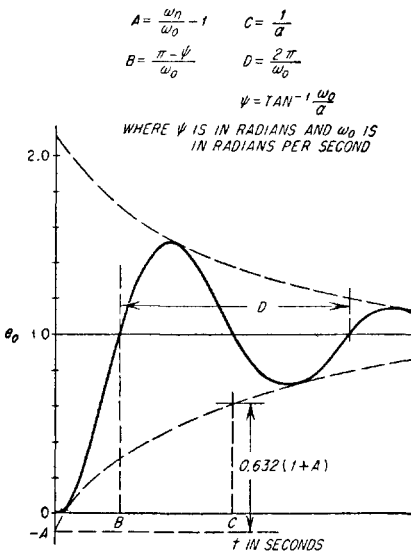
FIG. 18.12. General characteristics of the output function of an underdamped network of the type shown in Fig. 18.10 when the input is a step function.

FIG. 18.13. Dimensionless response functions of the circuit shown in Fig. 18.10 for a unit-step-function input and various damping ratios.

If the input function to the circuit shown in Fig. 18.10 is a step function and if the transfer-function poles are on the semicircle, the characteristics of the output response can be determined from the curve shown in Fig. 18.12.   It should be noted that the frequency in radians per second of the damped oscillations is given by ω_o and the rate at which they decay is determined by α.   A family of dimensionless response functions resulting from a step function input to the circuit shown in Fig. 18.10 is shown in Fig. 18.13.

Damping ratios less than unity are normally used in servomechanisms to reduce the rise time of the response function.   Although the overshoots associated with the reduced rise times are usually considered to be undesirable, an overshoot of 10 or 15 per cent is allowable in most systems.

FIG. 18.14. Circuit to illustrate how damping ratio affects damped oscillations in output response to a step-function input.

All quadratic systems which have complex-conjugate poles as shown in Fig. 18.11 will not necessarily have the same system response to the same input function.   As an example, consider the network shown in Fig. 18.14 which has the transfer function given by Eq. (18.24).

$$\frac{E_o(s)}{E_i(s)} = -\frac{g_m}{C}\frac{s}{s^2 + s/RC + 1/LC} \qquad (18.24)$$

where $\omega_n = 1/\sqrt{LC}$

$$\alpha = \frac{1}{2RC}$$

$$\zeta = \frac{1}{2R}\sqrt{\frac{L}{C}}$$

$$\omega_o = \sqrt{\frac{1}{LC} - \frac{1}{4R^2C^2}}$$

The locus of the poles of the system transfer function describes a semicircle as do the poles for the network shown in Fig. 18.10. The relationship between the positions of the poles in the s plane and the associated system damping characteristics, i.e., underdamping, critical damping, etc., are therefore the same for either the network shown in Fig. 18.10 or the network shown in Fig. 18.14. However, the networks will have different output responses to step-function inputs since the circuit shown in Fig. 18.14 has a zero at $s = 0$, consequently, only the circuit shown in Fig. 18.10 has a d-c term. Critical damping for either circuit is obtained when $\zeta = 1$ and the frequencies ω_o and ω_n are equal to the damped and undamped frequencies, respectively.

If the system transfer function contains several poles, the associated amplitude terms of the inverse Laplace transform will normally be greater for those poles which are closest to the $j\omega$ axis on the s plane. An exception can occur when the system transfer function has several poles which are relatively close together and one or two poles which are some distance away on the s plane. The amplitudes of the terms associated with the isolated poles may be relatively small compared to the terms associated with the poles falling in the main group even though the isolated poles are much closer to the $j\omega$ axis.

18.4c. Steady-state Stability Based on the Nyquist Criterion. A feedback system will not be stable if the system transfer function has any poles in the right half of the s plane. This is the same as stating that a feedback system will be unstable if the zeros of the denominator of the system transfer function lie in the right half of the s plane. In the case of a feedback amplifier (see Fig. 18.7), the system transfer function s written as shown in Eq. (18.25).

$$\frac{E_o(s)}{E_i(s)} = K_fG_f(s) = \frac{KG(s)}{1 - KG(s)K_1G_1(s)} \tag{18.25}$$

where $K_fG_f(s)$ = system transfer function with feedback

$\qquad KG(s)$ = system transfer function without feedback

$\quad KG(s)K_1G_1(s)$ = feedback transfer function

Consequently, if there are no zeros of $1 - KG(s)K_1G_1(s)$ in the right half of the s plane, the system will be stable. Often the form of the denominator of Eq. (18.25) is of such complex nature that it does not yield easily to analysis. The Nyquist technique is useful under these circumstances.

General Case. Nyquist has shown that with the aid of a graphical plot and the knowledge of the number of poles of $KG(s)K_1G_1(s)$ in the right half of the s plane, it is possible to establish the number of zeros of $1 - KG(s)K_1G_1(s)$ in the right half of the s plane. The first step in this type of stability analysis is the substitution of $j\omega$ for s in the expression $KG(s)K_1G_1(s)$ so as to obtain $KG(j\omega)K_1G_1(j\omega)$. The quantity $-KG(j\omega)K_1G_1(j\omega)$ is then plotted on the complex plane for all values of ω from $-\infty$ through 0 to $+\infty$ and is referred to as the *transfer locus plot.* This plot is examined to determine the number of counterclockwise encirclements N_{cc} of the point $-1 + j0$ as ω is varied from $-\infty$ through 0 to $+\infty$. The number of zeros of $1 - KG(s)K_1G_1(s)$ in the right half plane can then be found from Eq. (18.26).

$$Z = P - N_{cc} \tag{18.26}$$

where Z = number of zeros of $1 - KG(s)K_1G_1(s)$ in right half of s plane

P = number of poles of $KG(s)K_1G_1(s)$ in right half of s plane

N_{cc} = net number of counterclockwise encirclements of point $-1 + j0$ on complex plane by transfer locus plot as ω is varied from $-\infty$ through 0 to $+\infty$.

As has been stated, the feedback system will be unstable if there are any zeros of $1 - KG(s)K_1G_1(s)$ in the right half of the s plane. To illustrate a typical analysis, let the system transfer function of a sample feedback system be given by Eq. (18.27).

$$K_fG_f(s) = \frac{10/(s + 1)}{1 - \left[\dfrac{-10}{(s + 1)(s + 5)}\right]} \qquad (18.27)$$

where $KG(s) = 10/(s + 1)$

$K_1G_1(s) = -1/(s + 5)$

The term $KG(s)K_1G_1(s)$ is therefore equal to $-10/(s^2 + 6s + 5)$, and $KG(j\omega)K_1G_1(j\omega)$ is equal to $-10/(5 - \omega^2 + j6\omega)$. The plot of $-KG(j\omega)K_1G_1(j\omega)$ on the complex plane for all values of ω is shown in Fig. 18.15 and makes no encirclement of the critical point $-1 + j0$ as ω is varied from $-\infty$ through 0 to $+\infty$. The value of N_{cc} is therefore equal to 0. The number of poles P of $KG(s)K_1G_1(s)$ in the right half of the s plane is equal to zero since the roots of the denominator of $KG(s)K_1G_1(s)$ are equal to -1 and -5. From Eq. (18.26) the number of zeros of $1 - KG(s)K_1G_1(s)$ in the right half plane can then be established.
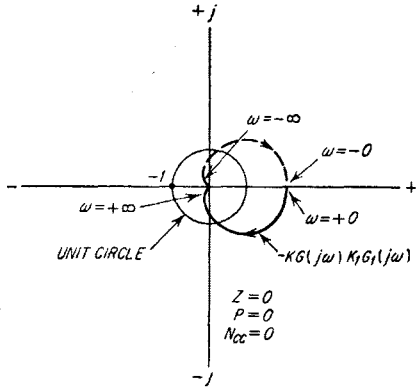
$$\begin{aligned}Z &= P - N_{cc} \\ &= 0 - 0 \\ &= 0\end{aligned}$$

FIG. 18.15. Transfer locus plot for the system transfer function defined by Eq. (18.27).

Therefore, the system is stable since $1 - KG(s)K_1G_1(s)$ has no zeros in the right half of the s plane. This is equivalent to stating that Eq. (18.27) has no poles in the right half of the s plane.

As a second example, consider a system in which the open-loop transfer function is given by Eq. (18.28).

$$KG(s)K_1G_1(s) = -7,000\frac{(s + 1)(s + 1.5)}{s^2(s + 0.1)(s + 20)(s + 30)} \qquad (18.28)$$

The plot of $-KG(j\omega)K_1G_1(j\omega)$ is shown in Fig. 18.16 and does not encircle the point $-1 + j0$. Since both P and N_{cc} are equal to zero, Z is also equal to zero and the system is stable.

With reference to Fig. 18.16, it should be noted that the transfer locus plot would not be a continuous curve if the points associated with $\omega = +0$ and $\omega = -0$ were not connected. It is therefore necessary to substitute in the basic equation, i.e., Eq. (18.28), infinitely small values of s as defined by a curve which detours about the origin into the right half of the s plane as shown in Fig. 18.17 to permit the completion of the transfer locus plot. If a discontinuity had been obtained at any other value of ω, it would have been necessary to make a similarly small detour into the right half of the s plane about that particular point. Open-loop transfer functions are

sometimes obtained which have transfer locus plots that are discontinuous at $\omega = +\infty$ and $\omega = -\infty$. In these cases it is necessary to substitute in the open-loop transfer
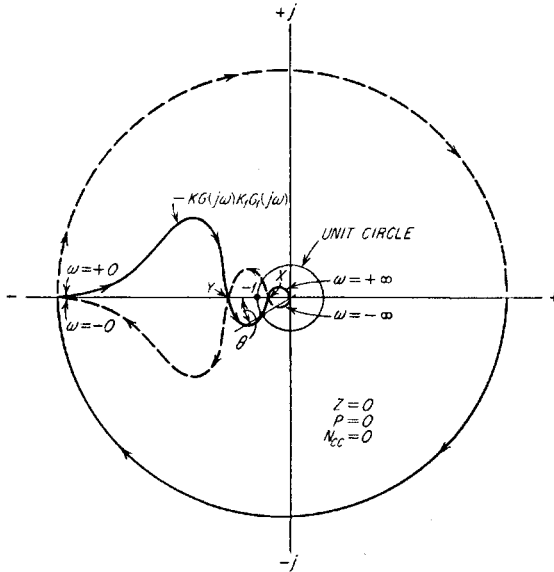
FIG. 18.16. Transfer locus plot for the open loop defined by Eq. (18.28).

functions the values of s which are defined by the curve shown in Fig. 18.17 to complete the plot.

Simplified Case. If a system has a stable open loop, $KG(s)K_1G_1(s)$ is known to have no poles in the right half of the s plane. In this case the system is known to be stable if the plot of $-KG(j\omega)K_1G_1(j\omega)$ does not encircle the point $-1 + j0$.

18.4d. Stability and Transient Analysis Based on the Amplitude and Phase versus Frequency Plots of the Open- and Closed-loop Transfer Functions. A feedback system's margin of steady-state stability and its transient response to transient input functions can be established by a simple method which involves the analysis of both the amplitude and phase characteristics of the open-loop system and the amplitude characteristics of the closed-loop system.

Open-loop Amplitude and Phase versus Frequency. In a negative feedback amplifier the signal inversion required to obtain negative feedback is the result of amplifying the signal through an odd number of signal-inverting stages. Although the

FIG. 18.17. Locus of values for s required to complete transfer locus plots when discontinuities are obtained at $\omega = 0$ and $\omega = \infty$.

inversion of a continuous sine wave can be considered as being identical to phase-shifting the sine wave by $\pm 180°$, the term phase shift as applied to an amplifier usually does not include the phase angle associated with phase inversions. Conse-

quently, in a negative feedback amplifier the phase shift around the feedback loop due to frequency variant elements will be considered to be equal to 0° at midband, even though the necessary signal inversion in the amplifier and feedback path for negative feedback will be equivalent to a phase shift of ±180°.

If the open loop of a feedback system is stable, it is possible to establish certain criteria which relate the characteristics of the open loop and the stability of the closed loop. For example, the closed loop will be stable if the phase shift of a signal through the open loop, disregarding the effective phase shift due to signal inversions in the different stages, is less than ±180° at the frequency at which the open-loop gain is equal to unity. Determination of the amplitude and phase characteristics of the stable open loop therefore establishes whether or not the closed-loop system will be stable. If the closed loop will not be stable, it is then possible to specify the characteristics of the equalizing network which must be placed in series with the open loop to obtain stability in the closed loop. If the forward path $KG(s)$ and the feedback path $K_1G_1(s)$ of a feedback system as shown in Fig. 18.18 contain only minimum-phase-shift networks[1] and if the open loop is stable, the over-all open-loop transfer function $KG(s)K_1G_1(s)$ will have no zeros or poles in the right half of the s plane. Since a known relationship exists between the amplitude and phase characteristics[2] of systems having no zeros or poles in the right half of the s plane, it is possible to calculate the phase

FIG. 18.18. Feedback system with the feedback loop shown open.

characteristics of this type of system provided the open-loop amplitude characteristics are known. If it is not known whether or not the stable open-loop system satisfies the criterion of having no zeros in the right half of the s plane, it is necessary to measure both the phase and amplitude characteristics. Examples of systems which have stable open loops and which utilize only minimum phase-shift networks are given by Eqs. (18.27) and (18.28).

The terms which are used to indicate the degree of steady-state stability of a negative feedback system are *gain margin* and *phase margin*. The term gain margin is normally defined as the decibel differential between the open-loop gain of 0 db and the open-loop gain at the frequency or frequencies at which the open-loop phase shift, because of frequency variant elements, becomes equal to ±180°. Phase margin is the phase difference between 180° and the phase shift in degrees at the frequency where the open-loop gain is unity. With reference to Fig. 18.16, the gain margins are given by the decibel differences in gain between the points X and -1 and between the points -1 and Y. The phase margin is given by the angle θ. In general, good stability in a feedback system requires that the gain margin be equal to approximately 10 db or more and that the phase margin be equal to approximately 40° or more. Shown in Fig. 18.19 are the open-loop amplitude and phase characteristics of a negative feedback amplifier. The lower and upper gain margins are 19.5 and 14 db, respectively, and both the phase margins are 36°.

If either the amplitude or phase characteristics of the open loop do not satisfy the conditions for steady-state stability of the closed loop, it is necessary to increase[3] or decrease the open-loop gain or to introduce an equalizing network somewhere in the

[1] Minimum-phase-shift networks have no zeros or poles in the right half of the s plane (see Sec. 18.5a).

[2] See Sec. 18.5b.

[3] In most cases a system can be stabilized by reducing the open-loop gain; however, the stability of some systems is dependent on a critical value of gain. An example of the second type is given by the transfer locus plot shown in Fig. 18.16.

open-loop path so that the closed-loop system will be stable.   In general, most networks are minimum-phase-shift networks, and their use as equalizing networks permits the determination of the phase shift associated with any given amplitude characteristic.   This permits the calculation of the modified open-loop amplitude and phase characteristics which will be obtained as a result of the addition of a minimum-phase-shift equalizing network to a feedback system in which the amplitude and phase characteristics are known.   The amplitude and phase characteristics of several minimum-phase-shift networks are given in Table 1.4.

Closed-loop Amplitude versus Frequency Response.   In Fig. 18.20 is a family of amplitude versus frequency curves for the simple circuit shown.   These curves illustrate how the damping ratio ζ affects the amplitude response.   The frequency of maximum amplitude peaking is approximately equal to the damped resonant frequency ω_o, and the degree to which the amplitude peaking introduces transient frequency components in the output, when the input is a step function, can be determined by referring to Fig. 18.13.

FIG. 18.19. Open-loop amplitude and phase plots of a negative feedback amplifier.
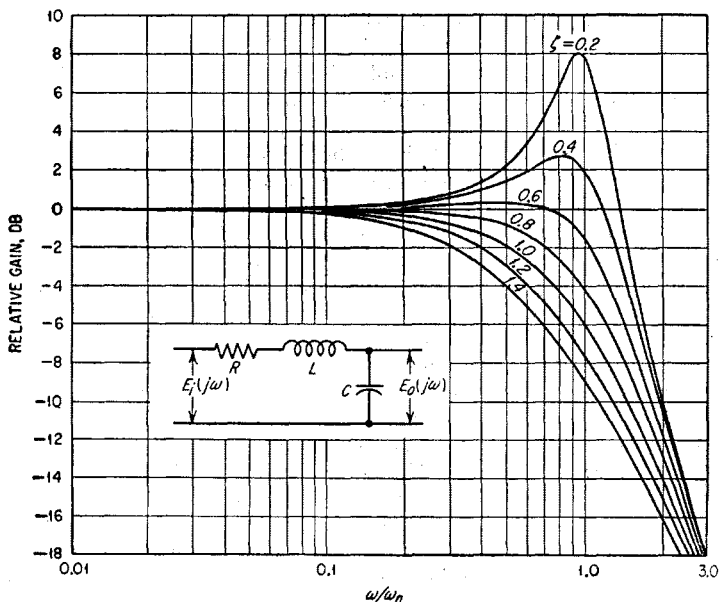
FIG. 18.20. Amplitude versus frequency curves for the circuit shown.

Although the curves shown in Fig. 18.20 are for a quadratic system, the damping ratio ζ of a higher order system can be approximated by comparing its closed-loop amplitude versus frequency-response curve with those shown in the figure.

18.4e. Locus of Roots Method for Graphically Establishing the Poles of the System Transfer Function. Equation (18.25) is the closed-loop system transfer function of the negative feedback amplifier shown in Fig. 18.7. The *locus of roots* method provides a convenient means of graphically determining the zeros of $1 - KG(s)K_1G_1(s)$, i.e., the poles of Eq. (18.25), on the s plane, and consequently establishing whether or not the system is stable.[1] Since the zeros of $1 - KG(s)K_1G_1(s)$ are the poles of Eq. (18.25), the system damping ratio ζ can be determined by locating the poles on the s plane and taking the cosine of the angle ψ as defined in Fig. 18.11. The locus of roots method is of particular value since the s-plane plot of the loci of the zeros of the system-transfer-function denominator clearly defines all possible values of the system-transfer-function poles as the frequency invariant product KK_1 is varied.

To illustrate an analysis by the locus of roots method, assume that the denominator of Eq. (18.25) is given by Eq. (18.29).

$$1 - KG(s)K_1G_1(s) = 1 + 20\frac{s}{s + \frac{1}{6}} \times 5\frac{s + \frac{1}{20}}{s + \frac{1}{3}} \qquad (18.29)$$

It follows that the zeros of $1 - KG(s)K_1G_1(s)$ are those values of s which satisfy Eq. (18.30).

$$\frac{s(s + \frac{1}{20})}{(s + \frac{1}{6})(s + \frac{1}{3})} = \frac{1}{KK_1} = -\frac{1}{100} \qquad (18.30)$$

The first step consists of locating on the s plane the poles and zeros of the left-hand portion of Eq. (18.30). The zeros are at $s = 0 + j0$ and $s = -0.05 + j0$, and the poles are at $s = -0.17 + j0$ and $s = -0.33 + j0$. These are shown in Fig. 18.21.
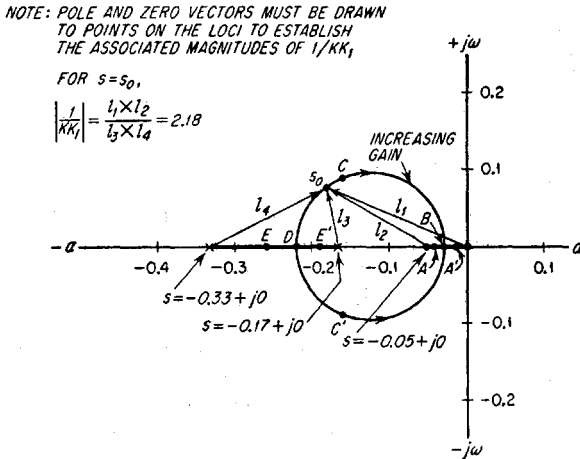
FIG. 18.21. Loci of the system-transfer-function poles by means of the locus of roots method.

The second step consists of experimentally determining and plotting all the possible values of s which, when substituted in Eq. (18.30), will cause the left-hand portion of the equation to have the same sign as the right-hand portion. The left-hand portion of Eq. (18.30) must therefore have a negative sign which is equivalent to a phase angle of $\pm 180°$. As an example of the procedure in experimentally determining these values of s, assume that s has a value of $-0.02 + j0$. If, on Fig. 18.21, vectors were drawn from the two zeros and the two poles to the assumed value of s, the phase

[1] Poles of the system transfer function in the right half of the s plane indicate that the system is unstable.

angle associated with the left-hand portion of Eq. (18.30) would be found to be 180°. This can be explained by the fact that the phase angle of each vector is measured from a horizontal line extending to the right of the particular zero or pole. The phase angle of the over-all expression is obtained by adding the sum of the phase angles measured from the zeros and subtracting the sum of all the phase angles measured from the poles. In this example, the sum of the phase angles associated with the vectors drawn from the two zeros to the point $-0.02 + j0$ is equal to $0 + 180$, or 180°. The sum of the phase angles associated with the two vectors drawn from the two poles to the point $-0.02 + j0$ is equal to $0 + 0$, or 0°. The sum of the phase angles associated with the zero vectors less the sum of the phase angles associated with the pole vectors is $180 - 0$, or 180°.

The loci of all values of s which cause the left-hand portion of Eq. (18.30) to have a negative value, i.e., a phase angle of $\pm 180°$, are shown in Fig. 18.21. An examination of the entire s plane will reveal that there is no other point or locus of points which satisfies the 180° phase angle criterion. Therefore the values of s which cause the left-hand portion of Eq. (18.30) to equal the right-hand portion must fall somewhere on the established loci since the right-hand portion of the expression has a negative sign. If the sign of either K or K_1 had been different so as to cause the sign of the right-hand portion of Eq. (18.30) to be positive, it would have been necessary to determine the loci of points which would cause the left-hand quantity to have a phase angle of 0°.

Since the loci plotted in Fig. 18.21 define all values of s which will cause the left-hand portion of Eq. (18.30) to have a negative value, the final step is that of graphically determining the particular values of s on the loci which satisfy the amplitude criterion dictated by the actual system value of $1/KK_1$. The process consists of assuming values of s on the loci and testing for the satisfaction of Eq. (18.30). A graphical procedure for testing an assumed value of s for satisfaction of Eq. (18.30) is that of drawing vectors on the s plane from both the two zeros and the two poles of $KG(s)K_1G_1(s)$ to the assumed value of s on the loci and dividing the product of the two zero vector lengths by the product of the two pole vector lengths. If the result is equal to the magnitude of $1/KK_1$, the assumed value of s is known to be one of the system-transfer-function poles.[1] The number of system-transfer-function poles to be determined is always equal to the number of poles of $KG(s)K_1G_1(s)$. Therefore, different values of s on the loci must be tested until the known number of system-transfer-function poles are determined. Since the particular function $KG(s)K_1G_1(s)$ given in Eq. (18.29) has two poles, the number of associated system-transfer-function poles is also equal to 2. Since $1/KK_1$ is equal to 0.01, the values of the two system-transfer-function poles A and A' which satisfy this criterion are $-0.0422 + j0$ and $-0.0117 + j0$.

With reference to Fig. 18.21, note that as the system gain is decreased; i.e., as the magnitude of $1/KK_1$ is increased, the system poles move from A and A' toward one another along the negative real axis and for one specific value of $1/KK_1$ appear as a second-order pole at point B. As the magnitude of $1/KK_1$ is further increased, the system-transfer-function poles become complex-conjugate poles as shown at points C and C' and then appear as a second-order pole at point D and finally separate again along the negative real axis as shown by points E and E'. The positions of the system-transfer-function poles for different values of the feedback-transfer-function gain constant KK_1 indicate that the system is overdamped for both small and large values of KK_1 and is underdamped for intermediate values of KK_1.

By the procedure outlined it is possible to establish the values of KK_1 associated with all the values of s on the loci. In this manner the system transient and steady-

---

[1] The result obtained by dividing the product of the zero vector lengths by the product of the pole vector lengths is equal to the magnitude of the left-hand portion of Eq. (18.30) for the assumed value of s.

state stability characteristics as functions of the product of the gain constants K and K_1 are clearly indicated. For example, the loci associated with certain feedback transfer functions may extend into the right half of the s plane. It is obvious that the system will be unstable for those values of KK_1 which cause the poles to lie in the right half of the s plane. The location of poles in the left half of the s plane will affect the transient response as discussed in Sec. 18.4b.

The locus of roots method is frequently used to solve for the roots of higher-order equations. As an example, consider Eq. (18.31).

$$As^4 + Bs^3 + Cs^2 + Ds + E = 0 \qquad (18.31)$$

If Eq. (18.31) were solved by the locus of roots method, it would be necessary to first write the following equations which result from the breakdown of Eq. (18.31).

$$s^4 + \frac{B}{A} s^3 + \frac{C}{A} s^2 + \frac{D}{A} s + \frac{E}{A} = 0 \qquad (18.32)$$

$$s \left(s^3 + \frac{B}{A} s^2 + \frac{C}{A} s + \frac{D}{A} \right) = -\frac{E}{A} \qquad (18.33)$$

$$s \left(s^2 + \frac{B}{A} s + \frac{C}{A} \right) = -\frac{D}{A} \qquad (18.34)$$

$$s \left(s + \frac{B}{A} \right) = -\frac{C}{A} \qquad (18.35)$$

Step 1 consists of determining the values of s which satisfy Eq. (18.35). The zeros of the left-hand portion of Eq. (18.35), that is, $0 + j0$ and $-B/A + j0$, are first plotted on the s plane. The locus representing all possible values of s which will cause the left-hand portion of Eq. (18.35) to have a sign like that of $-C/A$ is then plotted. The locus is next tested for values of s which will cause the product of the two zero vectors to the assumed value of s to have a magnitude equal to C/A. Careful examination of the locus will show that there are two values of s which satisfy this criterion.

Step 2 consists of solving for the values of s which satisfy Eq. (18.34). The zeros of the left-hand portion of Eq. (18.34) are next plotted on a new s-plane plot and are equal to $0 + j0$ and the two roots of Eq. (18.35) determined in step 1. In a manner similar to that taken in step 1, the loci representing all possible values of s which will cause the left-hand portion of Eq. (18.34) to have the same sign as $-D/A$ are then constructed on the s plane. All values of s on the loci are then tested for satisfaction of the amplitude criterion which states that the product of the zero vector lengths must be equal to the magnitude of D/A. Careful examination of the loci will disclose that Eq. (18.34) has three roots.

Step 3 consists of solving for the values of s which satisfy Eq. (18.33) The zeros of the left-hand portion of Eq. (18.33) must first be plotted on a new s-plane plot and are equal to $0 + j0$ and the three roots of Eq. (18.34) determined in step 2. As in the preceding steps, the loci representing all values of s which will cause the left-hand portion of Eq. (18.33) to have the same sign as the right-hand portion, that is, $-E/A$, are plotted on the s plane. Once more, all the values of s on the loci are tested to see if they satisfy the amplitude criterion; i.e., the product of the zero vector lengths to the assumed value of s must be equal to the magnitude of $-E/A$. There are four values of s which satisfy this criterion, and they are the roots of Eq. (18.31).

18.5. Minimum-phase-shift Equalizing Networks. To stabilize a feedback amplifier it is frequently necessary to add an equalizing network in either the amplifier or the feedback path. If the open-loop amplitude and phase characteristics of the unstable system are known, it is possible to establish the necessary modifications in either the amplitude or phase characteristics to stabilize the system (see Sec. 18.4d).

If the network to be added is a minimum-phase-shift network, the phase characteristics associated with any prescribed amplitude characteristic can readily be established without any knowledge of the network itself. For this reason, minimum-phase-shift networks are usually used to stablize feedback amplifiers.

18.5a. Definition of a Minimum-phase-shift Network. A minimum-phase-shift network has the minimum possible phase shift for any specified amplitude characteristic. Examples of two theoretical passive networks, each having a single pole and a single zero, which have identical amplitude characteristics as a function of frequency are given by the two transfer functions β_1 and β_2.

$$\beta_1 = K\frac{s + 1/T_1}{s + 1/T_2} \qquad (18.36)$$

$$\beta_2 = K\frac{s - 1/T_1}{s + 1/T_2} \qquad (18.37)$$

At any given frequency ω, the amplitude of β_1 is equal to the amplitude of β_2 and is given by Eq. (18.38).

$$\beta_1 = \beta_2 = K\frac{\sqrt{\omega^2 + (1/T_1)^2}}{\sqrt{\omega^2 + (1/T_2)^2}} \qquad (18.38)$$

The phase angle associated with β_1 is θ_1, and the phase angle for β_2 is θ_2.

$$\theta_1 = \tan^{-1}\omega T_1 - \tan^{-1}\omega T_2 \qquad (18.39)$$
$$\theta_2 = -\tan^{-1}\omega T_1 - \tan^{-1}\omega T_2 \qquad (18.40)$$

θ_1 will always be less in magnitude than θ_2. The minimum-phase-shift network is consequently the one defined by Eq. (18.36). It would appear that the same amplitude and magnitude of phase shift as a function of frequency would be given by a transfer function as defined by Eq. (18.41).

$$\beta_3 = K\frac{s - 1/T_1}{s - 1/T_2} \qquad (18.41)$$

This particular transfer function cannot be obtained with a passive network since it has a pole in the right half of the s plane and is unstable. In summation, passive networks may have zeros in either the right half of the s plane [nonminimum-phase-shift networks, e.g., Eq. (18.37)] or the left half of the s plane [minimum-phase-shift networks, e.g., Eq. (18.36)], but they will never have poles in the right half of the s plane.

Any ladder-type network is a minimum-phase-shift network. Those networks which in some instances may be nonminimum-phase-shift networks are those in which there are multiple signal paths between the input and output of the network. Examples of networks which may be nonminimum-phase-shift networks are lattice networks (all-pass filters), bridged networks, and those having distributed constants.

18.5b. Relationship between Amplitude and Phase in Minimum-phase-shift Networks. The phase shift introduced by a network at any given frequency is a function of the slope of the amplitude response curve in decibels per octave, the number of octaves over which the particular slope exists, and the attenuation slopes at other frequencies. Minimum-phase-shift networks having constant attenuation slopes of ± 6 db per octave for an infinite number of octaves will have a phase shift of $\pm 90°$, respectively. Networks whose attenuation slopes approach ± 6 db per octave will have a phase shift approaching $\pm 90°$, respectively, provided the slope exists for many octaves and provided the phase shift is not measured at or near the frequencies at which the attenuation slope changes. Under the same limiting conditions, a slope of ± 12 db per octave

is associated with a phase shift of $\pm 180°$, ± 18 db per octave with a phase shift of $\pm 270°$, etc.

It is possible to approximate the amplitude versus frequency response of a minimum-phase-shift network by a series of connected straight lines. Amplitude plots of this type are often referred to as *Bode plots*. As an example, consider the transfer function of the network given by Eq. (18.42).

$$\frac{E_o(s)}{E_i(s)} = K \frac{(s + A)}{(s + B)(s + C)} \tag{18.42}$$

Each factor of Eq. (18.42), that is, $s + A$, $s + B$, and $s + C$, can be represented by a straight line with a slope of 6 db per octave. The straight lines start at the frequencies[1] $\omega = A$, $\omega = B$, and $\omega = C$. If the factor is in the numerator, the slope is $+6$ db per octave, and if the factor is in the denominator, the slope is -6 db per octave. Assuming that $B < A < C$, the approximation of the amplitude versus frequency response would be constructed as shown in Fig. 18.22. The value of the transfer function at zero frequency is obtained by letting $s = 0$ in Eq. (18.42). Since B is less than either A or C, the first factor to affect the amplitude response is the factor $s + B$ which introduces, at $\omega = B$, a slope of -6 db per octave. This slope continues to the frequency at which $\omega = A$, and at this frequency a slope of $+6$ db per octave is added to the existing slope of -6 db per octave. The result is a slope of 0 db per octave. At the frequency $\omega = C$, a new slope of -6 db per octave is introduced and continues with a permanent slope since there are no additional factors in the expression. The straight-line synthesis can be compared with the actual response as shown in Fig. 18.22 where $K = 1$, $A = 5$, $B = 1$, $C = 25$ radians/sec.
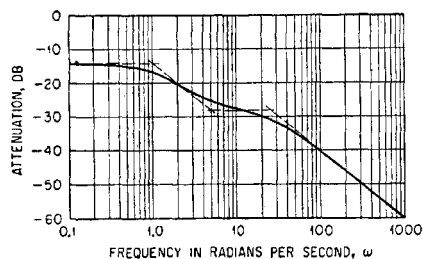
Fig. 18.22. Attenuation characteristics of the network defined by Eq. (18.42).
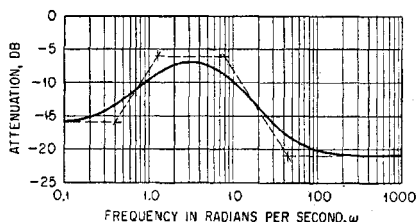
Fig. 18.23. Attenuation versus frequency characteristics of a sample minimum-phase-shift network.

The approximate phase characteristics of a minimum-phase-shift network can be established if the amplitude versus frequency characteristics are known. The procedure consists of first approximating the amplitude versus frequency response with a series of connected straight lines over a frequency range which extends from frequencies that are at least 10 times lower and 10 times higher than those frequencies for which the phase shift is desired. The slopes of the straight lines must always be equal to 0 db per octave or some multiple of ± 6 db per octave. The next step consists of writing the equation for the established curve, and the final step is that of solving the equation for the associated phase shift.

To illustrate a typical solution, assume that it is necessary to determine the phase shift associated with the minimum-phase-shift network having the amplitude versus frequency response shown in Fig. 18.23. The broken line in Fig. 18.23 represents an attempt to approximate the actual amplitude response. It should be noted that the constructed curve is composed of a series of straight lines each of which has a slope

[1] In the case of amplitude versus frequency analysis, the complex variable s is equal to $j\omega$.

equal to 0 db per octave or some multiple of ± 6 db per octave. The break points along the curve and the attenuation at some specific frequency permit the establishment of the value of K. Since the curve in Fig. 18.23 has four break points, the equation has four factors which are given by Eq. (18.43).

$$\frac{E_o(s)}{E_i(s)} = K \frac{(s + 0.4)(s + 46)}{(s + 1.3)(s + 8)} \tag{18.43}$$

At zero frequency the network introduces an attenuation of 16 db, consequently the value of K is 0.0895 obtained by letting $s = 0$ in Eq. (18.43) and solving for the value of K necessary to cause the equation to introduce an attenuation of 16 db. The value of K has been determined only for the purpose of obtaining the complete transfer
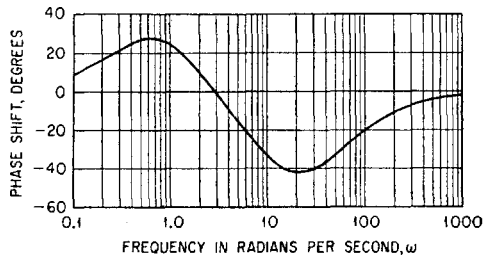
FIG. 18.24. Calculated phase characteristics for the attenuation curve shown in Fig. 18.23.

function. It is not necessary to establish the value of K in the determination of the phase shift through the network. If $j\omega$ is substituted for s, Eq. (18.43) can be rewritten as shown in Eq. (18.44).

$$\frac{E_o(j\omega)}{E_i(j\omega)} = K \frac{18.4 - \omega^2 + j46.4\omega}{10.4 - \omega^2 + j9.3\omega} \tag{18.44}$$

The approximate phase shift θ associated with the curve shown in Fig. 18.23 is therefore given by Eq. (18.45) and is shown in Fig. 18.24.

$$\theta = \tan^{-1} 46.4\omega/(18.4 - \omega^2) - \tan^{-1} 9.3\omega/(10.4 - \omega^2) \tag{18.45}$$

The transfer functions given in Table 1.4 are all minimum-phase-shift networks, and consequently a knowledge of their amplitude characteristics defines their phase
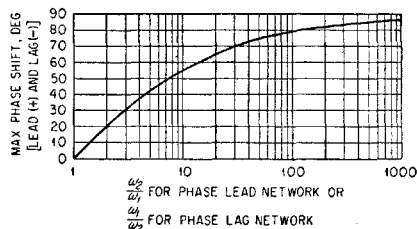
FIG. 18.25. Maximum phase shift for a simple phase lead or lag network having two break frequencies ω_1 and ω_2 as typified in circuits #4 and #12 of Table 1.4. The frequency at which the maximum phase shift occurs is equal to $\sqrt{\omega_1\omega_2}$.

characteristics. Figure 18.25 can be used to determine the maximum phase shift which can be obtained with a simple phase lead or lag network that has two break frequencies ω_1 and ω_2. The frequency at which the maximum phase shift occurs is the geometric center of the two break frequencies, that is, $\sqrt{\omega_1\omega_2}$.

Principles of Analog Computers and Servomechanisms

19.1. Analog Computers. An electronic analog computer[1] performs mathematical operations on electrical quantities which are the actual electrical quantities and/or equivalent electrical quantities which have been converted from the physical variables involved in the problem. An analog computer accepts, utilizes, and produces data in continuous form. This is contrasted to digital computation in which data representing the quantities involved in the computation are represented by a series of pulses signifying the discrete magnitudes of the quantities. The term computer as used herein refers not only to a complete multipurpose equipment capable of performing many grouped operations but more particularly to the simple computing elements used in electronic equipment for the purpose of performing mathematical operations such as addition, multiplication, or integration of electrical signals. Analog elements have application where it is desired to compute to an accuracy of three or four significant figures. Where greater accuracy is required and justified by the accuracy of the input data, digital techniques are necessary. In general, the cost and complexity of analog computation are less than digital computation within the accuracy limitation mentioned. A further advantage of analog computation is that once the basic building blocks are understood, a combination of elements to perform a given computation can be synthesized by engineers who have not specialized in computer design.

Analog computation is convenient in that items of physical equipment not normally associated with the computer can be most easily combined with the analog computer elements to perform an over-all computation based on the specific characteristics of the physical equipment. The effect of changing system parameters can be readily observed, and the design of physical equipment can often be optimized without lengthy calculations.

19.1a. System Considerations. The design of an analog computing system is influenced by such factors as the requirements of the problem or equation to be solved, computation time, scale factors, and the allowable over-all error. Factors determining the selection of specific computing elements must also be considered. For example, analog operations can be performed by mechanical, electromechanical, or purely electronic means. Electronic computing elements are usually preferred over purely mechanical elements because the response time can be much faster, the over-all computation time can be reduced, and the cost of electronic elements is usually less for the same flexibility. Additional factors such as availability, inherent errors, freedom from wear, stability, and ease of adjustment enter into the determination of the specific elements to be used.

To simplify the description of an analog computing system, block diagrams are usually drawn to show the interrelations of the specific computing elements. Table 19.1 gives the notations commonly used in the preparation of such block diagrams.

---

[1] For general references on electronic computers, see: I. A. Greenwood, Jr., J. Vance Holdam, Jr., Duncan MacRae, Jr., "Electronic Instruments," McGraw-Hill Book Company, Inc., New York, 1948. G. A. Korn and T. M. Korn, "Electronic Analog Computers," McGraw-Hill Book Company, Inc., New York, 1952. Computer Issue, *Proc. IRE*, "High-speed Computing Devices," McGraw-Hill Book Company, Inc., New York, 1950.

TABLE 19.1. SYMBOLS FOR ANALOG ELEMENTS
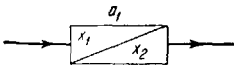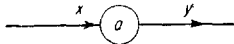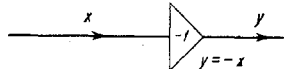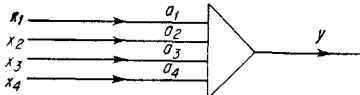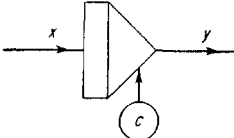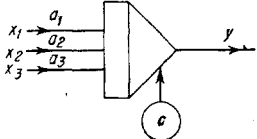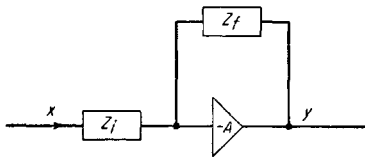
Element	Symbol	Function
Transducer		Data translation: $x_1 =$ input quantity $x_2 =$ output quantity $a_1 =$ scale factor $x_2 = a_1 x_1$
Potentiometer	$y = ax$ $a \leq 1$	$a =$ scale factor by which input is multiplied to form output
Inverting amplifier	$y = -x$	Inversion of sign
Summing amplifier	$y = -(a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4)$	Summing several quantities with inversion and required scale changes
Integrator amplifier	$y = -a \int_0^{t_1} x \, dt + c$	Integration of a function x with respect to time. $c =$ initial condition (integration constant), and $t_1 =$ integration time. $a =$ scale factor $= 1/RC$
Summing integrator	$y = -\int_0^{t_1} (a_1 x_1 + a_2 x_2 + a_3 x_3) \, dt + c$	Combined summing and integration. The scale factors $a_n = 1/R_n C$
General operational amplifier		General computing transfer function $\frac{y}{x} \simeq -\frac{Z_f}{Z_i}$ for $A \gg 1$

TABLE 19.1. SYMBOLS FOR ANALOG ELEMENTS (*Continued*)

Element	Symbol	Function
Servo-mechanism		Shaft positioning in accordance with an input signal
Special		Special function generation or operation denoted by label on box

The first requisite in the specification of requirements is the reduction of the problem to a written explicit or implicit mathematical expression or set of expressions to be solved by the computer. The physical significance of each item of datum required for the solution should then be noted, including the range of values expected. If input data are available as a physical quantity but are nonelectrical in nature, notation should be made of the transducers required to convert these quantities into electrical equivalents suitable for insertion into the computer. If input data are not conveniently available, they can often be simulated within the computer.

The desired form of the solution or other output data should also be specified (i.e., graphical recordings, meter readings, potentiometer settings, etc.). In addition, the over-all error allowable in the solution should be given. This should be consistent with the known input-data error, including transducer error, and is often the determining factor in arriving at the specific form of instrumentation to be used.

19.1b. Computation Time and Scale Factors. Scale factors describe the relationship between the magnitude and dimensions of one quantity and the magnitude and dimensions of another quantity in the computer. The scale factors used depend upon the accuracy required and the maximum magnitudes which can be handled by each analog element without overloading or limiting. The problem variables are denoted by lower-case letters and the corresponding equipment variables by upper-case letters, e.g.,

$$a_x = X/x \qquad (19.1)$$

For example, if a particular variable is bearing angle from 0 to 360° and if this is to be represented by a voltage excursion in the computer of 100 volts peak to peak, the scale factor a is

$$a = (100 \text{ volts})/360° = 0.278 \text{ volt/deg}$$

This over-all scale factor includes the scale factor of the particular transducer used. For example, if a transducer is available to convert angular rotation to voltage at a scale factor of 1 volt/deg, the input is 360 volts for 360° rotation. To limit the maximum input to 100 volts, a 3.6:1 voltage division must take place between the transducer output and the computing element.

If time t appears in the problem equation and is to be converted to an analog voltage T, it will have a scale factor just as any other variable. This is not to be confused with the time reference τ of the computer. The computer operates on "real time," τ, i.e., rates of integration and computation times are actual values. The time scale

τ may be different from the time variable t of the problem equation and is also related to t by a scale factor.

$$a'_t = T/t \tag{19.2}$$

$$a_t = \tau/t \tag{19.3}$$

where a'_t = scale factor relating problem time to equivalent computer analog voltage

$\quad\quad a_t$ = scale factor relating problem time to equivalent computer time

For example, if $a_t = 100$, rates of change are reduced 100:1 in the computation as compared to the indicated rate in the problem equation. Thus 1 sec in the problem equation requires 100 sec in the computer. The quantity a_t is usually selected to have as small a value as possible for integrators since their errors increase with time, and to have a large value for narrow-bandwidth computer elements (e.g., electromechanical servos) requiring a relatively long computation time. The value of a_t must, however, be the same for the entire computation. The speed of computation is therefore dictated by the requirement of the slowest element. If servos are not used, the maximum speed of computation is often limited by stability problems associated with high gain electronic amplifiers. The limitation on increasing computation time is usually allowable integration drift. In some cases, it is desirable to make $a_t = 1$ so that the problem will progress in the computer in accordance with "real time." The value of a_t must be unity on any computation using actual equipment in conjunction with the computer unless it is known that the actual equipment can be operated on a modified time scale without introducing dynamic errors.

The choice of scale factor will depend also upon the computation involved. For example, if the maximum allowable level at any portion of the computer is 100 volts and if 2 voltages are to be multiplied, the scale factors must be adjusted so that the *product* of the two variables does not exceed 100 volts. In a similar manner, care must be exercised in the establishment of a computer time scale to ensure that the output of differentiation does not exceed the maximum allowable value. For example, if the problem calls for $y = dx/dt$ and if dx/dt is 1,000 volts/sec, either x must be scaled to 10:1 or a_t must exceed 10 so that $dX/d\tau$ is less than 100 volts for the time interval over which the derivative is taken.

19.1c. Error Analysis. An error analysis should be made of proposed computer configurations to determine if the probable error is consistent with the requirements of the problem. In general, this consists of a tabulation of all known sources of error. Representative sources of error are adjustment tolerances, component tolerances, approximations involved in the computer equation, calibration errors, nonideal operation of analog elements, input-data errors, and output-data-reading errors. Errors must be reviewed to determine if there are interrelations which cause two errors to be mutually compensating, or which, on the other hand, might cause the errors to always be directly additive. The computer configuration should be examined for opportunities to decrease the total error.

In general, all errors will not add directly, and some means of establishing a "probable" error must be used. A basis which has been suggested is to take one-third to one-fifth of the square root of the sum of the squares of the peak values of the individual errors. An additional refinement is to assign "cross-correlation" functions describing the dependence between errors when known. The function would have a value of $+2$ where the individual errors were directly related and were additive in the same direction. This function would have a value of -2 where the two individual errors were directly related but were compensating and the effects were subtractive. This function would have a value of zero for any two errors not related. For two errors, this results in:

$$(\Delta\epsilon)_{\text{probable}} \simeq \frac{\sqrt{(\Delta\epsilon_1)^2 + (\Delta\epsilon_2)^2 + \Delta_{12}\Delta\epsilon_1\Delta\epsilon_2}}{4} \tag{19.4}$$

where Δ_{12} is the cross-correlation function relating $\Delta\epsilon_1$ to $\Delta\epsilon_2$. ($\Delta\epsilon_1$, and $\Delta\epsilon_2$ are peak values.)

An important simplification can sometimes be used in determining probable errors if the *relative error* between successive computations is of importance and the *absolute error* is not. In this case, Eq. (19.4) need contain only those factors significant in the computation of relative error.

19.2. Operational Amplifiers. An operational amplifier is a computing element employing a high-gain d-c amplifier together with input and feedback impedances as shown in Fig. 19.1a. The input impedance Z_i and feedback impedance Z_f can be any

(*a*) BASIC FORM OF OPERATIONAL AMPLIFIER

(*b*) OPERATIONAL AMPLIFIER USED   (*c*) OPERATIONAL AMPLIFIER FOR
FOR ADDITION OF e_1 AND e_2 INTEGRATION OF e_i
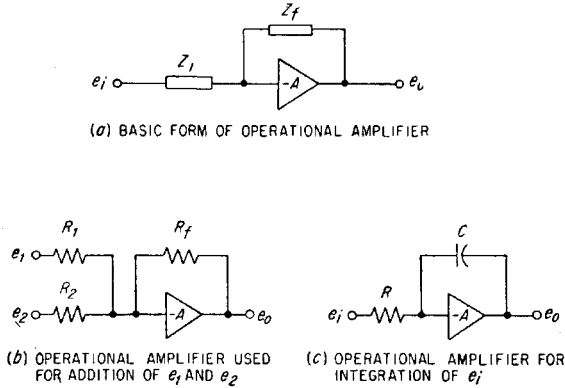
FIG. 19.1. Operational amplifiers.

electrical network subject to the requirement that the amplifier with the associated impedances be stable. The "transfer function" of this configuration is

$$\frac{e_o}{e_i} = \frac{-Z_f}{(Z_i + Z_f)/A + Z_i} \tag{19.5}$$

where Z_i = transfer impedance function[1] of input impedance
       Z_f = transfer impedance function of feedback impedance
If the gain is very large, Eq. (19.5) simplifies to be the ratio of the feedback impedance to the input impedance as shown in Eq. (19.6):

$$\frac{e_o}{e_i} \simeq \frac{-Z_f}{Z_i} \tag{19.6}$$

Examples of operational amplifiers as computing elements are shown in Fig. 19.1b and c.

Operational amplifiers are used in analog computation to provide a required transfer function in accordance with Eq. (19.6), to provide isolation and/or inversion, to provide scale changes, or to solve implicit functions. These subjects are discussed in Sec. 19.3.

A majority of the problems associated with the design of an operational amplifier are concerned with the design of the high gain d-c amplifier. Of particular importance are minimization of drift in the d-c amplifier and proper shaping of the frequency response to achieve stability in the operational amplifier (see Sec. 3.19c). Amplifier drift can be especially undesirable; for example, in an operational amplifier used as an integrator, very small differences from zero input potential can be integrated to produce large output errors. This type of error can be minimized by "balancing" the d-c amplifier, i.e., by adjusting it for zero output with the input grounded through a resistance equal to the signal source resistance, just prior to performing the desired

[1] See Table 1.5 for a tabulation of transfer impedance functions.

integration.   Even if the balancing is done just prior to computation, a small drift rate can be objectionable if the computation time is appreciable.   A reduction in drift and an increase in low-frequency gain can be realized by "chopper stabilization" (see Sec. 3.19c).   The design of d-c amplifiers suitable for use as operational amplifiers is covered in Sec. 3.19.   Assuming that a stable, drift-free high-gain d-c amplifier is available, the computation accuracy of an operational amplifier is determined by the accuracy and stability of the input and feedback impedances and the absolute gain of the amplifier.

The general requirements for the stabilization of feedback systems are discussed in Sec. 18.   Stabilizing networks applicable to operational amplifiers and servomechanisms are discussed in Sec. 19.7c.

19.3. Mathematical Operations.   There are several representative circuit techniques for performing mathematical operations in analog computation.[1]   The simplest computing networks contain only passive elements; however, the use of active circuits such as high-gain amplifiers or servomechanisms increases the accuracy of the computation and overcomes the attenuation associated with passive networks at the penalty of increased cost and complexity.

19.3a. Explicit and Implicit Solutions.   A given problem can be solved explicitly or implicitly.   An explicit solution requires the formation of the direct analog of the

(a) EXPLICIT DIVISION

(b) IMPLICIT DIVISION

(c) GENERAL IMPLICIT COMPUTATION
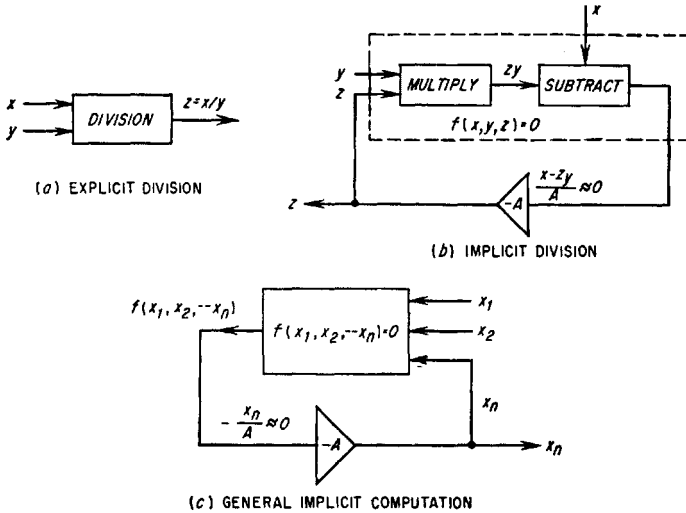
FIG. 19.2. Explicit and implicit computation.

explicit equation for the required unknown.   An implicit solution allows the solution of a rearranged version of this equation which may be easier to instrument.   For example, the problem of determining the value of z from the relationship $z = x/y$ can be solved by explicitly dividing x by y to obtain z.   An implicit solution for z is obtained from the instrumentation of the equation $zy - x = 0$.   In this case, the implicit solution requires multiplication where the explicit solution requires division.   In a similar manner, explicit subtraction can be replaced by implicit addition, integration by differentiation, powers by roots, etc.   Figure 19.2a and b shows each method of solution.   The use of a high-gain amplifier as shown in Fig.

[1] For a detailed description of mechanical computing elements, see M. Fry, Designing Computing Mechanisms, *Machine Design*, August, 1945, to February, 1946, or Antonin Svoboda, "Computing Mechanisms and Linkages," McGraw-Hill Book Company, Inc., New York, 1948.

19.2b is characteristic of implicit solutions. The general method of implicit computation is indicated schematically in Fig. 19.2c. The function $f(x_1, x_2, \ldots, x_n) = 0$ is formed from the explicit equation to be solved. For example, if $z = (x^2 + y^2)^{1/2}$ is to be solved for z, $f(x_1, x_2, \ldots, x_n) = 0$ is $(z^2 - x^2 - y^2) = 0$. The computation solves the equation:

$$f(x_1, x_2, \ldots, x_n) = -\frac{x_n}{A} \tag{19.7}$$

If the gain A of the amplifier is very high, x_n/A is very small and a value of x_n is obtained which satisfies the expression

$$f(x_1, x_2, \ldots, x_n) \simeq 0 \tag{19.8}$$

The error Δx_n inherent in this form of computation is

$$\Delta x_n \simeq -\frac{x_n}{A} \frac{1}{\dfrac{\partial f(x_1, x_2, \ldots, x_n)}{\partial x_n}} \tag{19.9}$$

Care must be exercised in the solution of implicit expressions to be sure that the closed loop is stable (see Sec. 19.6) and that the expression does not have complex (i.e., real and imaginary) solutions.

Example 19.1

Diagram the solution by implicit means of the explicit expression:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Solution

1. Determine the implicit expression for x. The required expression for

$$f(x_1, x_2, \ldots, x_n) = 0$$

is

$$ax^2 + bx + c = 0$$

2. Determine the configuration required to instrument $(ax^2 + bx + c) = 0$.

This is illustrated in Fig. 19.3a. The value of x determined by this system is a solution of the original explicit expression provided a real solution exists. The explicit solution of this problem is shown in Fig. 19.3b for comparison.

This example illustrates the possible simplification in instrumentation which can be achieved by implicit techniques. Implicit computation can be complicated by the problem of ensuring that the configuration seeks the desired solution if more than one solution is possible.

19.3b. Specific Mathematical Operations. The methods most generally employed for performing specific mathematical operations in analog computers are discussed herein. In each application, it is necessary to consider such factors as input loading, output impedance, mutual coupling, bandwidth and type of signal to be used, accuracy required, and allowable complexity.

Addition and Subtraction. Addition networks are shown in Fig. 19.4. In general, addition networks can be used for subtraction by inverting the sign of the input to be subtracted by use of an inverting amplifier. The passive parallel impedance network of Fig. 19.4a is simple and inexpensive. Any number of inputs can be added in this manner. The principal disadvantage of this circuit is the attenuation between input and output. The output voltage is given by

$$e_o = R_o \left(\frac{e_1}{R_1} + \frac{e_2}{R_2} + \cdots + \frac{e_n}{R_n} \right) \left(\frac{1}{1 + R_o/R_1 + R_o/R_2 + \cdots + R_o/R_n} \right) \tag{19.10}$$

The second parenthetical expression can be neglected for $R_o(1/R_1 + 1/R_2 + \cdots$
$1/R_n) \ll 1$; that is, R_o much smaller than the parallel combination of the input
resistances. This allows the input voltages to be simply added with an amplitude
factor associated with each term equal to R_o/R_n.

(a) IMPLICIT SOLUTION

(b) EXPLICIT SOLUTION
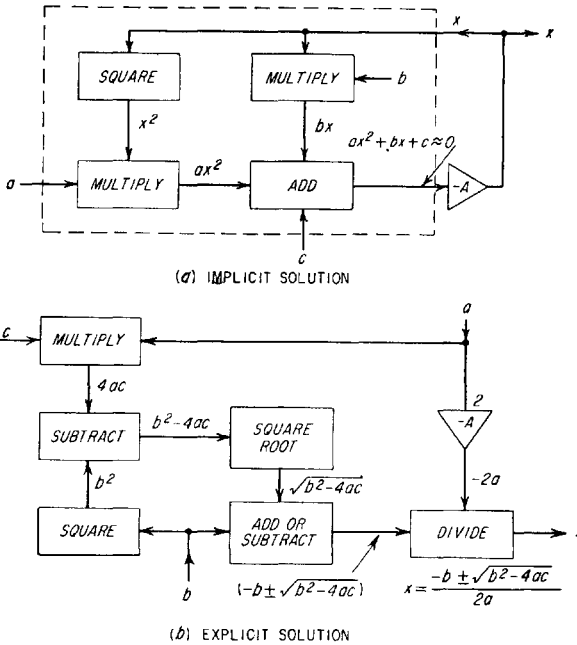
FIG. 19.3. Figure for Example 19.1.

The summing amplifier shown in Fig. 19.4b overcomes the attenuation objection
and produces less error for the same scale factor. The output voltage of this circuit
is given by

$$e_o = -R_o \left(\frac{e_1}{R_1} + \frac{e_2}{R_2} + \cdots + \frac{e_n}{R_n} \right) \tag{19.11}$$

for $|A| \gg 1 + R_o \left(\dfrac{1}{R_1} + \dfrac{1}{R_2} + \cdots + \dfrac{1}{R_n} \right)$, where A is the gain of the amplifier.

This network inverts the algebraic sign of the sum.

The passive-series addition network of Fig. 19.4c is useful where identical scale
factors are desired for each input. The output voltage is given by

$$e_o = (e_1 + e_2 + \cdots + e_n) \left(\frac{R_o}{R_o + R_1 + R_2 + \cdots + R_n} \right) \tag{19.12}$$

The sources $e_1, e_2, \ldots, e_n$ must be isolated from a common terminal in this net-
work. For a-c signals, transformer coupling can be used.

The bridge network shown in Fig. 19.4d is useful as an addition circuit. If all the
arms have equal resistance values, the output voltage is given by

$$e_o = \frac{R}{2} \left(\frac{e_1}{R + r_1} + \frac{e_2}{R + r_2} \right) \tag{19.13}$$

where r_1, r_2 = source resistances of e_1 and e_2, respectively

Although one source must be isolated from a common terminal, this circuit has the advantage of providing negligible coupling between the two sources. Additional terms can be added by utilizing additional bridge networks.

Tube circuits shown in Fig. 19.4e have the advantage of providing signal gain, or at least negligible attenuation, depending on the input and output connections; however, the accuracy does not compare with the circuits shown in Fig. 19.4a through d due to both the tube nonlinearities and drift in characteristics. Also, the output appears on a d-c base which may be undesirable. There is negligible coupling between inputs and between input and output circuits. The multiple-input circuit of Fig. 19.4e(1) requires only a single tube, and only small signals can be handled without appreciable

(a) PASSIVE PARALLEL NETWORK

(b) OPERATIONAL AMPLIFIER NETWORK

(c) PASSIVE SERIES NETWORK

(d) BRIDGE NETWORK

(e) VACUUM TUBE METHODS
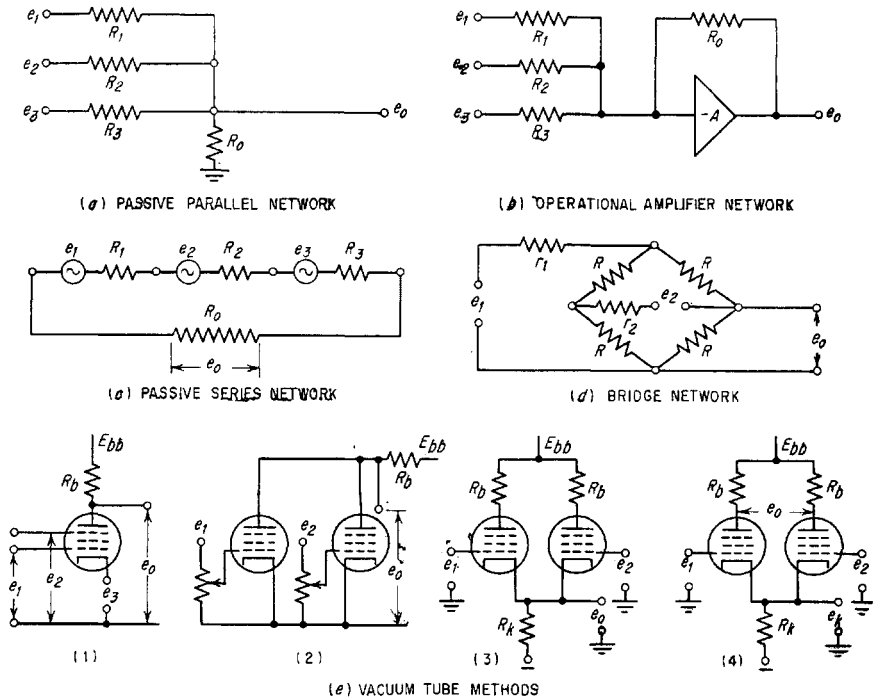
FIG. 19.4. Addition and subtraction networks.

error. The output voltage e_o is a function of the sum of the grid inputs e_1 and e_2 from which the cathode input e_3 is subtracted. The impedance of the source e_3 must be low, or the gain of the stage will be reduced by cathode degeneration. An additional inverting amplifier is required if it is necessary to restore the original polarity of the input signals.

The common-plate addition circuit of Fig. 19.4e(2) has the advantage of allowing a number of inputs limited only by the number of tubes that can be operated in parallel. The scale factor of each input can be controlled by the adjustment of the grid circuit potentiometers. If the scale factor of each input is to be independent of the number of other inputs, the plate resistance of each tube must be much larger than R_b. Therefore pentodes are usually used.

The common-cathode addition circuit of Fig. 19.4e(3) is similar to the common-plate circuit except for the additional advantages of no signal inversion, low output impedance, and less variation due to tube characteristics. The scale factor of each

input can be adjusted by the addition of either grid circuit potentiometers or a separate cathode resistor between the cathode of each tube and the common-cathode resistor R_k.

The differential amplifier shown in Fig. 19.4e(4) provides a voltage between plates proportional to the difference between grid signals, and a voltage at the cathode proportional to the sum of the grid voltages.

Multiplication and Division. Representative methods of performing multiplication and division are shown in Fig. 19.5. A problem in the design of division circuits is the
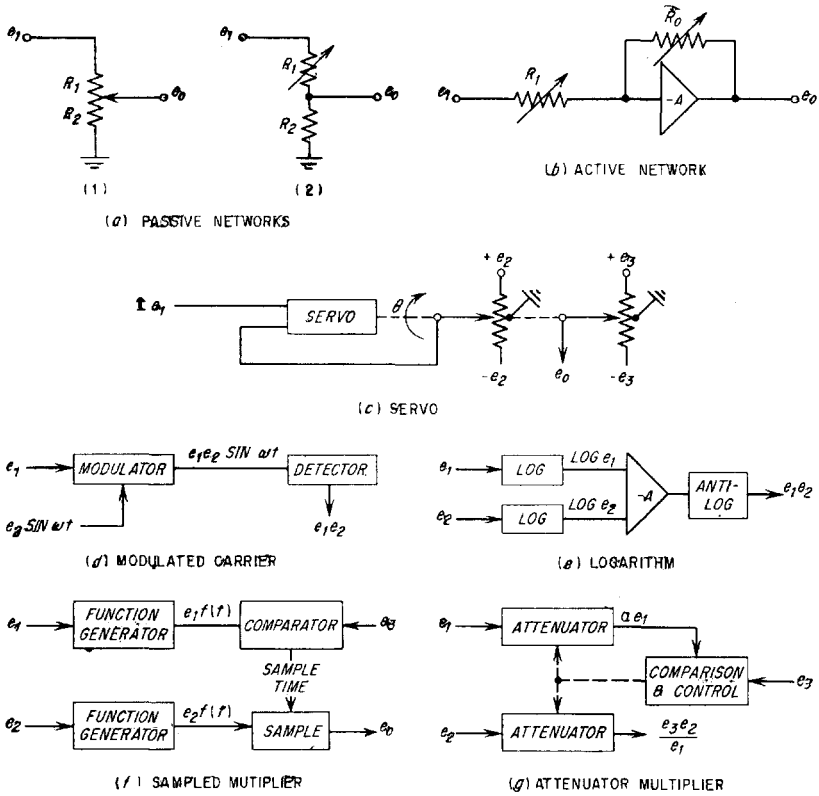
FIG. 19.5. Multiplication and division.

limiting or overloading which can occur for small values of the divisor. A factor to be considered in multiplying circuits is the allowable algebraic sign of each input. If the sign of neither input can be reversed, the multiplier is a single-quadrant device. If one of the input voltages can have either a positive or negative sign and if the sign of the product changes accordingly, the device is a two-quadrant multiplier. If the multiplier accepts either sign for each input and produces the proper sign of the product, it is a four-quadrant multiplier.

Multipliers or dividers using servo or carrier systems are, in general, capable of accuracies of 0.1 per cent but have relatively long response times. Electronic multipliers are capable of rapid multiplication or division. Complex electronic multipliers can have accuracies comparable to servo multipliers; simple vacuum-tube multipliers, however, are relatively inaccurate.

The passive networks of Fig. 19.5a are used for multiplying or dividing by a constant. The output voltage is given by

$$e_o = e_1 R_2 / (R_1 + R_2) \qquad (19.14)$$

In the first network, the value of R_2 is selected by the position of the slider arm of the potentiometer. The circuit multiplies e_1 by a constant proportional to R_2. If R_1 is the adjusted element in the second circuit, the output will be divided by $(R_1 + R_2)$. In practice, the loading effect of impedances connected to e_o must be considered. This is shown in Fig. 19.8 where α is the ratio of loading resistance to total potentiometer resistance and x is the ratio of R_2 to $R_1 + R_2$. These circuits have the disadvantage of attenuating the input signal. Since the impedance of the second network is variable, its loading effect of the input voltage e_1 may be important.

The operational amplifier of Fig. 19.5b multiplies the input e_1 by a constant scale factor. In this case,

$$e_o = \frac{-R_o e_1}{(R_1 + R_o)/A + R_1} \simeq \frac{-R_o e_1}{R_1} \qquad \text{for } |A| \gg 1 \qquad (19.15)$$

This network therefore inverts and changes scale by the factor R_o/R_1. If R_o is servo-controlled by a second input e_2, the circuit multiplies e_1 by e_2. If R_1 is servo-controlled by e_2, the circuit divides e_1 by e_2. In the case of division the effect of a changing input resistance on the source e_1 must be considered.

The servo system shown in Fig. 19.5c drives the two potentiometer arms by an angular amount depending upon the input e_1. The first potentiometer provides a feedback comparison voltage for the servo. The voltage of the second potentiometer arm which is mechanically connected to the first potentiometer is

$$e_o \simeq \frac{e_1 e_3}{e_2} \qquad \text{for } e_2 > e_1 \qquad (19.16)$$

The over-all accuracy achievable with this system is about 0.1 per cent.

The modulated carrier system shown in Fig. 19.5d represents one version of a general class of multiplier.[1] One input e_1 is used to modulate a second input $e_2 \sin \omega t$ which consists of an a-m carrier signal. The doubly modulated carrier is then detected,[1] and the detector output is proportional to $e_1 e_2$. A limitation of any carrier system is that the highest important frequency component of e_1 and e_2 must be substantially lower than the carrier frequency.

The configuration shown in Fig. 19.5e represents a general method of multiplication by auxiliary functions. This circuit solves the equation

$$-\log e_1 e_2 = \log e_1 + \log e_2 \qquad (19.17)$$

By reversing the sign of one of the inputs to the summing amplifier, division can be performed in accordance with Eq. (19.18):

$$\log (e_1/e_2) = \log e_1 - \log e_2 \qquad (19.18)$$

Elements for generating logarithmic and antilogarithmic functions are rectifiers, potentiometers, vacuum tubes, and nonlinear resistors.

Multiplication can be accomplished by integration as shown by Eq. (19.19):

$$e_1 e_2 = \int e_1 \, d(e_2) + \int e_2 \, d(e_1) \qquad (19.19)$$

This method of multiplication can be very accurate if precision integrators are employed.

[1] See, for example, G. D. McCann, C. H. Wilts, and B. N. Locanthi, "Electronic Techniques Applied to Analog Methods of Computation," *Proc. IRE*, vol. 37, pp. 954–961, August, 1949.

The sampled multiplier[1] shown in Fig. 19.5f has a full-scale accuracy better than 0.2 per cent and is capable of rapid, time-shared multiplication or division. The function generators produce time variant outputs which are identical in form (i.e., linear, exponential, etc.) but which differ in amplitude in accordance with the respective inputs e_1 and e_2. The magnitude of $e_1 f(t)$ is compared with a third input e_3 in a comparator. When these two voltages are equal, the amplitude of the voltage $e_2 f(t)$ is instantaneously sampled and stored as the output voltage. The output voltage e_o is given by

$$e_o = \frac{e_2 e_3}{e_1} \tag{19.20}$$

An attenuator-multiplier[2] is shown in Fig. 19.5g. The signals e_1 and e_2 are passed through identical variable attenuators. The output of one attenuator is compared to a third signal e_3. The comparator produces an attenuation control signal to cause αe_1 to be equal to e_3 where α is the attenuation. Under these conditions, the output from the e_2 attenuator is also given by Eq. (19.20). It should be noted that the output signal will always be less than e_2.

Many other forms of multipliers are possible. For example, "area integration" can be used wherein the amplitude and duration of a signal are each proportional to the quantities to be multiplied. The integral of the signal is proportional to the product of the two quantities.[3]

Integration and Differentiation. Networks for these operations are shown in Fig. 19.6. Differentiation is usually obtained in analog computation implicitly, i.e., by integration. This is because differentiators accentuate noise components of the input signal and, in addition, introduce serious closed-loop stability problems (see Sec. 3.19c).

The passive differentiator of Fig. 19.6b has a transfer function given by

$$\frac{E_o(s)}{E_i(s)} = \frac{RCs}{1 + RCs} \tag{19.21}$$

A "perfect" differentiator has a transfer function equal to s. The passive differentiator is effective for frequencies where $\omega RC < 1$. To differentiate high-frequency components of the input signal, RC should be small. .However, a small RC product causes high attenuation of the input signal. To avoid these problems, the operational amplifier differentiator of Fig. 19.6d can be used. The transfer function of this network is

$$\frac{E_o(s)}{E_i(s)} = -\frac{A}{1 + A} \frac{RCs}{[RC/(1 + A)]s + 1} \simeq -RCs \text{ for } A \gg (1 + RCs) \tag{19.22}$$

The high-gain feedback amplifier has the effect of increasing the frequency range of the differentiator by the factor $(1 + A)$ without appreciably affecting the attenuation.

The passive integrator of Fig. 19.6a has a transfer function given by

$$\frac{E_o(s)}{E_i(s)} = \frac{1}{RCs + 1} \tag{19.23}$$

---

[1] Complete circuits and discussions are given in: H. Freeman and E. Parsons, A Time Sharing Analog Multiplier, *Trans. IRE Professional Group on Electronic Computers*, March, 1954, pp. 11–17; John Broomall and Leon Riebman, A Sampling Analog Computer, *Proc. IRE*, vol. 40, pp. 568–572, May, 1952.

[2] See Korn and Korn, *op. cit.*, pp. 223–224, or Greenwood, et al., *op. cit.*, pp. 50–53.

[3] For a detailed description of a highly accurate multiplier of this general type, see E. A. Goldberg, "A High Accuracy Time Division Multiplier," Project Cyclone Symposium II, pt. 2, April 28–May 2, 1952, Reeves Instrument Co., sponsored by the U.S. Navy. Also, in the same report, W. A. McCool, "An AM-FM Electronic Analog Multiplier."

Eq. (19.23) shows that the transfer function reduces to $1/RCs$ for frequencies where $s \gg 1/RC$. A "perfect" integrator has a transfer function of $1/s$. In Fig. 19.6a, the quality of integration is improved as RC and/or ω is increased. An increase in RC in Eq. (19.23) increases the attenuation between E_i and E_o. To overcome this difficulty the operational amplifier integrator of Fig. 19.6c is used. This basic circuit is

(a) PASSIVE INTEGRATION

(b) PASSIVE DIFFERENTIATION

(c) OPERATIONAL AMPLIFIER INTEGRATION

(d) OPERATIONAL AMPLIFIER DIFFERENTIATION

(e) SUMMING INTEGRATOR

(f) HOLDING AMPLIFIER

(g) SERVO INTEGRATOR
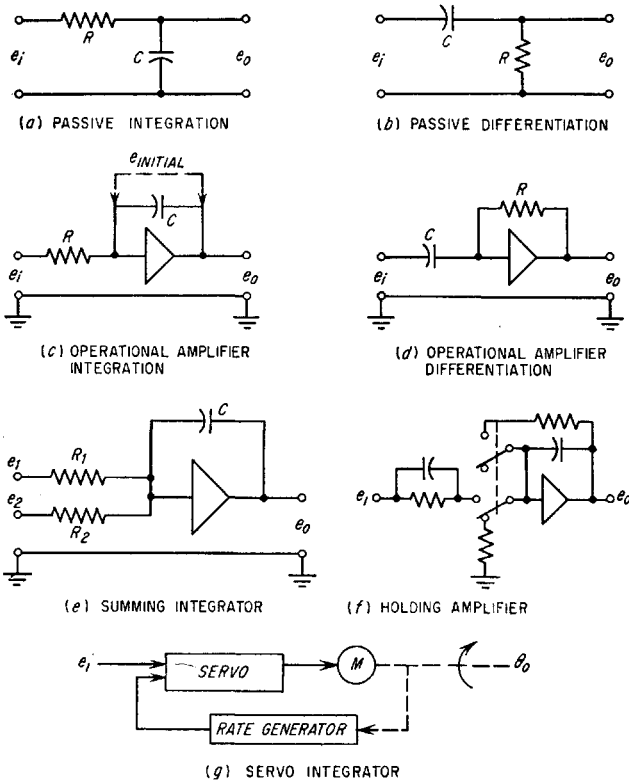
FIG. 19.6. Integration and differentiation.

one of the most useful analog computing elements. The transfer function of this circuit is

$$\frac{E_o(s)}{E_i(s)} = \frac{-A}{(1 + A)RCs + 1} \simeq \frac{-1}{RCs} \quad \text{for } A \gg (1 + RCs) \quad (19.24)$$

The rate of integration is approximately the same as the initial rate of integration in a passive integrator. The integration error, however, can be substantially less than that of a passive integrator.

Practical considerations such as grid current (see Sec. 3.19), temperature drift, and capacitor leakage are very important in the design of a high-quality integrator. The effects of other circuit elements, such as stray capacitance, source resistance, load resistance, and amplifier input resistance on the values of R and C must be included. The input resistance R is usually a high-precision low-temperature-coefficient resistance. The feedback capacitance is usually made with polystyrene or some other very low leakage dielectric material. The integration constant, or "initial condition,"

is added by charging the feedback capacitor to the initial-condition voltage just prior to initiation of the integration period.

If the integration is to be taken with respect to a variable y instead of time, the integration can be performed as shown in Eq. (19.25):

$$\int x \, dy = \int \left(x \frac{dy}{dt} \right) dt \qquad (19.25)$$

The time derivative of y is first obtained. This is multiplied by the variable x, and the product is integrated with respect to time to give the required integral of x with respect to y.

The summing integrator shown in Fig. 19.6e is a useful device for combining the operations of summing and integration. The output signal is equal to

$$E_o(s) \simeq -\frac{1}{Cs} \left(\frac{E_1(s)}{R_1} + \frac{E_2(s)}{R_2} + \cdots + \frac{E_n(s)}{R_n} \right) \qquad (19.26)$$

A special application is illustrated in Fig. 19.6f. This "holding amplifier" is used to store a signal value at a chosen instant. This can be used to obtain a "point" solution to a problem during a computation, to store a value for later use in the computation, or to check the computer during the problem. The normal connection is with the relay arms in the up position. The input and feedback impedances are then composed of resistances shunted by capacitors of such values that the time constant of each network is the same. In this way the amplifier acts as an inverting amplifier and can be connected to the circuit whose output it is desired to "hold." At the hold instant, the resistance in the feedback impedance is disconnected and the input impedance is replaced with an equivalent resistance to ground at the input to the amplifier. There is zero input signal under this condition, and the stored signal is the last output signal on the feedback capacitor.

A servo integrator is illustrated in Fig. 19.6g. The system consists of a servo amplifier driving a motor. The motor speed is proportional to the voltage from the servo amplifier. A tachometer or rate generator is mechanically coupled to the motor shaft so that it generates a voltage proportional to the motor speed. This voltage is fed back into the servo amplifier where it is compared to the input signal. The motor speed is controlled by the servo amplifier so that the rate voltage approaches the input voltage. Under this condition, the angular displacement of the motor shaft is approximately proportional to the integral of the input signal. A serious limitation of servo integrators is the response time due to the inertia of the rotating components. This restricts the use of such integrators to applications wherein the signals have very slow rates of change compared to the response time. Another restriction is the relative inaccuracy of the servo integrator due to tachometer inaccuracies.

Function Generator. Another mathematical operation required in analog computation is function generation. Function generators can be formed from simple potentiometers, electromechanical servo systems, or electronic circuitry of varying complexity. Representative methods[1] of function generation are illustrated in Fig. 19.7.

The "photoformer"[2] of Fig. 19.7a consists of a cathode-ray tube and associated deflection circuitry, a mask placed over the tube face cut to the shape of the desired function, and a phototube with an associated amplifier. The input signal inserted into the horizontal deflection amplifier causes the cathode-ray trace to sweep horizontally across the tube. The vertical deflection circuitry causes the trace to move

[1] For a discussion of specific methods of performing squaring and square-root operations, see B. Chance et al., "Waveforms," chap. 19, McGraw-Hill Book Company, Inc., New York, 1949, or Greenwood et al., *op. cit.*, chap. 6.

[2] See D. E. Sunstein, "Photoelectric Waveform Generator," *Electronics*, vol. 22, p. 100, 1949.

upward until it is visible in the unmasked portion of the cathode-ray tube.   The
phototube detects the presence of the trace above the mask and feeds back a signal
which moves the beam downward.   In this manner, the cathode-ray beam follows
the contour of the mask as it is swept horizontally in accordance with the input signal.
The vertical deflection signal is proportional to the required function and is used as
the output signal of the photoformer.
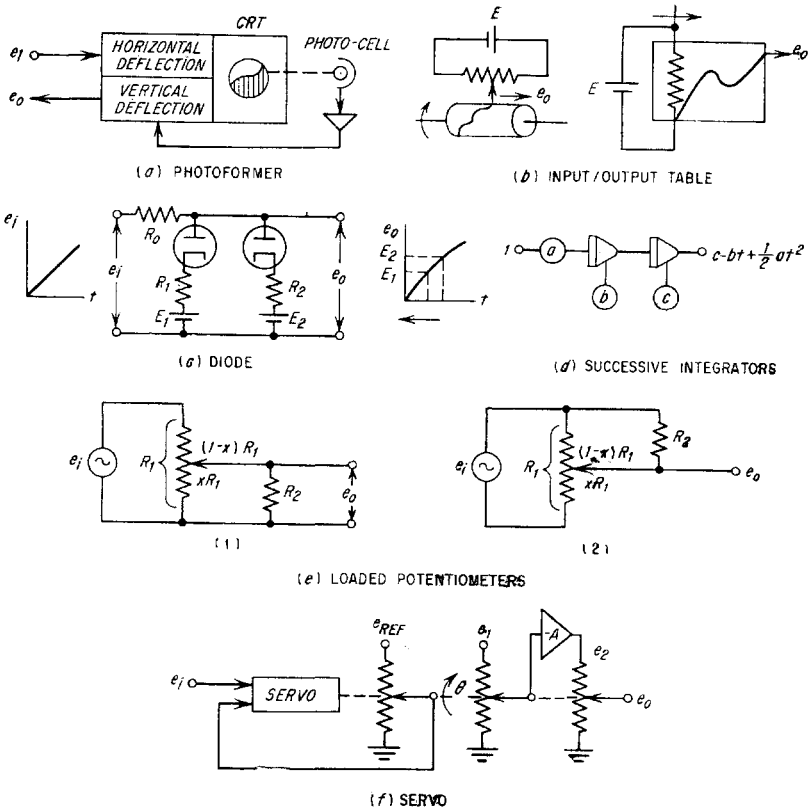
Fig. 19.7. Function generators.

A mechanical equivalent of the photoformer can be constructed by using an input/
output table as shown in Fig. 19.7b.   Cylindrical and rectangular tables are dia-
grammatically represented in the figure with the required function attached to the
table in the form of a contact wire.   The wire serves as the contact to a potentiometer
as shown.   The required function is available from the contact as a voltage varying
with drum rotation or linear displacement of the resistance element in accordance
with the function plotted by the wire.

A diode function generator is illustrated in Fig. 19.7c.   The required function is
approximated by a series of straight lines.   A series of diodes are caused to have
successive limit levels[1] such that the sum of the output of the diodes approximates the
required function.

[1] See Sec. 12 for a discussion of the design of diode limiters.

A method of generating polynomials of time by the use of successive integrators is shown in Fig. 19.7d.   The input to the successive integrators in the figure has been assumed to be unity.

Potentiometers[1] are of general use in generating linear signals as a function of shaft rotation.   The shaft rotation is often servo-controlled in accordance with some input signal to the servo amplifier.   Over-all potentiometer linearities of better than 0.1 per cent are obtainable.   A significant source of error is the effect of the load[2] resistance on the potentiometer output signal.   This effect is used to advantage in the generation of nonlinear functions with linear potentiometers.   The possible transfer functions for the two configurations shown in Fig. 19.7e are plotted in Fig. 19.8 as a
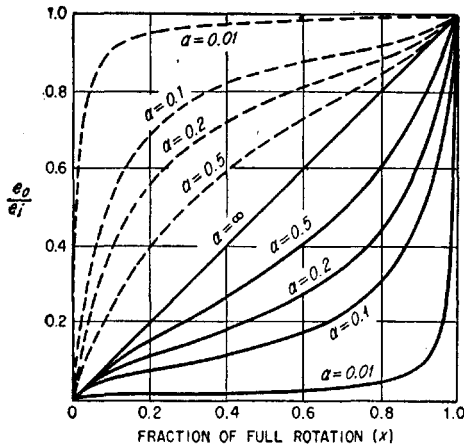
FIG. 19.8. Loaded potentiometer functions.

function of α, where $\alpha = R_2/R_1$.   The solid lines apply to circuit (1).   The dashed lines apply to circuit (2).   The curve for $\alpha = \infty$ applies for either case.

Nonlinear functions can also be obtained by tapped potentiometers[3] and potentiometers with nonlinear resistance windings.

Another general type of function generator is the servo-controlled cascaded-potentiometer arrangement of Fig. 19.7f.   If the potentiometers are linear, the connection shown produces a function proportional to the square of the input voltage.   The operational amplifier shown between the potentiometers is used to provide isolation. If the signal from the output potentiometer instead of the signal from the reference potentiometer is fed into the servo amplifier for comparison with e_i, the shaft rotation is proportional to the square root of e_i.

Straight-line functions are useful in the simulation of physical conditions such as mechanical stops, hysteresis effects, backlash, frictional effects, and absolute-value

[1] For a detailed discussion of potentiometer construction, see J. F. Blackburn, "Components Handbook," chap. 8, McGraw-Hill Book Company, Inc., 1948.   For discussions of general techniques of function generation with potentiometers, see: Greenwood, et al., *op. cit.*, chap. 5; Korn and Korn, *op. cit.*, chap. 6; and Harold Levenstein, Generating Nonlinear Functions with Linear Potentiometers, *Tele-Tech*, October, 1953, p. 76.

[2] For discussions of loading error and methods of minimizing this error, see J. F. Nettleton and P. E. Dole, Reducing Potentiometer Loading Error, *Rev. Sci. Instr.*, vol. 18, pp. 332–341, May, 1947.

[3] See Korn and Korn, *op. cit.*, pp. 261–271.

determination.   Diodes or relays are convenient elements to use for this purpose.[1]
Diodes have the advantage of usually requiring less space and power, but their
operation is "soft" (i.e., they only approximate straight-line characteristics).   Relays,
on the other hand, when used in conjunction with high-gain amplifiers, have good
straight-line characteristics.
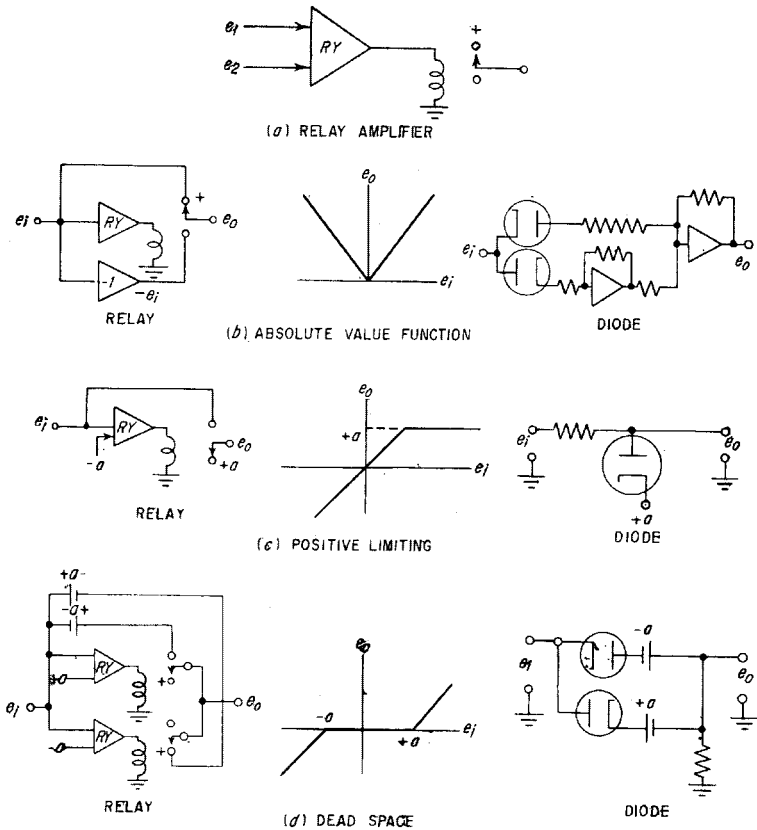
Fig. 19 9. Straight-line function generation by relays and diodes.

The relay amplifier shown diagrammatically in Fig. 19.9a is a high-gain d-c amplifier
used to actuate a high-speed relay.   The amplifier is polarity-sensitive, and only one
polarity of input signal actuates the relay.   Representative diode and relay arrange-
ments for the formation of straight-line functions are shown in Fig. 19.9b, c, and d.
For these examples, a positive net input voltage connects the relay to the plus terminal.

   Trigonometric Functions and Coordinate Transformation.   The generation of func-
tions to transform the coordinates of a problem is illustrated in Fig. 19.10.   Figure
19.10a defines the coordinates of two rectangular systems displaced in angle with

---

[1] For the design of diode circuits, see Sec. 12.   For a discussion of diode computer func-
tions, see C. D. Morrell and R. V. Baum, "The Role of Diodes in an Electronic Differential
Analyser," Project Cyclone Symposium II on REAC Techniques, April 28–May 2, 1952,
sponsored by the U. S. Navy.   For a discussion of relay function generation, see R. R.
Bennett, "The Generation of Straight Line Transfer Relationships," Project Cyclone
Symposium I on REAC Techniques, March 15–16, 1951.   The notation and principles of
relay amplifier operation discussed in this section are adopted from Bennett's article.
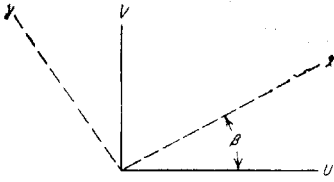
(a) ROTATION OF COORDINATES
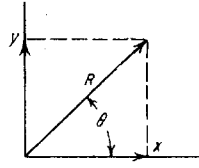
(b) RECTANGULAR AND POLAR COORDINATES
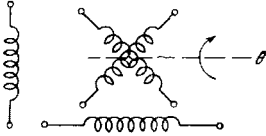
(c) RESOLVER
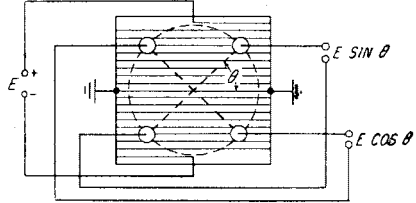
(d) SINE-COSINE POTENTIOMETER
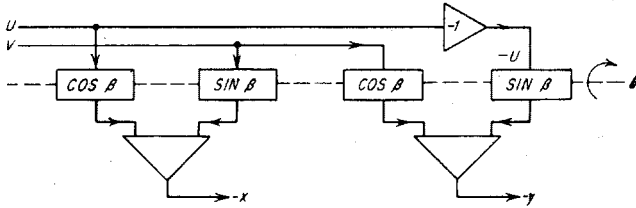
(e) ROTATIONAL COMPUTATION
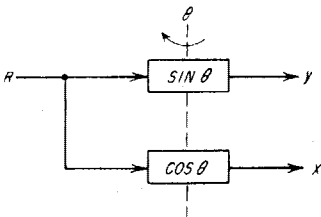
(f) POLAR TO RECTANGULAR
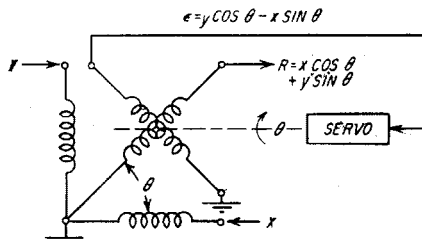
(g) RECTANGULAR TO POLAR
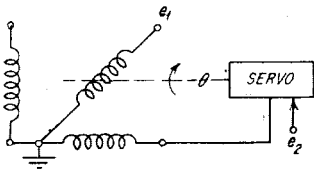
(h) INVERSE SINE AND COSINE
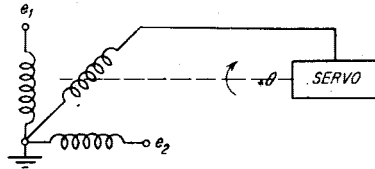
(i) INVERSE TANGENT

FIG. 19.10. Coordinate transformation and trigonometric functions.

respect to each other. It is sometimes necessary to convert variables of one coordinate system to variables of the displaced coordinate system. If the location of a point in the U, V coordinate system is known, the corresponding x and y coordinates are given by the following equations:

$$x = U \cos \beta + V \sin \beta$$
$$y = -U \sin \beta + V \cos \beta$$

(19.27)

Figure 19.10b defines the coordinates of rectangular and polar systems. These are related to each other by the following equations:

$$x = R \cos \theta$$
$$y = R \sin \theta$$

(19.28)

$$R = (x^2 + y^2)^{1/2}$$
$$\theta = \tan^{-1} (y/x)$$

(19.29)

The solution of equations of the types given by Eqs. (19.27) and (19.28) requires a convenient device to compute the sine and cosine of an angle. The resolver and the sine-cosine potentiometer[1] shown in Fig. 19.10c and d, respectively, are each generally useful for this purpose. The resolver consists of two stator windings in quadrature and two rotor windings in quadrature. If one of the stator windings is excited with an a-c voltage, voltages are induced in the rotor windings proportional to the sine and cosine of the angular position of the rotor. Conversely, if one of the rotor windings is excited, the sine and cosine functions are obtained at the stator windings.

The sine-cosine potentiometer consists of a flat card resistance winding with two contacts mechanically displaced by 90°. The terminals for the excitation voltage are 180 mechanical degrees apart. The positions of the ground taps are also 180 mechanical degrees apart, displaced 90° from the excitation voltage terminals. The voltages from the contacts will vary as the sine and cosine of the angle of rotation of the contacts.

A method of computing the coordinate rotation of Fig. 19.10a is shown in Fig. 19.10e. Signals proportional to U and V are fed into resolvers or sine potentiometers as shown. The angle β is the angular displacement between the coordinate systems. The outputs of the sine and cosine function generators are summed in accordance with Eq. (19.27) to produce the required value of the x and y coordinates. In a similar manner, the system in Fig. 19.10f transforms polar to rectangular coordinates in accordance with Eq. (19.28). The system of Fig. 19.10g will transform coordinates from rectangular to polar coordinates in accordance with Eq. (19.29). If the inputs to the stator windings are x and y, the output of one rotor winding is ($x \cos \theta + y \sin \theta$) which is equal to R. The output of the other rotor winding is ($y \cos \theta - x \sin \theta$) which is equal to zero when θ is equal to $\tan^{-1} (y/x)$. This is fed into a servo which rotates the resolver shaft until this expression is satisfied.

In a similar manner, the arcsine and arccosine functions can be produced by the arrangement shown in Fig. 19.10h. The rotor is excited by a voltage e_1. The output voltage ($e_1 \cos \theta$) from the resolver stator is compared to a voltage e_2, and the difference is fed into the servo amplifier. The servo causes the resolver shaft to rotate such that the shaft position θ is given by $\cos^{-1} (e_2/e_1)$.

The arctangent is obtained from the arrangement shown in Fig. 19.10i. In this case, both stators are excited by out-of-phase voltages e_1 and e_2. The rotor is servoed to a null by the rotor voltage which is equal to ($e_2 \cos \theta - e_1 \sin \theta$). The angle θ for a null is $\theta = \tan^{-1} (e_2/e_1)$. The arctangent is also given directly by the circuit of Fig. 19.10g.

[1] See Blackburn, *op. cit.*

19.3c. Solution of Equations.[1]   Many physical problems can be described by mathematical equations of a standard or easily recognizable form.   An important factor, therefore, in the establishment of an analog for a physical problem is the ability to instrument commonly encountered expressions.   This requires the use of the mathematical operations described in the previous sections and is illustrated in the following paragraphs.

Linear Simultaneous Equations with Constant Coefficients.   Simultaneous equations containing no derivative or integral terms are expressed in the following form:

$$\begin{aligned}
a_1x_1 + a_2x_2 + \cdots + a_nx_n + a_o &= 0 \\
b_1x_1 + b_2x_2 + \cdots + b_nx_n + b_o &= 0 \\
&\quad\quad\vdots \\
n_1x_1 + n_2x_2 + \cdots + n_nx_n + n_o &= 0
\end{aligned} \tag{19.30}$$

The general procedure for the solution of a set of equations such as this is as follows:

1. Solve for $x_1, x_2, \ldots, x_n$ in terms of the other variables.   The results appear as

$$x_1 = -\left(\frac{a_2}{a_1}x_2 + \frac{a_3}{a_1}x_3 + \cdots + \frac{a_n}{a_1}x_n + \frac{a_o}{a_1}\right)$$

$$x_2 = -\left(\frac{b_1}{b_2}x_1 + \frac{b_3}{b_2}x_3 + \cdots + \frac{b_n}{b_2}x_n + \frac{b_o}{b_2}\right)$$

$$\vdots$$

$$x_n = -\left(\frac{n_1}{n_n}x_1 + \frac{n_2}{n_n}x_2 + \cdots + \frac{n_{n-1}}{n_n}x_{n-1} + \frac{n_o}{n_n}\right)$$

2. Designate a summing amplifier to combine inputs to satisfy these equations for each unknown.

3. Interconnect these amplifiers in accordance with these equations.   The appropriate coefficients are determined by the input and feedback resistances of the summing amplifiers.

4. Add the constant terms a_o/a_1, b_o/b_2, etc., into each summing amplifier by appropriate d-c voltages.

This procedure is illustrated by the following example.

Example 19.2

Determine the analog required to solve the following set of equations.

$$\begin{aligned}
2x_1 + 2x_2 - 8 &= 0 \\
x_1 + 4x_2 + 8 &= 0
\end{aligned}$$

Solution

1. Solve the equations for expressions for x_1 and x_2.
From the first equation,

$$x_1 = -x_2 + 4$$

From the second equation,

$$x_2 = -\tfrac{1}{4}x_1 - 2$$

---

[1] For additional information and procedures for solving equation forms discussed in this section and additional forms such as nonlinear differential equations and equations with variable coefficients not discussed in the section, the reader is referred to: Korn and Korn, *op. cit.*, Project Cyclone Symposium I on REAC Techniques, March 15--16, 1951, and Project Cyclone Symposium II on REAC Techniques, April 28--May 2, 1952, sponsored by the U.S. Navy.

2. Connect summing amplifiers for x_1 and x_2 in accordance with these equations. This is shown in Fig. 19.11.

3. Add the constant terms. These are also shown in the figure.

The d-c voltages could have been unity, in which case the resistance ratios of the summing amplifiers would have to be adjusted to form the proper sum. As shown, the resistance ratios are all unity.

Linear Differential Equations with Constant Coefficients. Such equations are represented by the following general form:

$$\frac{d^n x}{dt^n} + a_1 \frac{d^{n-1} x}{dt^{n-1}} + a_2 \frac{d^{n-2} x}{dt^{n-2}} + \cdots + a_n x + a_o = 0 \qquad (19.31)$$

or in operational form (see Sec. 23.6)

$$s^n x + a_1 s^{n-1} x + a_2 s^{n-2} x + \cdots + a_n x + a_o = 0 \qquad (19.32)$$
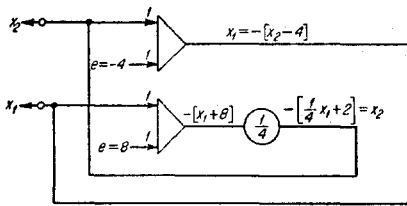
FIG. 19.11. Figure for Example 19.2

where s represents d/dt.

Equations of this form can, in theory, be instrumented by successive differentiations and additions. In practice, however, differentiation is usually accomplished implicitly by integration to avoid noise amplification and other problems inherent in differentiation systems. To solve equations such as (19.32) by successive integrations, the following procedure is suggested.

1. Equate the highest-order derivative term to the remainder of the equation. This results in the form:

$$s^n x = -(a_1 s^{n-1} x + a_2 s^{n-2} x + \cdots + a_n x + a_o) \qquad (19.33)$$

2. Designate n integrators to successively integrate $s^n x$, yielding lower-order derivatives $s^{n-1} x$, $s^{n-2} x$, etc., until a term corresponding to x is obtained.

3. Designate a summing amplifier to combine the outputs of the successive integrators to form the sum of Eq. (19.33), using inverting amplifiers to reverse the algebraic signs of terms as required. Note that the output of each integration is multiplied by $1/RC$ where RC is the time constant of the electronic integrator used to perform the integration. Note also that each integration reverses the algebraic sign. In general, approximately half as many inverting amplifiers as integrators are required.

4. The output of the summing amplifier is fed back with the proper algebraic sign into the first integrator, thereby providing the required input for the successive integrations.

5. The solution is available from the output of the last integrator. There will usually be more than one configuration which satisfactorily solves the equation.

This procedure is illustrated by the following example.

Example 19.3

Obtain an analog configuration to solve the following linear differential equation:

$$\frac{d^2 x}{dt^2} + a \frac{dx}{dt} + bx = 0$$

The initial conditions are that $x = c$ at $t = 0$ and $dx/dt = 0$ at $t = 0$.

Solution

1. Express the equation in operational form and equate the highest-order derivative term to the remainder of the equation.

$$s^2x = -(asx + bx)$$

2. Designate a number of integrators corresponding to the order of the equation and interconnect them so as to produce a zero-order differential term. The integrators and their interconnections are shown in Fig. 19.12. The initial condition for dx/dt is set into the summing integrator producing dx/dt, and similarly the initial condition for x is set into the integrator producing x. The solution of the equation is taken at the point where the zero-order differential appears, yielding a value for x.

Simultaneous Linear Differential Equations with Constant Coefficients. The previous techniques can be used to solve sets of linear differential equations. Successive integrations are used for each equation as required. Appropriate interconnections are made between integrator chains to provide the required summation for each equation. This is illustrated by the following example.

FIG. 19.12. Figure for Example 19.3.

Example 19.4

Determine the analog representation of the mechanical system shown in Fig. 19.13a. The masses m_1 and m_2 are restrained by springs of stiffness k_1, k_{12}, and k_2. The dashpots have damping coefficients f_1 and f_2. A driving force $F(t)$ is impressed on mass m_1. The quantities x_1 and x_2 are the displacements of m_1 and m_2, respectively.

Solution

1. Write the equations of motion.
The equation of motion of m_1 is

$$m_1 \frac{d^2x_1}{dt^2} + f_1 \frac{dx_1}{dt} + k_1x + k_{12}(x_1 - x_2) = F(t)$$

The equation of motion of m_2 is

$$m_2 \frac{d^2x_2}{dt^2} + f_2 \frac{dx_2}{dt} + k_2x_2 + k_{12}(x_2 - x_1) = 0$$

Assume that the initial conditions are

$$\left.\begin{array}{c} x_2 = \dfrac{dx_1}{dt} = 0 \\[2mm] x_1 = X_o \\[2mm] \dfrac{dx_2}{dt} = V_o \end{array}\right\} \text{ for } t = 0$$

2. Express each equation in operational form and solve for the highest-order derivative.

$$s^2x_1 = -\frac{f_1}{m_1} sx_1 - \frac{k_1}{m_1} x_1 - \frac{k_{12}}{m_1}(x_1 - x_2) + \frac{F(t)}{m_1}$$

$$s^2x_2 = -\frac{f_2}{m_2} sx_2 - \frac{k_2}{m_2} x_2 - \frac{k_{12}}{m_2}(x_2 - x_1)$$

3. Instrument each equation, designating the required inputs obtained from the solution of the other equation.   Figure 19.13b is the required diagram of the analog system.   Note that x_2 is used in the solution of x_1, and x_1 is used in the solution of x_2.

It should be noted that all the required interconnections are not shown in Fig. 19.13b. This is to avoid confusion in relatively complicated diagrams of analog systems.

(a) TWO MASS MECHANICAL SYSTEM

$$s^2 x_1 = -\frac{f_1}{m_1} s x_1 - \frac{k_1}{m_1} x_1 - \frac{k_{12}}{m_1} \left[x_1 - x_2 \right] + \frac{F(t)}{m_1}$$

$$s^2 x_2 = -\frac{f_2}{m_2} s x_2 - \frac{k_2}{m_2} x_2 - \frac{k_{12}}{m_2} \left[x_2 - x_1 \right]$$

(b) ANALOG FOR THE MECHANICAL SYSTEM
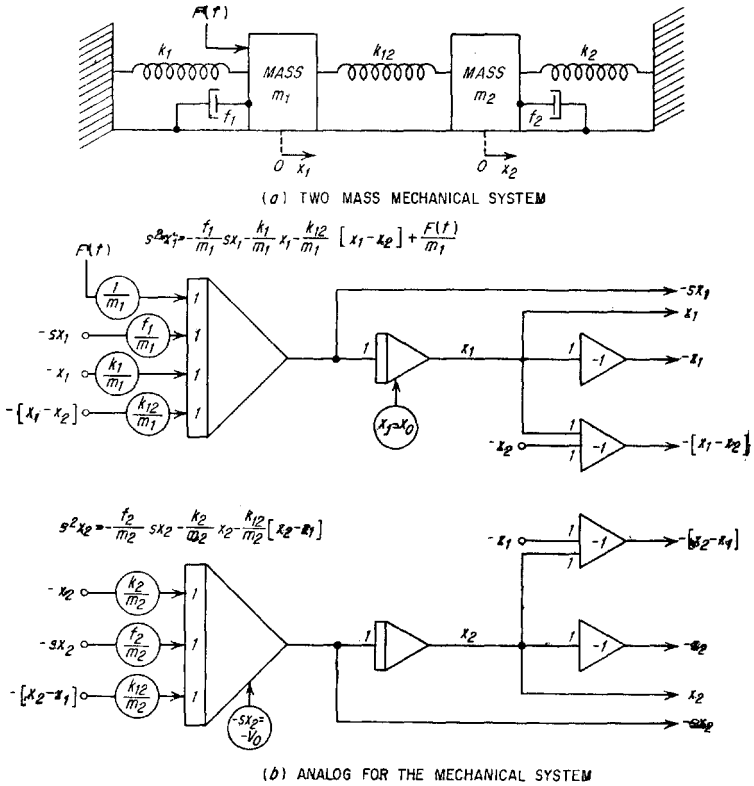
FIG. 19.13. Figure for Example 19.4.

19.4. Servomechanism Design Principles[1]

19.4a. General Characteristics.   A servomechanism is a power-amplifying device in which the element driving the output is actuated by a function of the difference between the input and the output.   Servomechanisms are used in a wide variety of applications such as remote control of physical apparatus, remote data transmission, power amplification, electrical to mechanical data transformation, and as analog computation elements.   The primary function of a servomechanism in these applications is to control automatically a given quantity or process in accordance with a given command.   Although the physical elements used in a servo system vary greatly with the application, the mathematical treatment required is similar for all applications in accordance with the basic principles of feedback systems.[1]   The type of servomechanism discussed in this section is the continuous-control type wherein a continuous action takes place to reduce the output error.

[1] In order to simplify the material presented in the remainder of this section, the fundamentals of feedback systems which form a necessary background to any discussion of servomechanisms have been omitted.   Readers lacking this background are referred to Sec. 18.

Although this discussion is limited to linear systems, the material is useful in predicting the performance of systems having certain nonlinear characteristics. Most servo systems contain elements that are to some extent nonlinear, but a satisfactory analysis of these systems can be accomplished for most purposes by the use of linear theory.

Description of a General Feedback Control System. A typical closed-loop control system is illustrated in Fig. 19.14. The important signals and elements of the system are:

Input signal θ_i. This is the input quantity controlling the output of the servo system. It is the independent variable of the system.

Output signal θ_o. This is the quantity being controlled. It is the dependent variable of the system.

(a) CONTROL SYSTEM ELEMENTS

(b) CONTROL SYSTEM TRANSFER FUNCTIONS
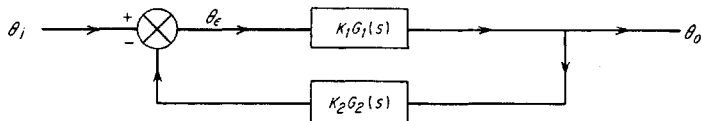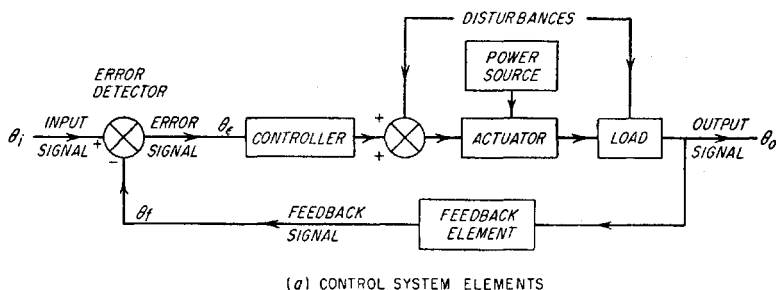
FIG. 19.14. General feedback control system.

Feedback signal θ_f. This is the signal fed back to be compared with the input signal. It is equal to θ_o modified by the transfer function of the feedback element.

Error signal θ_e. This is the difference between the input and feedback signals $(\theta_i - \theta_f)$.

Controller. This is the device which converts the error signal into a form suitable for controlling the actuator in the desired manner. In addition, the controller may have a transfer function which modifies the error signal to improve the system stability and accuracy or to obtain other desirable over-all characteristics.

Actuator. This is a source of mechanical power such as a motor, hydraulic piston, etc.

Load. This consists of the physical equipment causing inertial and/or frictional loads on the actuator. These loads can vary with the output of the servo or can be independent of the output. External factors such as wind loading and aerodynamic load are also a part of the total servo load.

Feedback element. This is an element used to change the form of the output signal so that it can be satisfactorily inserted into the error detector. In addition, the feedback element may have a transfer function which modifies the output signal to improve the over-all system stability.

Disturbances. In addition to the desired input signal θ_i there will, in general, be other signals unavoidably introduced into the servo loop which cause undesired

perturbations in the servo output. The performance of a servo is based upon its ability to respond properly to the desired input signal while minimizing the effects of these disturbing signals. Therefore, these signals must be considered as additional inputs to the system, and the response of the servo output to these disturbances must be determined. Various types of disturbances include:

1. Input uncertainties. These represent differences between the correct input and the actual input to the servo. Noise in the input signal is an example of such a disturbance.

2. Output disturbances. These are due to "stiction" (static friction), binding of gears and bearings, backlash, and nonrigid mechanical elements.

3. Internal errors. These may consist of errors such as nonlinearities of the various components; drift; variations of characteristics of elements with time, temperature, or pressure; backlash; hysteresis; friction; tube noise; unwanted electrostatic or magnetic coupling; deflections of mechanical components; and compressability of hydraulic fluids, lines, and control elements.

The detailed requirements and characteristics of servo-system elements are discussed in Sec. 19.5. Additional definitions of terms used in the analysis of servo systems involving the signals defined above and the transfer functions of Fig. 19.14*b* are defined as follows.

Forward Transfer Function $K_1G_1(s)$. This is the transfer function formed by the product of the transfer functions of the controller, actuator, and load. This is, in general, a term containing a gain factor K_1 independent of frequency and a frequency-dependent function $G_1(s)$. The forward transfer function is defined by

$$\frac{\theta_o(s)}{\theta_\epsilon(s)} = K_1G_1(s) \qquad (19.34)$$

Feedback Transfer Function $K_2G_2(s)$. This is the transfer function of the feedback element. It is defined as

$$\frac{\theta_f(s)}{\theta_o(s)} = K_2G_2(s) \qquad (19.35)$$

Open-loop Transfer Function. This is the product of the forward and feedback transfer functions and is defined by

$$\frac{\theta_f(s)}{\theta_\epsilon(s)} = K_1K_2G_1(s)G_2(s) \qquad (19.36)$$

Error Transfer Function.[1] This is defined as

$$\frac{\theta_\epsilon(s)}{\theta_i(s)} = \frac{1}{1 + K_1K_2G_1(s)G_2(s)} \qquad (19.37)$$

System or Closed-loop Transfer Function.[1] This is defined as

$$\frac{\theta_o(s)}{\theta_i(s)} = \frac{K_1G_1(s)}{1 + K_1K_2G_1(s)G_2(s)} \qquad (19.38)$$

19.4b. Analysis of a Second-order Servo System. The predominant characteristics of many servo systems can be described by a second-order differential equation. Servos which can be so described are called second-order systems. The results of the analysis of a particular second-order system are generally useful in that they can

---

[1] The sign of the second term in the denominator depends upon the manner in which the input and feedback signals are combined in the error detector to produce the error signal. When the inputs are added, a negative sign results (see Sec. 18). When the inputs are subtracted as shown in Fig. 19.4, a positive sign results as shown in these equations.

be applied to any other second-order system. For example, consider a position servo intended to cause the position of the output shaft θ_o to maintain correspondence with the input shaft θ_i, as shown in Fig. 19.15. The difference between the input θ_i and output θ_o is the error θ_ϵ. The input to the system is an angular displacement θ_i. The synchro generator converts this mechanical angle to an equivalent electrical signal. This electrical signal and the output mechanical angular displacement are combined in the synchro control transformer to produce an output voltage V_ϵ which is proportional to the difference between the input angle θ_i and the output angle θ_o. This voltage is proportional to θ_ϵ and is used as the input to the amplifier; the amplifier in turn supplies excitation to a servomotor. The output torque T_{mo} of the motor is assumed to be proportional to the motor input voltage. The moving parts of the system are considered to have a total inertia J at the output shaft and *viscous damping* f. The load torque on the output shaft is T_L.
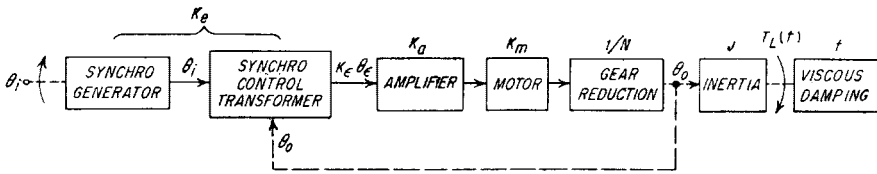
FIG. 19.15. Second-order position servomechanism.

For small displacement errors, the voltage output of the control transformer is given approximately by

$$V_\epsilon(t) = K_\epsilon \theta_\epsilon(t) \tag{19.39}$$

where K_ϵ = gain of synchro pair, volts/unit angle of error

Assuming a field-controlled d-c motor and neglecting any time delays inherent in the build-up of motor torque with applied voltage, the torque developed by the motor is

$$T_{mo}(t) = K_a K_m V_\epsilon(t) \tag{19.40}$$

where K_a = amplifier gain, volts/volt

K_m = motor characteristic, units of torque/volt

The torque $T_A(t)$ applied to the load through the gear reduction is

$$T_A(t) = N T_{mo}(t) = N K_\epsilon K_a K_m \theta_\epsilon(t) = K \theta_\epsilon(t) \tag{19.41}$$

where N = gear ratio $(1/N$ = gear reduction ratio$)$

$K = N K_\epsilon K_a K_m$, units of torque/unit angle of error

The equation of motion for the system is

$$J \frac{d^2\theta_o(t)}{dt^2} + f \frac{d\theta_o(t)}{dt} = T_A(t) - T_L(t) \tag{19.42}$$

or

$$J \frac{d^2\theta_o(t)}{dt^2} + f \frac{d\theta_o(t)}{dt} = K \theta_\epsilon(t) - T_L(t) \tag{19.43}$$

The error θ_ϵ is $(\theta_i - \theta_o)$, therefore

$$J \frac{d^2\theta_\epsilon(t)}{dt^2} + f \frac{d\theta_\epsilon(t)}{dt} + K \theta_\epsilon(t) = J \frac{d^2\theta_i(t)}{dt^2} + f \frac{d\theta_i(t)}{dt} + T_L(t) \tag{19.44}$$

and

$$J \frac{d^2\theta_o(t)}{dt^2} + f \frac{d\theta_o(t)}{dt} + K \theta_o(t) = K \theta_i(t) - T_L(t) \tag{19.45}$$

Application of the Laplace transformation (see Sec. 23.6) to Eqs. (19.44) and (19.45) results in

$$(Js^2 + fs + K)\theta_\epsilon(s) = (Js^2 + fs)\theta_i(s) + T_L(s) \qquad (19.46)$$

and $\qquad (Js^2 + fs + K)\theta_o(s) = K\theta_i(s) - T_L(s) \qquad (19.47)$

The above equations can be rewritten as

$$\theta_\epsilon(s) = \frac{s\left(s + \dfrac{f}{J}\right)\theta_i(s) + \dfrac{T_L(s)}{J}}{s^2 + (f/J)s + K/J} \qquad (19.48)$$

$$\theta_o(s) = \frac{(K/J)\theta_i(s) - T_L(s)/J}{s^2 + (f/J)s + K/J} \qquad (19.49)$$

The transient and stability characteristics of any system can be determined by the denominator of the system transfer function (see Sec. 18.4b). The characteristic equation of the system is

$$s^2 + \frac{f}{J}s + \frac{K}{J} = 0 \qquad (19.50)$$

This is obtained by setting the denominator of the expression for $\theta_o(s)$ [i.e., Eq. (19.49)] equal to zero. The roots r_1 and r_2 of this equation are

$$r_1, r_2 = \frac{-f \pm \sqrt{f^2 - 4JK}}{2J} \qquad (19.51)$$

For the system to be stable, the roots r_1 and r_2 must be either (1) unequal negative real, (2) equal negative real, or (3) complex conjugate with negative real parts.

To simplify the expressions, it is convenient to introduce the parameters *undamped natural frequency* ω_n and the *damping ratio* ζ. The quantity ω_n is the frequency of oscillation of the system with no damping, defined by

$$\omega_n = \sqrt{K/J} \qquad \text{radians/sec} \qquad (19.52)$$

In this example, the *attenuation factor* α is equal to $f/2J$ which is equal to zero when f equals zero. When α equals zero, the roots of the characteristic equation are

$$r_1, r_2 = \pm j \sqrt{K/J} = \pm j\omega_n \qquad (19.53)$$

The parameter ζ is the ratio of actual damping to critical damping. Critical damping occurs when $r_1 = r_2$. The value of damping f for critical damping is

$$f_c = 2 \sqrt{KJ} \qquad (19.54)$$

The damping ratio is then

$$\zeta = \frac{\text{actual damping}}{\text{critical damping}} = \frac{f}{2 \sqrt{KJ}}$$

The characteristic equation in terms of ω_n and ζ is

$$s^2 + 2\zeta\omega_n s + \omega_n^2 = 0 \qquad (19.55)$$

where $\alpha = \zeta\omega_n$ and the roots are

$$r_1, r_2 = -\zeta\omega_n \pm j\omega_n \sqrt{1 - \zeta^2} \qquad (19.56)$$

The damped frequency, which is the actual frequency of damped oscillation, is ω_o where

$$\omega_o = \omega_n \sqrt{1 - \zeta^2} \qquad (19.57)$$

The positions of the roots r_1 and r_2 in the s plane are shown in Fig. 19.16.   Note that the locus of roots for constant damping ratio is a straight line passing through the origin in the complex plane and the locus of roots for constant ω_n is a circle with its center at the origin.   For critical damping ($\zeta = 1$), the roots of the characteristic equation occur at the intersection of this circle with the negative real axis. Equations (19.48) and (19.49) can be rewritten as follows:

$$\theta_\epsilon(s) =$$

$$\frac{s(s + 2\zeta\omega_n)\theta_i(s) + T_L(s)/J}{s^2 + 2\zeta\omega_n s + \omega_n{}^2} \quad (19.58)$$

$$\theta_o(s) = \frac{\omega_n{}^2\theta_i(s) - T_L(s)/J}{s^2 + 2\zeta\omega_n s + \omega_n{}^2} \quad (19.59)$$

FIG. 19.16. Complex plane plot of roots of second-order system.

These equations are useful in determining the response for specific input functions.

Effect of a Step Displacement Input.   If a step displacement ϕ is given to $\theta_i(t)$,

$$\mathcal{L}[\theta_i(t)] = \theta_i(s) = \frac{\phi}{s} \quad (19.60)$$

For this condition and assuming $T_L(s) = 0$, Eq. (19.58) becomes

$$\theta_\epsilon(s) = \frac{(s + 2\zeta\omega_n)\phi}{s^2 + 2\zeta\omega_n s + \omega_n{}^2} \quad (19.61)$$

The solution of the error as a function of time, $\theta_\epsilon(t)$, is made by determining the inverse transform according to the principles outlined in Sec. 23.6.   The result assuming complex conjugate roots is given by

$$\theta_\epsilon(t) = \frac{\phi}{\sqrt{1 - \zeta^2}} e^{-\zeta\omega_n t} \sin(\omega_n \sqrt{1 - \zeta^2}\, t + \psi) \quad (19.62)$$

where $\psi = \tan^{-1}\dfrac{\sqrt{1 - \zeta^2}}{\zeta}$

Note that as t approaches infinity the steady-state error becomes zero.   Figure 19.17 gives the error as a function of time for various values of damping ratio.

Effect of a Step Input Velocity.   If a step velocity ω_i (the input is instantaneously changed from zero velocity to the value ω_i) is given to $\theta_i(t)$, with all other initial conditions equal to zero,

$$\mathcal{L}[\theta_i(t)] = \theta_i(s) = \frac{\omega_i}{s^2} \quad (19.63)$$

Inserting this in Eq. (19.58) and assuming $T_L(s)$ to be equal to zero,

$$\theta_\epsilon(s) = \frac{(s + 2\zeta\omega_n)\omega_i}{s(s^2 + 2\zeta\omega_n s + \omega_n{}^2)} \quad (19.64)$$

The final value theorem[1] is applied to Eq. (19.64) to find the steady-state error $(\theta_\epsilon)_{ss}$. The result is

$$(\theta_\epsilon)_{ss} = \frac{2\zeta\omega_i}{\omega_n} = \frac{f\omega_i}{K} \quad (19.65)$$

[1] See Sec. 23.6.

Equation (19.65) states that the steady-state error for a step-input velocity is proportional to the damping and inversely proportional to the system gain. The error is

$$\theta_\epsilon(t) = (\theta_\epsilon)_{ss} \left[1 - \frac{e^{-\zeta \omega_n t}}{2\zeta \sqrt{1 - \zeta^2}} \sin \left(\omega_n \sqrt{1 - \zeta^2}\, t + \psi' \right) \right] \tag{19.66}$$

where $\psi' = \tan^{-1} \dfrac{2\zeta \sqrt{1 - \zeta^2}}{2\zeta^2 - 1}$

Figure 19.18 gives the ratio between the actual error and the steady-state error as a function of $\omega_n t$.
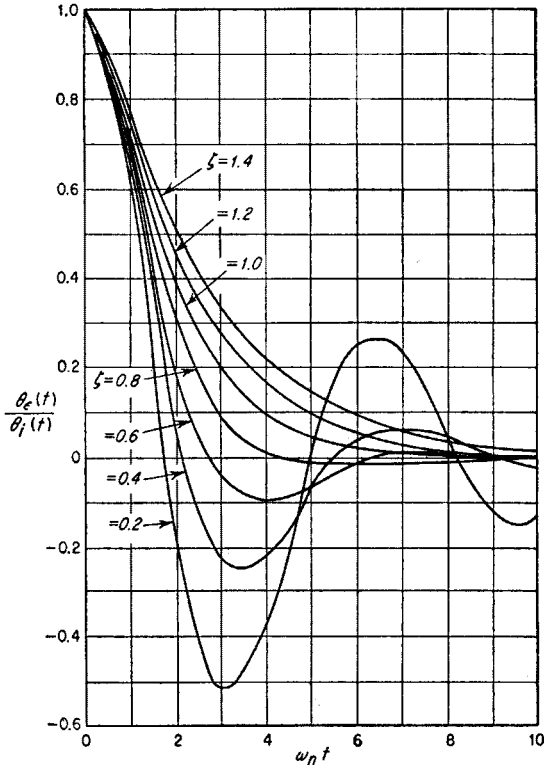
FIG. 19.17. The error response of a second-order system subjected to a step input.

Effect of a Load Disturbance. If a step load disturbance T_L is impressed on the output shaft with zero input θ_i, then

$$\mathcal{L}[T_L(t)] = T_L(s) = \frac{T_L}{s} \tag{19.67}$$

Inserting in Eq. (19.58):

$$\theta_\epsilon(s) = \frac{T_L/J}{s(s^2 + 2\zeta \omega_n s + \omega_n^2)} \tag{19.68}$$

and the steady-state error is

$$(\theta_\epsilon)_{ss} = \frac{T_L}{J \omega_n^2} = \frac{T_L}{K} \tag{19.69}$$

Equation (19.69) states that the steady-state error for an applied load torque is inversely proportional to the system gain. The error as a function of time for this input is

$$\theta_\epsilon(t) = (\theta_\epsilon)_{ss} \left[1 - \frac{e^{-\zeta\omega_n t}}{\sqrt{1 - \zeta^2}} \sin \left(\sqrt{1 - \zeta^2}\, \omega_n t + \psi\right) \right] \qquad (19.70)$$

where $\psi = \tan^{-1} \dfrac{\sqrt{1 - \zeta^2}}{\zeta}$

This error response is plotted for various values of ζ in Fig. 19.19.
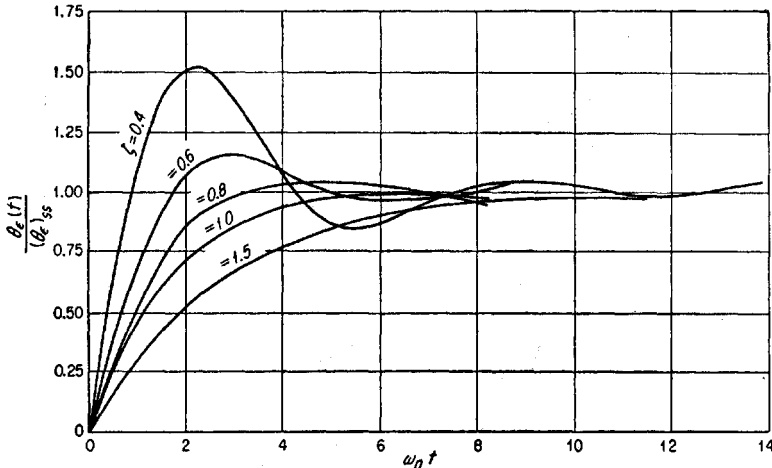
FIG. 19.18. Dimensionless transient curves of the error response of a second-order system when subjected to a step velocity input.

Transient Response Curves. The following general remarks may be made from an inspection of the dimensionless curves of Figs. 19.17 to 19.19.

1. The system eventually reaches a steady state for all values of ζ.

2. The steady-state value is approached but never exceeded for $\zeta = 1$, that is, there is no overshoot.

3. If $\zeta \leq 1$, the system oscillates about the steady-state value. The initial overshoot is small, and oscillations die out quickly for values of ζ near unity, but the overshoot increases and the oscillations continue for longer periods for values of ζ much less than unity.

4. The rise time of the system response is shortest for ζ much less than unity, and the rise time is longest for ζ much greater than unity. Some overshoot is usually accepted as the penalty for faster rise time. Most practical servos employ values of damping ratios from approximately 0.4 to 0.8.

In servomechanism work, speed of response is defined as the time required for the system output to approach within some arbitrary percentage of its steady-state value. It depends only on the factor $e^{-\zeta\omega_n t}$. This is illustrated in Fig. 18.12, where $t = C$. By this definition the time constant T of the system is

$$T = \frac{1}{\alpha} = \frac{1}{\zeta\omega_n} \qquad (19.71)$$

The response is within approximately 37 per cent of the final value for $t \geq T$ and is within approximately 2 per cent of the final value for $t \geq 4T$. A servomechanism has a high speed of response for large values of the product $\zeta\omega_n$. It is desirable,

therefore, to make $\zeta\omega_n$ as large as possible to provide rapid system response.   However, regardless of the value of ω_n, the damping ratio ζ is usually restricted to a value between 0.4 and 0.8 to achieve the best compromise between system rise time and overshoot.   Therefore, ω_n is the parameter which indicates the speed of response of a correctly damped system.

In an actual servo the speed of response can be increased only by increasing the damping f or decreasing the inertia J.   If the damping ratio is to be maintained constant, a change in gain K is necessary when a change is made in either f or J.
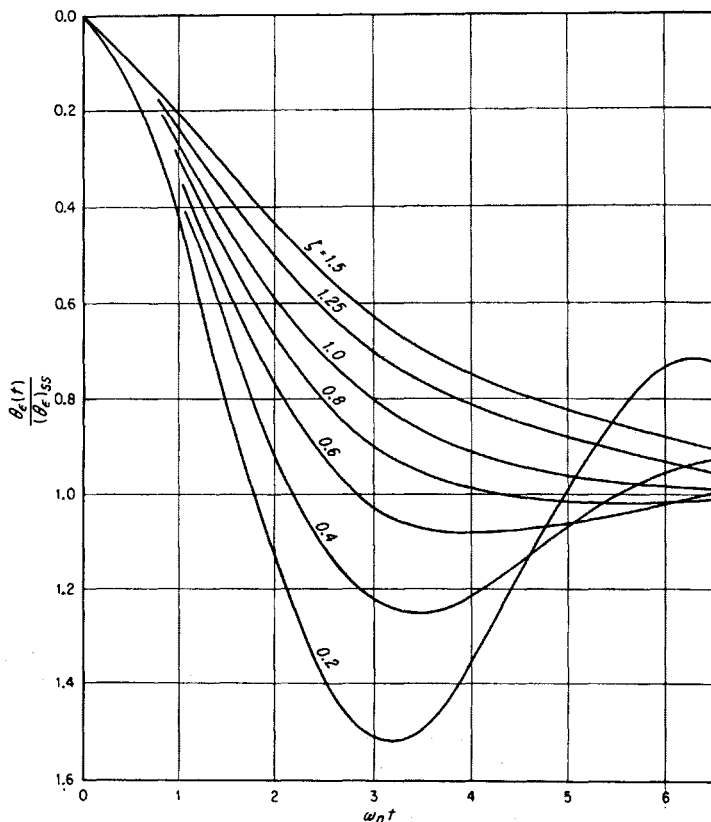
FIG. 19.19. Dimensionless transient curves of the error response of a second-order system subjected to a suddenly applied load torque.

For a specified f/J ratio, optimum response (rise time and overshoot) is obtained by increasing the loop gain K (and consequently raising ω_n and reducing ζ) until the minimum tolerable value of ζ is obtained.   An increase in viscous damping f allows an increase in system gain K without a corresponding decrease in the damping ratio ζ. This increase in viscous damping results in a loss of energy dissipated as heat in the viscous drag.   Methods other than viscous damping such as the insertion of phase-compensating networks in the loop are used for stabilization purposes.   These networks allow increased gain and speed of response while maintaining the damping ratio within the desired limits.   See Sec. 19.7c for the design and application of these stabilizing networks.

Frequency Response of a Second-order System. The difficulty of analyzing closed-loop systems by the transient method of analysis soon becomes apparent with systems higher than second order. The frequency-response method of analysis and synthesis
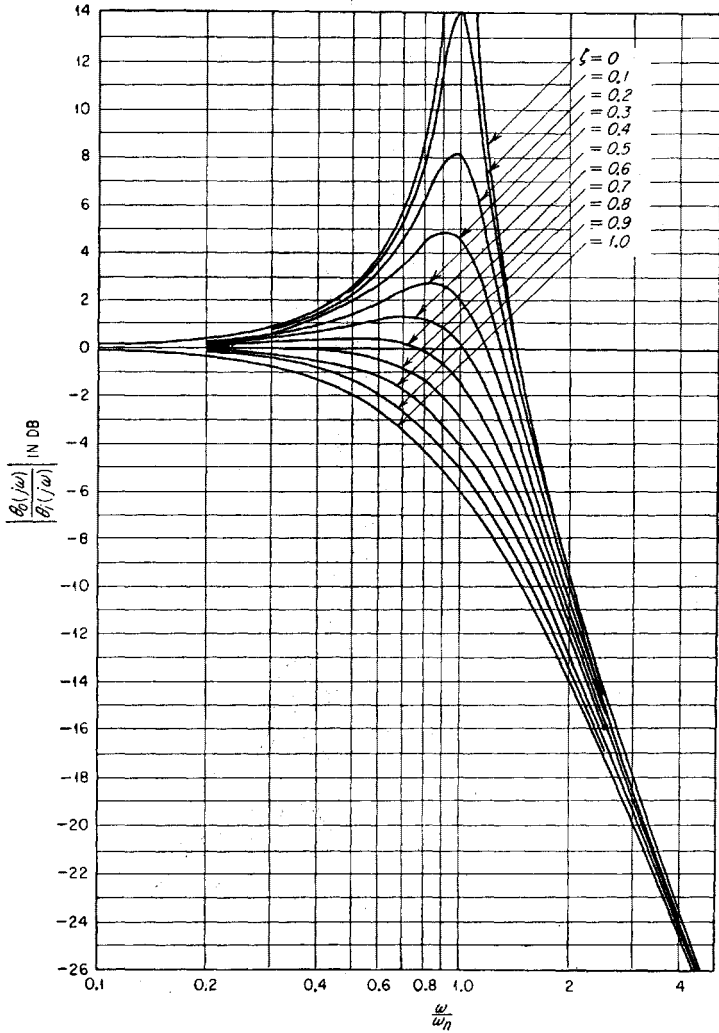
Fig. 19.20. Amplitude response of a second-order system.

overcomes many of the shortcomings of the transient analysis method but does not give factors such as the time response of the system directly.

The steady-state frequency response of the system in Fig. 19.15 is obtained by substituting $j\omega$ for s in Eq. (19.59). The result for T_L equal to zero is

$$\frac{\theta_o}{\theta_i}(j\omega) = \frac{\omega_n{}^2}{(\omega_n{}^2 - \omega^2) + j(2\zeta\omega_n\omega)} \qquad (19.72)$$

Equation (19.72) can be expressed in terms of an amplitude and a phase function as

$$M = \frac{\theta_o}{\theta_i}(j\omega) = \frac{\omega_n^2}{\sqrt{(\omega_n^2 - \omega^2)^2 + (2\zeta\omega_n\omega)^2}} \qquad (19.73a)$$

and

$$\phi(j\omega) = -\tan^{-1}\frac{2\zeta\omega_n\omega}{\omega_n^2 - \omega^2} \qquad (19.73b)$$

where M is called the *magnification ratio*. M and ϕ are plotted in Figs. 19.20 and 19.21, respectively, as functions of the ratio ω/ω_n. Figure 19.22 is a plot of the output
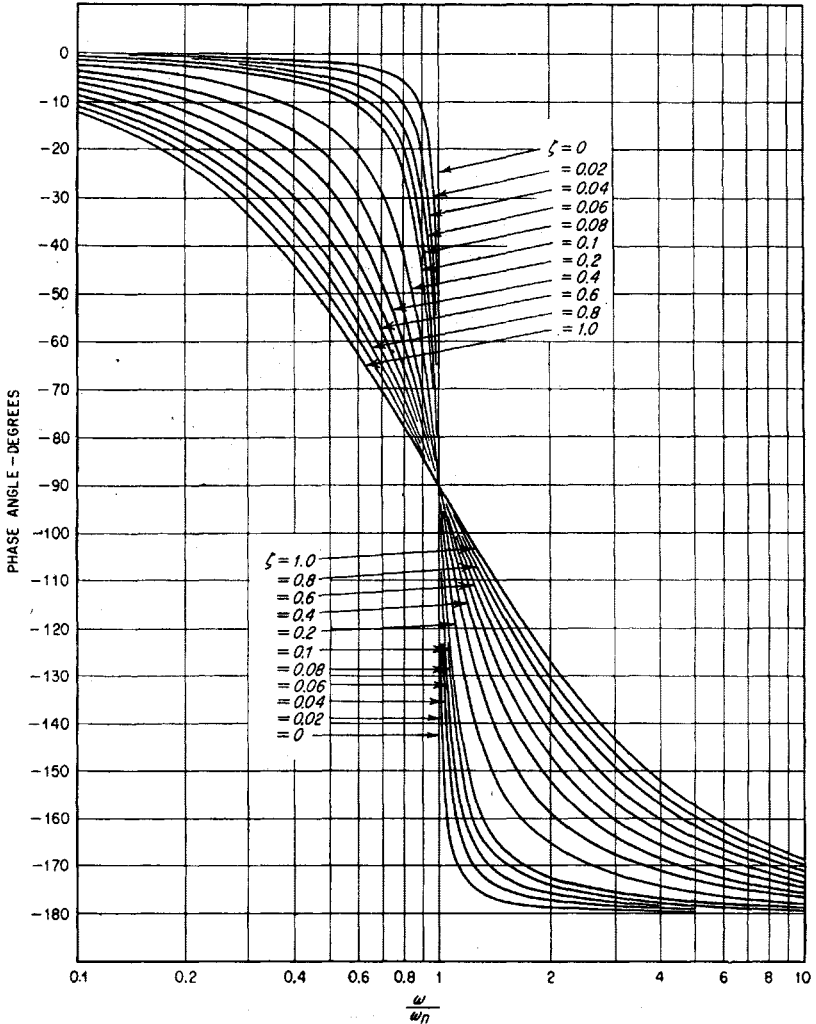
Fig. 19.21. Phase response of a second-order system.

of a second-order system as a function of time for a step displacement input. A comparison with Fig. 19.20 shows a correspondence between the peak overshoot of the output and the peak magnitude M_p of the frequency response. It can be seen that both the peak overshoot in the transient response and the peak magnitude of the

frequency response vary inversely with damping ratio. Figure 19.23 is a plot of M_p versus the damping ratio ζ. The frequency ω_p associated with M_p is given by

$$\omega_p = \omega_n \sqrt{1 - 2\zeta^2}$$

The majority of higher-order servomechanisms behave in a manner similar to second-order systems. Even though the transient response of a higher-order system contains damped sinusoids of several frequencies, one frequency usually predominates. The peak magnitude M_p of the frequency response is a measure of the damping of the predominant oscillation of the transient response. For these reasons a higher-order system can usually
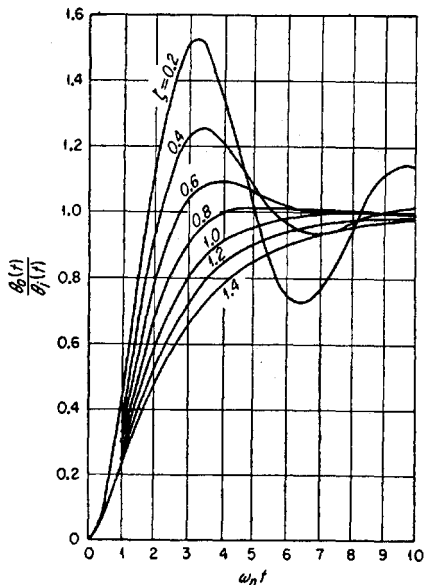
FIG. 19.22. Dimensionless response functions of a second-order system for a unit step input.
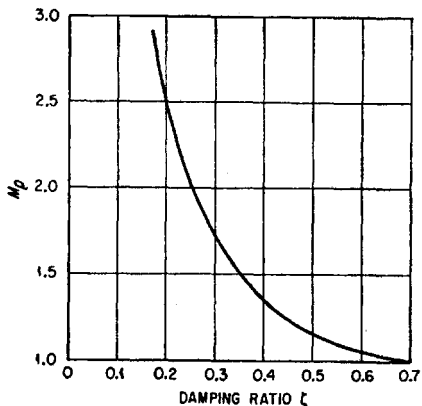
FIG. 19.23. Peak magnification ratio M_p as a function of damping ratio for a second-order system.

be analyzed by using the response curves of a second-order system. In addition, the frequency band over which the amplitude response has substantially a constant magnitude is a measure of the speed of response; i.e., a high speed of response corresponds to a high upper cutoff frequency of the amplitude response.

19.5. Servomechanism Elements. The following paragraphs discuss representative servo elements which might be used in the design of an electromechanical servo system. There are many possible variations. Factors to be considered in the final selection are the physical form of the data to be used in the input and output of each element, the accuracy requirements of each element, response time, desirable and undesirable effects of the element transfer function on the servo performance, and the suitability of such characteristics of the element as cost, availability, size, and variations in performance with temperature and wear.

19.5a. Error Detectors. The error signal θ_ϵ of Fig. 19.14 is the difference between the input signal θ_i and the feedback signal θ_f. The error detector is the servo element which obtains this difference and is indicated by the symbol shown in the figure. The + and − signs adjacent to the error detector indicate the variation in sign which occurs to each quantity as it goes through the error detector. As shown, $+\theta_i$ and $-\theta_f$ produce an output of $(+\theta_i - \theta_f)$. Several commonly used error detectors are shown in Fig. 19.24. The differential amplifier (see Sec. 3.17) is also commonly used as an error detector.

(a) POTENTIOMETER BRIDGE

(b) E TRANSFORMER

(c) SYNCHRO UNITS

(d) MECHANICAL DIFFERENTIAL

(e) TACHOMETER

(f) GYROSCOPE

(g) LIGHT SPLITTER
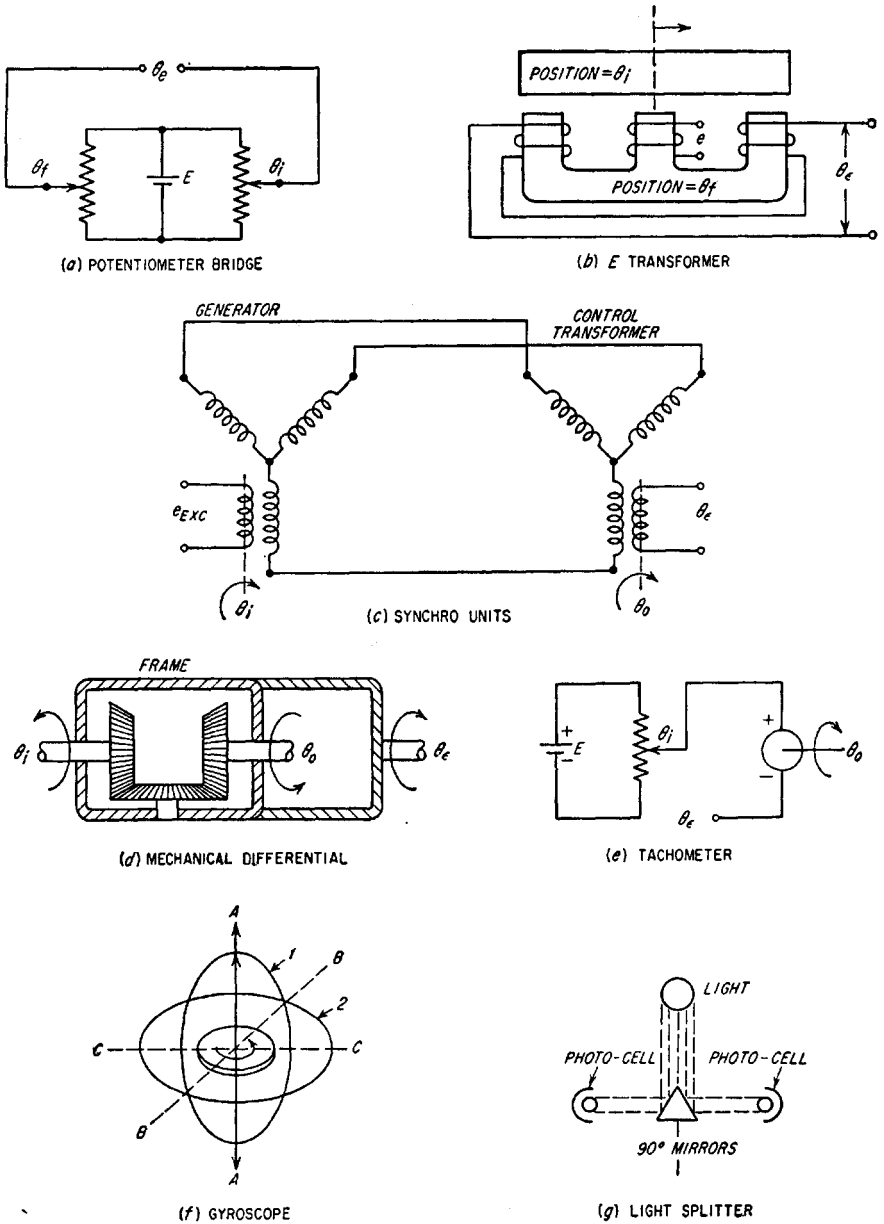
FIG. 19.24. Error detectors.

The potentiometer bridge of Fig. 19.24a is a convenient means of producing an electrical error voltage as the difference between two shaft positions θ_f and θ_i.

The E transformer of Fig. 19.24b produces an error voltage proportional to the relative displacement of the two portions of the magnetic circuit as shown. The center arm of the E section is excited by an a-c voltage. The outer windings are

similar and connected in series opposition. The error voltage is zero for a symmetrical position of the movable arm. As the movable arm is displaced either side of the symmetrical position, the reluctance of each magnetic circuit changes, producing the resultant error voltage.

The synchro error detector of Fig. 19.24c produces an electrical error signal which is a function of the difference in shaft positions of the two units.

The mechanical differential of Fig. 19.24d consists of a mechanical arrangement of gears as shown. The center gear is free to rotate, but the shaft is attached to the frame. The angular rotation of the frame is proportional to the difference in angular rotation between the other gears, assuming positive directions of rotation of θ_i and θ_o as shown in the figure.

In the tachometer error detector shown in Fig. 19.24e the error voltage is proportional to the difference between the potentiometer setting and the derivative of the output shaft position. The error voltage therefore controls the velocity of the output shaft.

The gyroscope shown diagrammatically in Fig. 19.24f can also be used as an error detector. The gyroscope is composed of a high-speed rotor mounted on gimbals (1) and (2) as shown. The mass rotates about axis A and, in the ideal case, maintains this axis in a fixed position. Gimbal (1) is free to rotate about axis B. A torque applied to gimbal (2) about axis C will cause gimbal (1) to process about axis B with an angular velocity proportional to this torque. Conversely, if gimbal (1) is rotated about axis B with a fixed angular velocity, a torque will appear acting on gimbal (2) about axis C. Angular position and rate data are measured by maintaining axis A in a fixed position and measuring angular displacements about axis B and/or C with respect to an external reference frame.

The light splitter shown in Fig. 19.24g is an optical means of producing an error signal which is a function of linear displacement along the axis between the phototubes. The outputs from the phototubes are compared in a bridge circuit. The output of the bridge circuit is the error signal and depends upon the difference in light intensity received by each phototube.

The transfer function of error detectors is equal to $K_e G_e(s)$ where K_e is the gain change and accounts for the transducer action changing the physical form of the variables. $G_e(s)$ is a function of frequency defining the dynamic properties of the error detector. It can usually be assumed equal to unity over the frequency band of interest, although this must be determined for a particular error detector in a particular application.

19.5b. Controllers. The controller is used to modify the physical form and the phase and amplitude characteristics of the error signal in order to drive the servo actuator. For small servos where the actuator power requirements are small, conventional electronic amplifiers are generally used. Magnetic amplifiers can be used for small or large power applications. The main requirements for the amplifier are that it provide sufficient power to drive the actuator and that it introduce negligible phase shift over the frequency band of interest. If a d-c amplifier is used, phase shift is usually not a serious problem. If an a-c or modulated-carrier amplifier is used to drive a two-phase induction motor, care must be taken to ensure that the phase of the carrier of the amplifier output and the reference phase applied to the motor are approximately 90° apart.

The controller usually contains whatever stabilization networks are necessary for proper operation of the complete servo. Stability analysis and the use of stabilization networks are discussed in Secs. 19.6 and 19.7. The gain of the amplifier is primarily determined by the system accuracy requirements. Variations in gain under operating conditions should be minimized so as to decrease problems of stabilization.

The choice between a-c and d-c controllers is dependent upon many factors such

as the type of signals available and actuator considerations. An a-c amplifier is sometimes preferred to avoid drift problems inherent in d-c amplification. On the other hand, a-c stabilization networks are sensitive to changes in the carrier frequency, and consequently stabilization with d-c networks is often preferred. These considerations often lead to a requirement for conversion between d-c and a-c signals in

(a) CHOPPER MODULATION

(b) DOUBLE DIODE   (c) BRIDGE

(d) TRIODE MODULATOR   (e) TRIODE DEMODULATOR

(f) SIGNAL CONVERSION CONTROLLER
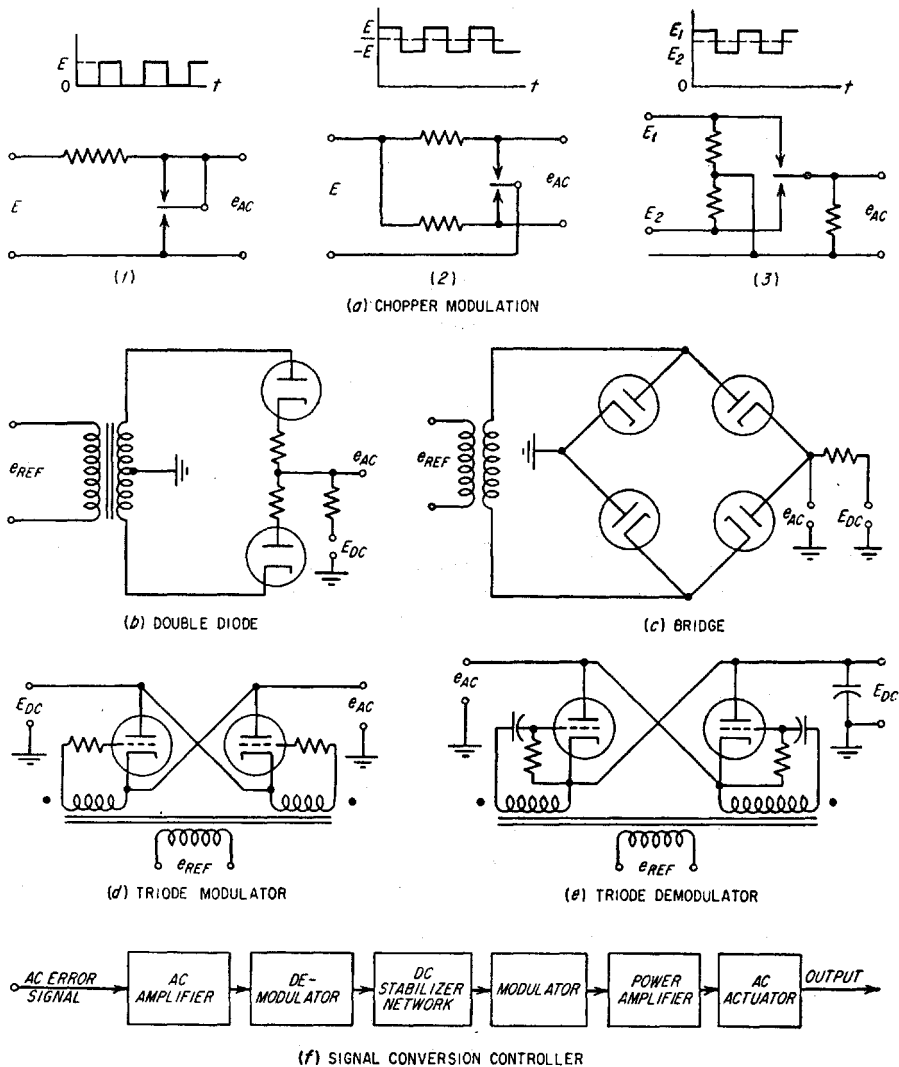
FIG. 19.25. Modulators and demodulators for signal conversion.

servo controllers. Modulators and demodulators are used to provide these conversions. When these devices are used, it is important to consider their effects on the controller transfer function.

Representative modulators and demodulators are shown in Fig. 19.25. The chopper modulators of Fig. 19.25a employ a reed vibrating at the carrier frequency. The output of the first circuit is single-ended and uses a single-ended input. A

chopper modulator can produce a double-ended output from a single-ended input as shown in the second circuit.   The third circuit has a single-ended output which is the difference between the two input signals.

The double-diode modulator is equivalent to the first chopper circuit.   Both diodes conduct simultaneously on alternate half cycles of the reference signal and effectively ground the output during conduction.   The instantaneous value of e_{ac} is equal to the input E_{dc} when the diodes are not conducting.   The difference between two input signals can be obtained by inserting the second signal between the center tap of the input transformer and ground.   The modulator can then be used as the error detector.

The bridge circuit of Fig. 19.25c is another example of a diode modulator.

The triode modulator of Fig. 19.25d is another form of an a-c–actuated switch. The grids are connected in phase.   During the positive half cycle of the reference signal on the grids of the tubes, one of the two tubes conducts (depending on the polarity of the input signal), and this results in an output signal.   During the negative half cycle of the reference signal, neither tube can conduct and the output voltage is zero.   The phase of the a-c output is a function of the input polarity, and the output amplitude is proportional to the input amplitude.

The triode demodulator shown in Fig. 19.25e is a half-wave phase-sensitive detector.   The d-c output is proportional to the a-c amplitude, and the d-c polarity is dependent upon the relative phase between the input and reference signals.   The grids are connected in phase.   If the grids and the input signal are in phase, the first tube conducts and the d-c signal is positive.   If the grids and the signal are 180° out of phase, the second tube conducts and the output signal polarity is negative. Grid-leak bias is employed so that grid conduction takes place only during the peak of the positive half cycle of the reference signal.   It is important that the a-c input and reference signals have a 0 or 180° phase relationship.

Many other demodulators can be used: For example, the choppers used for modulation can be used as demodulators by synchronizing the chopper with the a-c signal so as to connect the a-c signal to the output only on alternate half cycles.   The output, which can then be filtered, is a unipolar signal with a polarity dependent upon the relative phase between the chopper and the a-c signal.   The main requirements for a demodulator are that it (1) provide amplitude and polarity of the d-c output as a function of amplitude and phase of the a-c signal being modulated, (2) have satisfactory linearity over the range of signals to be encountered, and (3) have low d-c drift.

The use of signal conversion is shown in Fig. 19.25f.   The over-all servo controller utilizes modulation and demodulation to avoid a-c stabilization networks which may be sensitive to a-c carrier frequency changes, yet take advantage of drift-free a-c amplification.   The chopper or "carrier" frequency should be at least ten times the highest significant component of the control signal.[1]

19.5c. Actuators.   The actuator of a servomechanism can be an electric, pneumatic, or hydraulic motor, or any other device capable of converting the commands of the controller into the proper output response.   Only electric-motor servo actuators are considered here.

There are many factors to consider in the selection of the proper motor for a servo application.[2]   These factors can be classified as follows.

System.   These include requirements dictated by the application such as type of output (e.g., displacement, constant velocity, etc.), limits of acceleration, allowable time delay, allowable error, maximum speed, available power source, probable

[1] A maximum phase shift of approximately 36° can occur when a frequency ratio of 10:1 is used.   See Fig. 5.47.
[2] See, for example, R. S. Edwards, Selecting Electric Servomotors, *Machine Design*, January, 1949.

duty cycle, need for reversibility, smoothness, and system damping or stability considerations.

Load. These include requirements dictated by the load such as power required and torque required to handle the inertial and frictional characteristics of the load.

Environmental. These include factors such as temperature, humidity, and vibration.

The ability of a servomotor to accelerate the load is usually of prime importance. This depends upon the torque available from the motor and is inversely proportional to the inertia of the motor and load. Accordingly, the torque-to-inertia ratio T/J

(a) DC-SHUNT WOUND MOTOR

(b) TWO-PHASE INDUCTION MOTOR

(c) MOTOR CLUTCH ACTUATOR
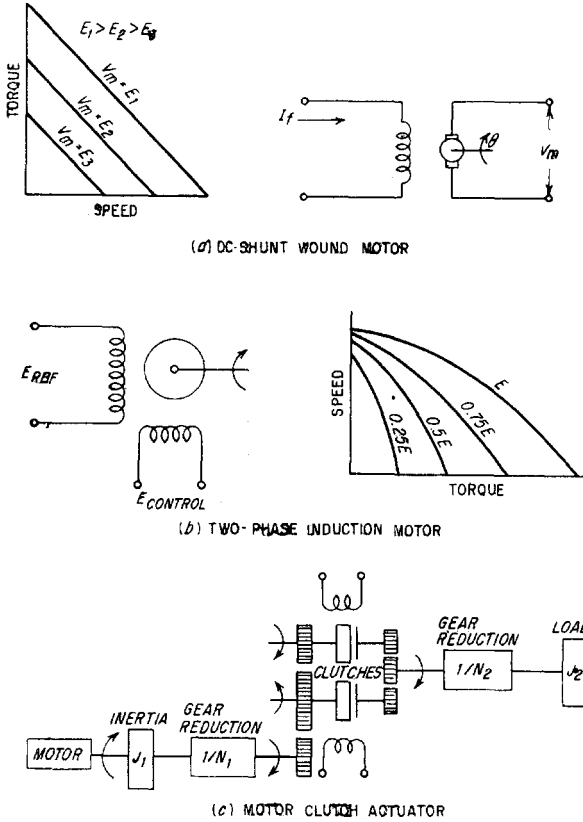
FIG. 19.26. Servomotor characteristics.

is an important figure of merit for servomotors. Under some conditions, conventional d-c or a-c motors can be used, but usually a special design is required to obtain the desired high T/J ratio for servo applications.

Servomotors can be either series or shunt-wound d-c motors or a-c induction motors. The advantages of d-c motors are less weight for the same power and higher starting torque. The advantage of an a-c motor is greater reliability due to freedom from commutation problems such as noise, wear, and failure at high altitude.

1. D-C Shunt Motor. A schematic is shown in Fig. 19.26a. A servomotor of this type can be controlled by means of either the armature or the field winding. If armature control is used, the field is excited from a constant-current source and the armature voltage is supplied from the servo controllers. In high-power applications, the armature control power is usually supplied from a separate motor generator with

the generator field controlled by the servo amplifier. For lower-power applications the armature can be supplied by thyratron or vacuum-tube circuits[1] or by magnetic amplifiers. The armature-controlled motor has the disadvantage of requiring all the motor power to be supplied by the controller. It has the advantage of a fast response time, however, because of the low armature inductance and relatively high starting and reversing torques. The torque equation for an armature controlled d-c motor is

$$T_{mo} = \frac{V_m - K_e \omega_m}{R_a} K_T \qquad (19.74)$$

where T_{mo} = motor torque
V_m = armature voltage
ω_m = motor speed
R_a = armature resistance
K_e = ratio of back emf to motor speed
K_T = ratio of torque to current

Idealized torque speed curves are shown in Fig. 19.26a as a function of terminal voltage. The reduction in torque proportional to the speed of the motor (due to back emf) is considered as viscous damping. The damping coefficient f is equal to minus the slope of the torque-speed characteristic of the motor. Motor reversal is accomplished by reversing the polarity of the armature terminal voltage.

The transfer function for armature-controlled constant-field d-c motors[2] is

$$\frac{\omega_m(s)}{V_m(s)} = \frac{1/K_e}{T_m s + 1} \qquad (19.75)$$

where T_m = motor time constant given by Eq. (19.76)

$$T_m = J/f = J \left/ \left(\frac{K_e K_T}{R_a} \right) \right. \qquad (19.76)$$

The relationship between motor shaft position θ_m and applied voltage is

$$\frac{\theta_m(s)}{V_m(s)} = \frac{1/K_e}{s(T_m s + 1)} \qquad (19.77)$$

2. Field-controlled D-C Generators. The transfer function of a field-controlled generator running at constant speed is

$$\frac{V_a(s)}{E_f(s)} = \frac{K_G}{R_f(T_f s + 1)(T_a s + 1)} \qquad (19.78)$$

where V_a = armature output voltage
K_G = output voltage per ampere of field excitation
E_f = input field excitation voltage
R_f = field resistance
T_f = field time constant = ratio of field inductance to total resistance in field circuit
T_a = armature time constant = ratio of armature inductance to total resistance in armature circuit.

In most cases, the armature time constant is much less than the field time constant and can be neglected. If the field is excited from a high-resistance source, however,

[1] See I. A. Greenwood, Jr., J. Vance Holdam, Jr., and Duncan MacRae, Jr., "Electronic Instruments," pp. 405–427, McGraw-Hill Book Co., Inc., New York, 1948.
[2] See H. Chestnut and R. W. Mayer, "Servomechanisms and Regulating System Design," pp. 113–117, John Wiley & Sons, Inc., New York, 1951.

the effective field time constant can be much less than the armature time constant. The over-all transfer function between input voltage to a generator and shaft position of a motor driven by the generator is the product of Eqs. (19.77) and (19.78).

3. D-C Series Motor. Series-wound d-c motors can be used for servo applications requiring very high starting and reversing torques and high viscous damping. They have the disadvantage of poor speed regulation with torque. For bidirectional control, split windings are required.

4. Permanent-magnet Motors. Compact and efficient d-c servomotors are constructed by using permanent magnets instead of the field windings. The characteristics of these motors are the same as shunt-wound armature-controlled motors. Very large armature currents which might demagnetize the permanent magnets must be avoided in the use of these motors.

5. A-C Induction Motors. A two-phase induction motor is shown in Fig. 19.26b. This motor utilizes control and reference windings. The control winding is excited by the controller which has an output voltage displaced in phase 90° from the reference winding. The reference winding is connected directly to the power source. The starting torque is proportional to the product of control and reference winding voltages. Representative torque-speed curves for various control voltages are shown in Fig. 19.26b. The slope of these curves at any point is the speed-to-torque ratio which is equal to the negative of the viscous damping coefficient f. The motor time constant T_m is equal to J/f. Equations (19.75) and (19.77) also give the transfer function of an a-c motor. The value of K_e for an a-c motor is the ratio of the back emf to the shaft velocity and is constant only for the case when the motor speed is sufficiently below the maximum speed.

6. Clutch Actuators.[1] A conventional motor can be used in servo applications by the use of clutches as shown in Fig. 19.26c. The clutches, which are electrically actuated, are used to couple power from a constant-speed motor to the load. By this method, energy which is stored in the inertia J_1 by a small, high-speed, continuously running motor can be transferred to the load in a short time by one of the clutches. The clutches usually have a very high torque-to-inertia ratio. The servomotor does not require a large torque-to-inertia ratio in this application, yet very fast response times are possible with this system. Bidirectional operation can be achieved through the use of two clutches and the appropriate gearing as shown in the figure. The output torque of a clutch is approximately proportional to the current of the actuating coil as given by Eq. (19.79).

$$T_c = K_c I_c \qquad (19.79)$$

where I_c = clutch coil current

T_c = torque output

K_c = proportionality constant relating clutch torques to clutch current

No damping is inherently present in a clutch mechanism. This causes the output to accelerate to the speed of the input shaft for any excitation current, assuming T_c is greater than the load torque.

19.6. Stability and Methods of Analysis. One of the primary requirements for any automatic control system is that it be stable. The controlled variable or output of a servo system must have a definite relationship to the input. A system may be defined as stable if a temporary change in the input or any temporary disturbance produces only a temporary change in the controlled variable.

Graphical methods for determining the stability of a system are usually the most useful in servomechanism analysis and design. These methods determine whether or not a system is stable and, in addition, can be used to determine the degree of

[1] See I. A. Greenwood, Jr., J. Vance Holdam, Jr., Duncan MacRae, Jr., "Electronic Instruments," pp. 393–397, McGraw-Hill Book Company, Inc., New York, 1948.

stability.   Graphical procedures also lead to suggestions as to methods of improvement of servo systems and permit the effects of various system elements on performance to be considered independently.

The various graphical methods most useful to servo design engineers are reviewed[1] briefly in the following discussion.

19.6a. Nyquist Diagram.   To apply Nyquist's stability criterion, it is necessary to determine if the open-loop transfer function $KG(s)$ is itself stable, i.e., contains no poles in the right half of the s plane.   If poles of $KG(s)$ do appear in the right half plane, then the open loop is unstable by itself.   This does not necessarily result in

(a) NYQUIST DIAGRAMS OF STABLE AND UNSTABLE SYSTEMS

(1) STABLE $K=A$ (2) UNSTABLE $K<A$ (3) UNSTABLE $K<A$

$$KG(s) = \frac{K(T_2 s + 1)^2}{s(T_1 s + 1)^2(T_3 s + 1)(T_4 s + 1)}$$ $T_1 > T_2 > T_3 > T_4$

(b) NYQUIST DIAGRAM OF A CONDITIONALLY STABLE SYSTEM
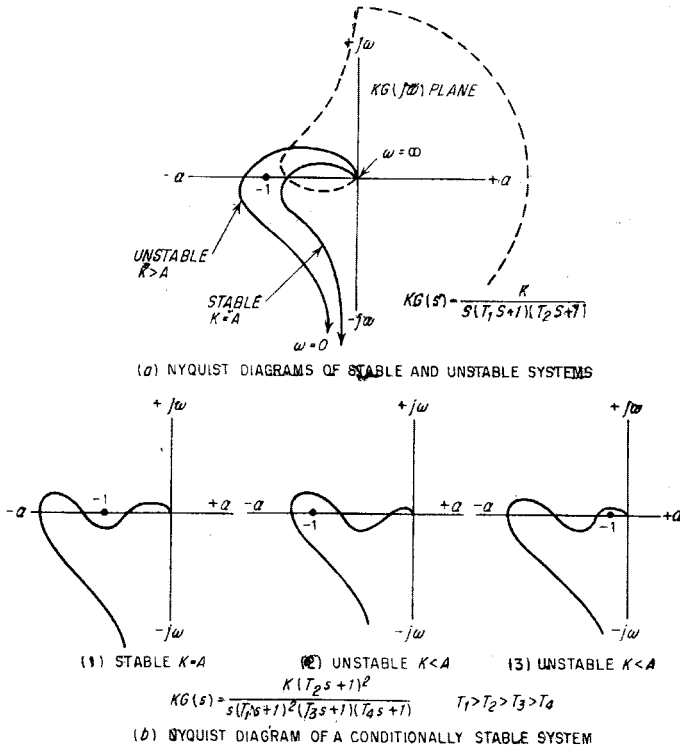
FIG. 19.27. Nyquist diagrams.

servo-system instability, however, since all of the poles of the closed loop can be in the left half plane even though there are poles of $KG(s)$ in the right half plane.

Most practical servo systems do not have poles of $KG(s)$ in the right half plane since the designer usually chooses stable components and also stabilizes subsidiary portions of a system.   Thus, it is seldom that a pole of $KG(s)$ appears in the right half plane in a practical servomechanism unless deliberately inserted by the designer. This leads to the simplified Nyquist criterion which states that if no poles of $KG(s)$ exist in the right half of the s plane, a system is stable if there are no encirclements of the -1, $j0$ point on the $KG(j\omega)$ plane as $j\omega$ is varied from $-j\infty$ to $+j\infty$ (see Sec. 18.4c[2]).   The complete plot is shown in Fig. 19.27a.   The portion of the locus from

[1] See Sec. 18.4 for a development of graphical methods of stability analysis.

[2] Note that in Sec. 18.4c the open-loop transfer function consists of the control or forward transfer function $KG(s)$ and the feedback transfer function $K_1G_1(s)$.   In the material herein, the open-loop transfer function has been referred to as $KG(s)$ for the sake of simplicity.

0 to $+j\infty$ is usually shown without the mirror image from 0 to $-j\infty$ and the semi-circle in the right half plane completing the closed contour.

Typical Nyquist plots of stable and unstable systems are illustrated in Fig. 19.27. The *conditionally stable* system of Fig. 19.27b is stable for a value of gain $K = A$. As the gain is increased or decreased from this value, the -1 point is encircled by the locus and the system becomes unstable.

19.6b. Inverse Nyquist Plot. It is sometimes found more convenient to use the inverse Nyquist plot. This is a plot of the inverse open-loop transfer locus $1/[KG(j\omega)]$.

Assuming there are no poles of $KG(s)$ in the right half plane, the conditions for stability are demonstrated in Fig. 19.28. The conditions for stability are reversed from the normal Nyquist plot, and stability is achieved only when the (-1) point is encircled.
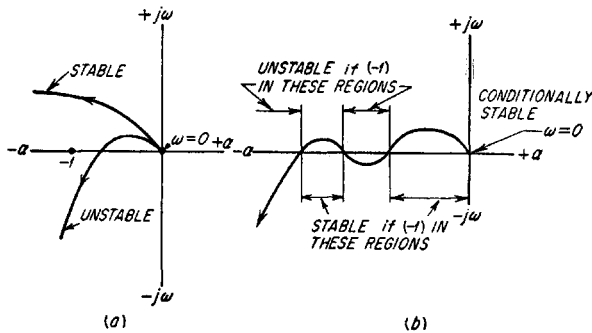
FIG. 19.28. Inverse Nyquist plots.

19.6c. Graphical Determination of the Open-loop Transfer Locus. The coordinates of points on a Nyquist plot of the open-loop transfer function can be graphically determined from a plot of the poles and zeros of the transfer function on the s plane. For example, consider the function

$$KG(s) = K \frac{\omega_n^2(T_1 s + 1)}{s(T_2 s + 1)(s^2 + 2\zeta\omega_n s + \omega_n^2)} \tag{19.80}$$

or, rearranged

$$KG(s) = K\omega_n^2 \frac{T_1}{T_2} \frac{s + 1/T_1}{s(s + 1/T_2)(s + \zeta\omega_n - j\omega_o)(s + \zeta\omega_n + j\omega_o)} \tag{19.81}$$

At a frequency $s = j\omega_1$:

$$KG(j\omega_1) = |KG(j\omega_1)|/\underline{\phi(j\omega_1)} \tag{19.82}$$

Equation (19.82) can be evaluated by locating the poles and zeros of Eq. (19.81) on the s plane and measuring lengths and angles to the point on the $j\omega$ axis corresponding to the frequency ω_1 at which $KG(j\omega)$ is being evaluated. This is illustrated in Fig. 19.29. From the figure,

$$KG(j\omega_1) = K\omega_n^2 \frac{T_1}{T_2} \left[\frac{L_3}{L_1 L_2 L_4 L_5} \right] \tag{19.83}$$

and

$$\phi(j\omega_1) = -[\phi_1 + \phi_2 - \phi_3 + \phi_4 + \phi_5] \tag{19.84}$$

where $L_1 = |j\omega_1|$

$L_2 = \left| j\omega_1 + \dfrac{1}{T_2} \right|$

$L_3 = \left| j\omega_1 + \dfrac{1}{T_1} \right|$

$L_4 = |\zeta\omega_n + j(\omega_1 - \omega_o)|$

$L_5 = |\zeta\omega_n + j(\omega_1 + \omega_o)|$

and $\phi_1 = 90°$
$\phi_2 = \tan^{-1} \omega_1 T_2$
$\phi_3 = \tan^{-1} \omega_1 T_1$

$\phi_4 = \tan^{-1} \dfrac{\omega_1 - \omega_o}{\zeta \omega_n}$

$\phi_5 = \tan^{-1} \dfrac{\omega_1 + \omega_o}{\zeta \omega_n}$

19.6d. Logarithmic Coordinates ("Bode" Diagram). Although a Nyquist diagram is useful for the determination of system stability, it is not easily interpreted in terms of circuit parameters and may require considerable labor.

By plotting $20 \log_{10} |K(j\omega)|$ on a linear scale as a function of ω on a logarithmic scale, a simpler and less laborious method can be used to determine system stability. Such a diagram will simplify the determination of steady-state performance and the
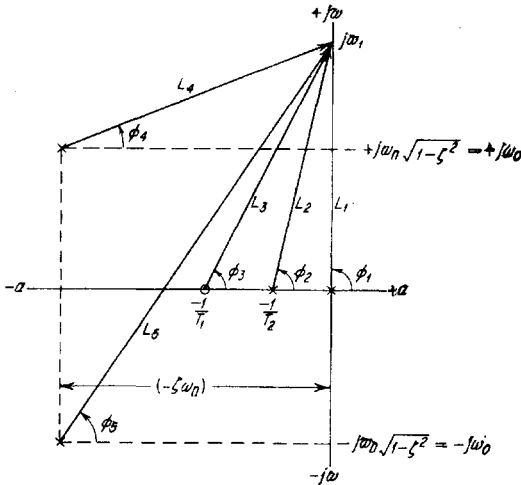
FIG. 19.29. Graphical determination of transfer functions.

visualization of the effects of adjustments of gain, bandwidth, and stabilization network parameters.

The use of such attenuation-log frequency diagrams for representing control element transfer functions is greatly facilitated by the use of straight-line approximations to the actual curve (see Sec. 18.5b). The resultant straight-line plots differ from the actual attenuation characteristic by only a few decibels, and with a little experience the system performance can be determined from the straight-line approximation with equal facility. This greatly reduces the labor of plotting and evaluating transfer functions. The straight-line functions commonly encountered are as follows:

1. *Frequency invariant factors* (K).

$$20 \log_{10} K = A \qquad \text{db} \tag{19.85}$$

This plots as a horizontal straight line as shown in Fig. 19.30a.
The phase angle ϕ of the factor K is equal to zero for all frequencies.
2. *Terms of the form* $(j\omega T + 1)$.
If $\omega T \ll 1$, then

$$20 \log_{10} |j\omega T + 1| = 20 \log_{10} 1 = 0 \qquad \text{db} \tag{19.86}$$

and if $\omega T \gg 1$

$$20 \log_{10} |j\omega T + 1| = 20 \log_{10} |j\omega T| = +6 \text{ db/octave} \tag{19.87}$$

Therefore, the term $(j\omega T + 1)$ may be represented by two straight-line asymptotes: a horizontal line at 0 db for low frequencies and a line of +6 db per octave at high

(a) PLOT OF FREQUENCY INVARIANT FACTOR K (b) PLOT OF $(j\omega T + 1)$ (c) PLOT OF $(j\omega T + 1)^{-1}$

(d) PLOT OF $(j\omega T)^{\pm 1}$ (e) PLOT OF $\left[(j\omega T)^2 + 2\zeta j\omega T + 1\right]^{-1}$
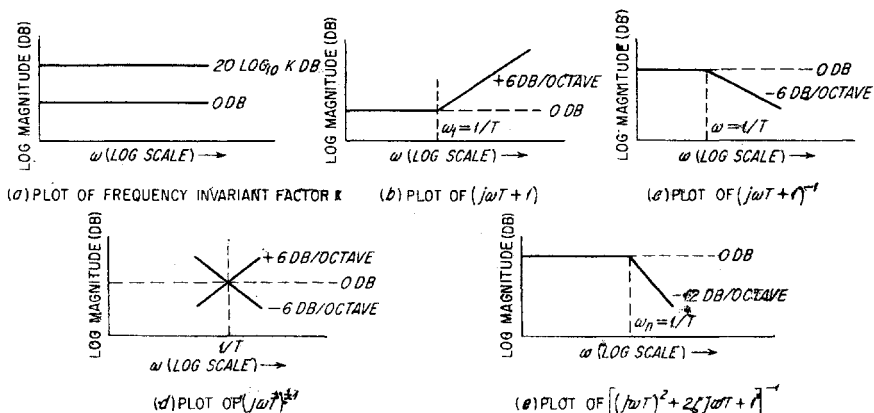
FIG. 19.30. Straight line approximations.

frequencies. These lines intersect at a break point corresponding to a break frequency ω such that $\omega T = 1$ as shown in Fig. 19.30b.

The asymptotes are in error -3 db at $\omega = 1/T$, -1 db one octave either side of the break frequency and -0.3 db two octaves either side of the break frequency.
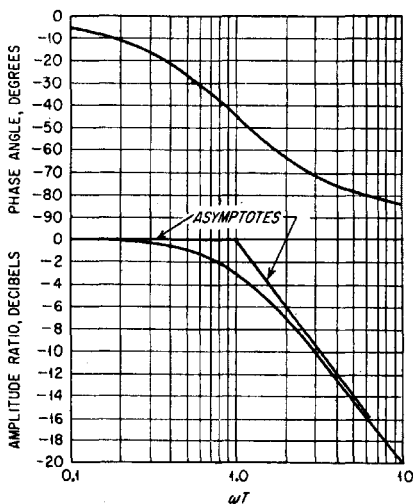
FIG. 19.31. Attenuation-phase plot of $(j\omega T + 1)^{-1}$.

4. *Terms of the form $j\omega T$ or $(j\omega T)^{-1}$.*

3. *Terms of the form $(j\omega T + 1)^{-1}$.* The function $(j\omega T + 1)^{-1}$ can be approximated by asymptotes having a slope of 0 db at low frequencies and -6 db per octave at high frequencies. These lines intersect at the break frequency ω such that $\omega T = 1$ as shown in Fig. 19.30c.

The asymptotes are in error +3 db at $\omega = 1/T$, +1 db one octave either side of the break frequency and +0.3 db two octaves either side of the break frequency.

The phase angles associated with $(j\omega T + 1)$ and $(j\omega T + 1)^{-1}$ are

$$\phi = \pm \tan^{-1} \omega T \qquad (19.88)$$

The phase angle is positive for $j\omega T + 1$ and negative for $(j\omega T + 1)^{-1}$. Figure 19.31 is a plot of phase angle and amplitude for the term $(j\omega T + 1)^{-1}$. This plot can be used for the $j\omega T + 1$ term by changing the sign of the ordinates.

$$\pm 20 \log_{10} |j\omega T| = \pm 6 \text{ db/octave} \qquad (19.89)$$

These factors have ± 6 db per octave slopes with a value of 0 db at such a frequency ω that $\omega T = 1$ as shown in Fig. 19.30d.

The phase angles associated with these terms are

$$\left.\begin{array}{l} \phi = +90° \text{ for } j\omega T \\ \phi = -90° \text{ for } 1/j\omega T \end{array}\right\} \text{ for all frequencies} \qquad (19.90)$$

5. *Terms of the form* $\dfrac{1}{(j\omega T)^2 + 2\zeta j\omega T + 1}$.

For $\omega T \ll 1$

$$-20 \log_{10} |(j\omega T)^2 + 2\zeta j\omega T + 1| = -20 \log_{10} 1.0 = 0 \text{ db} \qquad (19.91)$$

For $\omega T \gg 1$ the value becomes

$$-20 \log_{10} |(j\omega T)^2| = -12 \text{ db/octave} \qquad (19.92)$$

The typical plot of such a quadratic factor is shown in Fig. 19.30e. The break frequency occurs at $\omega_n = 1/T$. The error of the straight-line asymptotes depends on the value of the damping ratio ζ. If ζ is greater than 1, the roots are both real and the denominator is better handled by factoring into two first-order terms. The exact amplitude response as a function of ζ is shown in Fig. 19.20. The error between the straight-line asymptote and the true curve should always be taken into account for values of ζ less than 0.4 and for frequencies within an octave of ω_n.

(a) AMPLITUDE RESPONSE   (b) PHASE RESPONSE
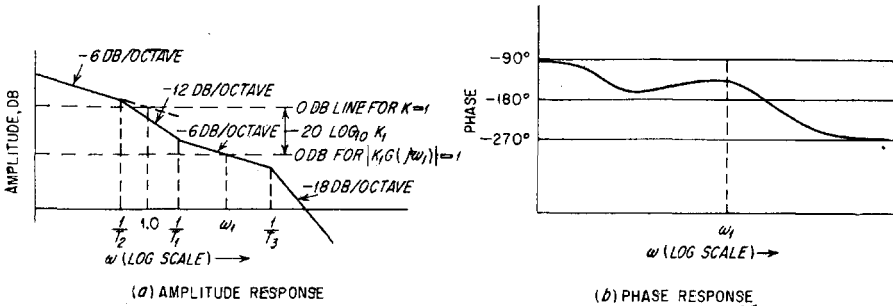
FIG. 19.32. Magnitude and phase plot of $KG(j\omega) = \dfrac{K(j\omega T_1 + 1)}{j\omega(j\omega T_2 + 1)[(j\omega T_3)^2 + 2\zeta j\omega T_3 + 1]}$.

The phase angle is also a function of ζ and is given by

$$\phi = -\tan^{-1} \frac{2\zeta\omega T}{1 - \omega^2 T^2} \qquad (19.93)$$

The phase angle for all values of ζ is $-90°$ when $\omega T = 1$ and asymptotically approaches 0 and $-180°$ as ω approaches zero and infinity. The phase angle is shown in Fig. 19.21 as a function of ζ and ω/ω_n.

Summary. The straight-line approximations of individual factors can be combined to form the over-all transfer function of several factors by the addition of the magnitude in decibels resulting from each term and the addition of the phase angle contributed by each term. A typical transfer function combining several of these factors is

$$KG(j\omega) = K \frac{j\omega T_1 + 1}{j\omega(j\omega T_2 + 1)[(j\omega T_3)^2 + 2\zeta j\omega T_3 + 1]} \qquad (19.94)$$

The resultant composite Bode diagram is shown in Fig. 19.32. The following factors are useful in determining the location of the 0-db gain line on the composite Bode plot:

1. The intersection of the -6-db-per-octave initial slope with the $\omega = 1$ line determines the location of the 0-db gain line for $K = 1$. In general, the 0-db line for $K = 1$ is determined by the intersection of the initial slope of the Bode plot with $\omega = 1$.

2. The intersection of the Bode plot with the value of $\omega = \omega_1$ such that

$$|K_1 G(j\omega_1)| = 1$$

gives the location of the 0-db line for gain $K = K_1$. The frequency ω_1 is called the cutoff frequency.

3. This line is displaced an amount 20 $\log_{10} K_1$, from the 0-db line at $K = 1$. The displacement is below the $K = 1$ line for $K_1 > 1$ and above the $K = 1$ line for $K_1 < 1$.

The phase of the function is obtained by

$$\phi(\omega) = \tan^{-1} \omega T_1 - \left[\frac{\pi}{2} + \tan^{-1} \omega T_2 + \tan^{-1} \frac{2\zeta\omega T_3}{1 - (\omega T_3)^2} \right] \qquad (19.95)$$

or by the graphical method shown in Fig. 19.29.

It should be noted that rapid changes in phase occur only in the vicinity of the break frequencies at which the slope of the amplitude characteristic changes. If the amplitude has a constant slope over any appreciable frequency range, then the associated phase is essentially constant. This correspondence of slope and phase is such that

1. Zero slope corresponds to 0° phase shift.
2. $\pm N$ 6 db per octave corresponds to $\pm N$ 90° phase shift.

This allows a quick check to be made on the correctness of the composite plot.

Occasionally it is necessary to determine experimentally the transfer function of a particular component of a complete servo system, e.g., by measuring the amplitude

(a) TYPICAL BODE PLOT   (b) CONDITIONALLY STABLE BODE PLOT
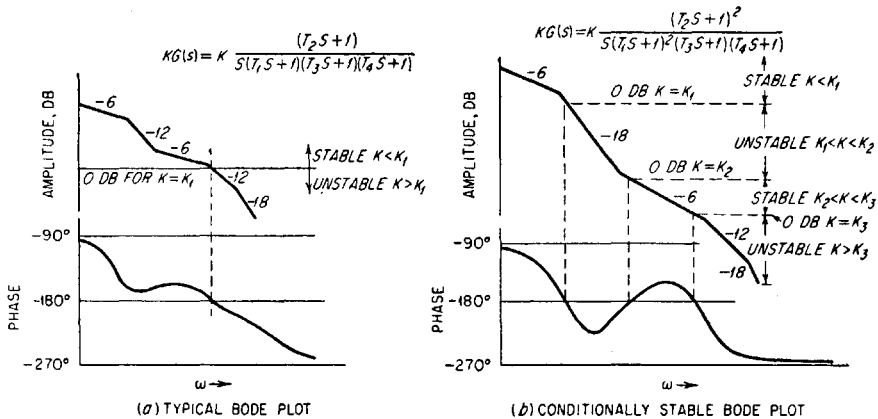
FIG. 19.33. Representative Bode plots for stability analysis.

and phase response of the component for sinusoidal imputs. By the approximation methods discussed in connection with Bode plots (see Sec. 18), the analytical expression for the transfer function can be obtained from the frequency-response data. This expression can then be combined with the transfer functions of other components in the system.

Stability. In a stable system, the open-loop amplitude curve must cross the 0-db line at a frequency where the corresponding phase lag is less than 180°. Figure 19.33 shows representative Bode diagrams illustrating how the stability can vary as a function of open-loop gain.

19.6e. Root-locus Method.[1] This method is extremely useful in the analysis and synthesis of closed-loop systems. The application of this method leads directly to the location of the poles of the closed-loop system transfer function in the *s* plane. The system is stable only if all of the poles appear in the left half plane.

The general expression for the system transfer function is (see Sec. 19.4a)

$$\frac{\theta_o(s)}{\theta_i(s)} = \frac{K_1 G_1(s)}{1 + K_1 G_1(s) K_2 G_2(s)}$$
$$= \frac{K_1 G_1(s)}{1 + KG(s)} \quad (19.96)$$

The poles of $\dfrac{\theta_o(s)}{\theta_i(s)}$ are determined by the values of *s* which make

$$1 + KG(s) = 0 \quad (19.97)$$
$$\text{or} \quad KG(s) = -1 = 1\underline{/180°} \quad (19.98)$$

The values of *s* which result in a net phase of 180° determine the locus of all possible roots of $1 + KG(s)$ on the *s* plane. The value of *K* determines the particular value of *s* on the locus of roots which satisfies $|KG(s)| = 1$ and thereby establishes the exact location of the poles on the locus.

Fig. 19.34. Vector representation of factors of $KG(s)$.

To illustrate this method, consider the open-loop transfer function

$$KG(s) = \frac{K(T_2 s + 1)}{s(T_1 s + 1)(T_3 s + 1)(T_4 s + 1)} \quad (19.99)$$

To visualize the application of the root-locus method more easily, rewrite Eq. (19.99) as

$$KG(s) = K \frac{T_2}{T_1 T_3 T_4} \left[\frac{s + 1/T_2}{s(s + 1/T_1)(s + 1/T_3)(s + 1/T_4)} \right] \quad (19.100)$$

The factors in the brackets are used to determine the locus of roots. Each factor can be represented as a vector as seen in Fig. 19.34. The locus of roots is found by determining the values of *s* such that

$$-\phi_1 - \phi_2 + \phi_3 - \phi_4 - \phi_5 = 180° \quad (19.101)$$

The value of *K* which will establish the location of the roots at a particular point on the locus is determined from Eq. (19.97); therefore

$$K_1 = \frac{T_1 T_3 T_4}{T_2} \left[\frac{L_1 L_2 L_4 L_5}{L_3} \right] \quad (19.102)$$

All the factors of the above equations can be measured directly from the plot. It should be noted that all the signs in Eq. (19.101) could be reversed and Eq. (19.97) would still be satisfied. Thus, there is another pole located at the mirror image about

[1] To the authors' knowledge, this method was first proposed by W. R. Evans. For additional details, see Sec. 18.4e. Note that in Sec. 18.4e the open-loop transfer function consists of the control or forward transfer function $KG(s)$ and the feedback transfer function $K_1 G_1(s)$. In the material herein, the open-loop transfer function has been referred to as $KG(s)$ for simplicity.

the real (α) axis of s in Fig. 19.34. The poles of a physically realizable network will always occur in conjugate pairs on the s plane if they are not located on the real axis.

Figure 19.35 illustrates the locus of roots method for determining system stability. Note that the number of poles of $\theta_o(s)/\theta_i(s)$ located on the plot is always equal to the number of poles of $KG(s)$.

19.6f. Degree of Stability. In the design of servo systems, the system must not only be stable but must be designed with a definite degree of stability. The degree of
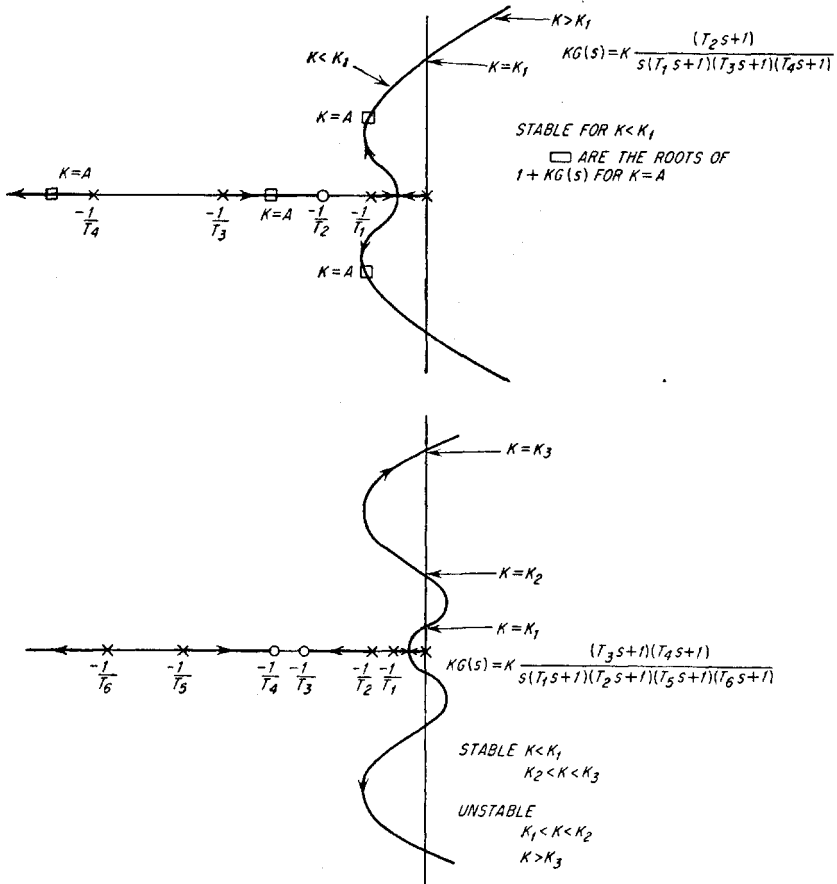
FIG. 19.35. Determination of stability from locus of roots.

stability is a measure of the closeness of the system to an unstable condition. Factors useful in determining the degree of stability from frequency-response data are phase margin, gain margin, and maximum magnification ratio, [see Eq. (19.73)] since the value of ζ cannot be found directly from the frequency plot of the open-loop transfer function.

Phase margin is defined as the difference between 180° and the phase of the open-loop transfer function $KG(j\omega)$ at unity gain. Gain margin is defined as the difference in decibels between unity gain and the magnitude of $KG(j\omega)$ at the frequency for which the phase of $KG(j\omega)$ is $-180°$. For a reasonable degree of stability, typical values for

phase and gain margin are 30 to 45° and −6 to −12 db, respectively. Figure 19.36 demonstrates the measurement of the gain and phase margins on Nyquist and Bode plots.

In most servo systems, satisfactory phase and gain margins are obtained by choosing the gain so that the 0-db gain line crosses the Bode plot through a slope of 6 db per octave at a point one or two octaves before the slope increases to 12 or more db per octave. If the slope is greater than 6 db per octave at frequencies both lower and higher than those frequencies at which the slope is 6 db per octave, the gain should be chosen such that the zero db point is at least two octaves above and two octaves below those frequencies at which the slope is greater than 6 db per octave.
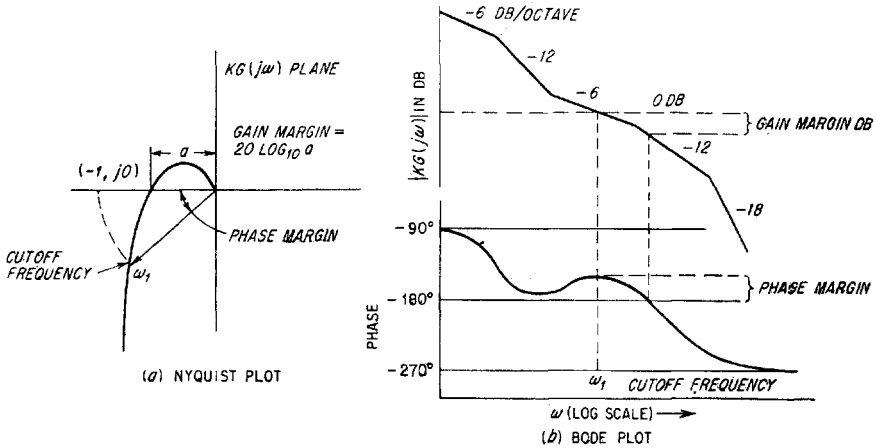
FIG. 19.36. Gain and phase margin.

The peak magnification ratio M_p is also a measure of the degree of stability. The inverse Nyquist plot can be used to determine M_p. The reciprocal of the closed-loop transfer function for a system with unity feedback[1] is

$$\frac{\theta_i(j\omega)}{\theta_o(j\omega)} = \frac{1}{KG(j\omega)} + 1 = \frac{1}{M}\underline{/-\phi} \qquad (19.103)$$

The loci of constant magnification ratios M are concentric circles about the $-1 + j0$ point on the inverse Nyquist plot. Radial lines from this point correspond to lines of constant ϕ. The value of M_p is the inverse of the radius of the smallest M circle which is tangent to the locus of the inverse open-loop transfer function $1/KG(j\omega)$ as shown in Fig. 19.37a. A complete plot of M versus ω can be obtained by determining the inverse of the radial line to the locus at each frequency. A typical result is shown in Fig. 19.37b.

The conversion of the open-loop transfer locus to the closed-loop transfer locus can also be obtained by the use of a Nichols chart.[2] The coordinates of this chart are $20 \log_{10} KG(j\omega)$ and the phase of $KG(j\omega)$ in degrees. Contours of constant M and ϕ are plotted on the chart, where

$$\frac{\theta_o(j\omega)}{\theta_i(j\omega)} = M\underline{/\phi} \qquad (19.104)$$

[1] For a general discussion of the case where the feedback is not unity, see Chestnut and Mayer, *op. cit.*, p. 236.
[2] See H. M. James, N. B. Nichols, and R. S. Phillips, "Theory of Servomechanisms," McGraw-Hill Book Company, Inc., New York, 1947.

The phase and amplitude of the open-loop frequency response, which can be obtained from a Bode plot, are plotted on the Nichols chart with ω as a parameter.   The values of M and ϕ are obtained from the intersection of the plot with the constant M and ϕ contours.   Shifting the plot parallel to the amplitude axis corresponds to changing the loop gain K, that is, an increased gain calls for an upward shift along the amplitude

(a) INVERSE NYQUIST PLOT

(b) AMPLITUDE RESPONSE
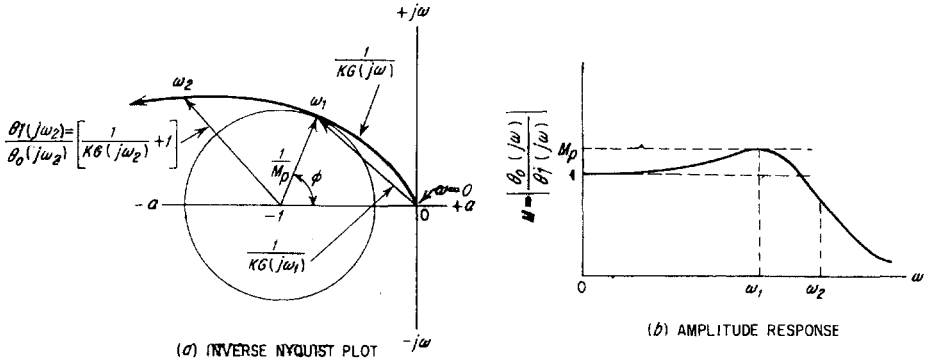
Fig. 19.37. Closed-loop amplitude response as related to the inverse Nyquist plot for a system having unity feedback.
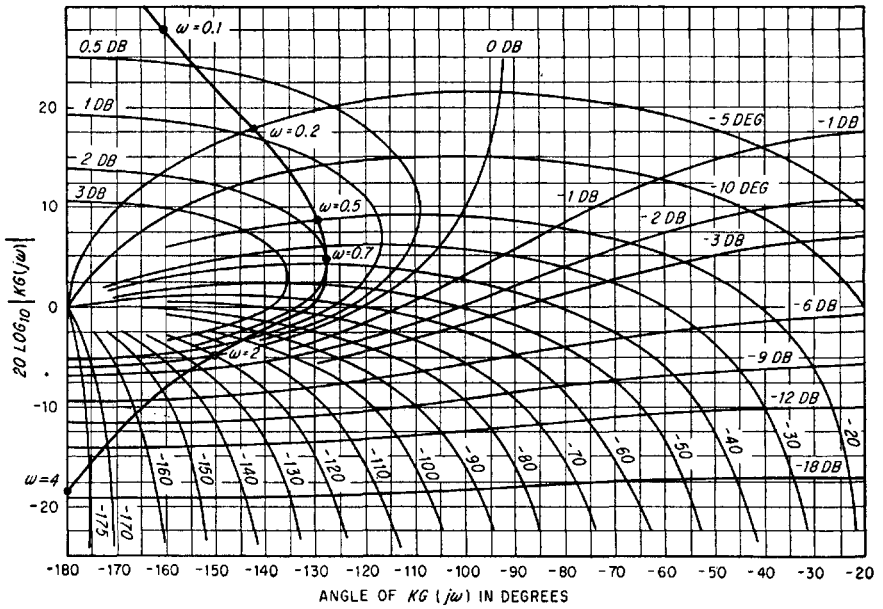
Fig. 19.38. Nichols chart.

axis by an amount equal to the gain change in decibels.   By this method, the open-loop gain for the desired maximum magnification ratio M_p can be determined.   Figure 19.38 is a typical plot of $KG(j\omega)$ on a Nichols chart.   The value of M_p for this plot is $+2$ db at $\omega = 0.7$ radian/sec.

A rule of thumb for most servo systems is that M_p should be limited to $1.2 < M_p < 1.4$ for optimum operation.   The value of ω for which the function $|KG(j\omega)| = 1$ is a measure of the speed of response of a system.   This is designated as the cutoff frequency ω_1 on Fig. 19.36b.

Since all poles of the closed-loop transfer function are located on a root-locus plot, the value of ζ and ω_o of predominant complex poles can be readily determined from the locus of roots. The desired degree of stability and speed of response can be defined by the selection of the area in the s plane where the predominant poles of a system should be located. The predominant poles are those contributing the greatest amplitude to the time response. The damping ratio ζ of the complex conjugate poles which are closest to the $j\omega$ axis is often used as a measure of stability (see Sec. 18.4b for an exception to this rule). In most servo systems $0.4 < \zeta < 0.8$ represents the region of best compromise between the speed of response and overshoot of the system to a step displacement input for a specified value of ω_n.

The root-locus method directly determines the locations of the poles of the closed-loop transfer function and provides a visualization of the effect on the closed-loop poles of changes in the location of the open-loop zeros and poles. The time response of the system can be determined from these roots by the methods of Sec. 23.6.

19.7. Error Coefficients and Stabilization Networks

19.7a. Types of Servo Systems. The type of servo system is defined by the number of poles n at the origin in the open-loop transfer function $KG(s)$. It also indicates the number of series integrations in the open loop.

Type 0 system. The open-loop transfer function $KG(s)$ has no net integrations or no poles at the origin. A system of this type is usually considered to be a regulating device and not a servomechanism.

Type 1 system. The open-loop transfer function has one net integration or a single pole at the origin.

Type 2 system. The open-loop transfer function has two net integrations or two coincident poles at the origin.

The general shapes of the Nyquist and Bode plots for frequencies near zero are illustrated by the solid lines in Fig. 19.39. Typical responses at higher frequencies are illustrated by the dotted lines. The initial slopes of the Bode plots are 0 db per octave, -6 db per octave, and -12 db per octave for Type 0, Type 1, and Type 2

Fig. 19.39. Representative open-loop system plots.

servos, respectively. It should be noted from Fig. 19.39 that the axis approached by the $KG(j\omega)$ locus as ω approaches zero on the Nyquist plots is dictated by the number of poles at the origin, increasing clockwise from the positive real axis one quadrant for each pole.

19.7b. Error Coefficients. The performance of a system is generally specified in terms of either its static or its dynamic accuracy. Static accuracy is a measure of system performance for desired values of the output in the steady state. Dynamic accuracy indicates the system performance for desired values of the output considering non-steady-state conditions and the effect of system time constants.

The following static error coefficients are defined.

1. Position error coefficient

$$K_p = \frac{\text{output}}{\text{actuating error}} \qquad (19.105)$$

for a constant value of the output.

2. Velocity error coefficient

$$K_v = \frac{\text{velocity of output}}{\text{actuating error}}$$ (19.106)

for a constant velocity of the output.

3. Acceleration error coefficient

$$K_a = \frac{\text{acceleration of output}}{\text{actuating error}}$$ (19.107)

for a constant acceleration of the output.

K_p is dimensionless; K_v has the units per second and K_a the units per second squared. The actuating error in each case is the error signal for fixed output position, uniform output velocity, and uniform output acceleration respectively.

The static error coefficients can also be defined as follows

$$K_p = \lim_{s \to 0} KG(s)$$ (19.108)

$$K_v = \lim_{s \to 0} sKG(s)$$ (19.109)

$$K_a = \lim_{s \to 0} s^2 KG(s)$$ (19.110)

These error coefficients can be interpreted in terms of the steady-state error by use of the final value theorem. The results are that K_p, K_v, and K_a are each inversely proportional to the steady-state error if the input is a unit step [$\theta_i(s) = 1/s$], unit ramp [$\theta_i(s) = 1/s^2$], and unit parabolic function [$\theta_i(s) = 1/s^3$] respectively. The static error coefficients for type 0, type 1, and type 2 second-order systems with unity feedback are, from Eqs. (19.108) to (19.110),

	Type 0	Type 1	Type 2
K_p	K	∞	∞
K_v	0	K	∞
K_a	0	0	K

where K is the gain term associated with the open-loop transfer function $KG(s)$.

Dynamic error coefficients obtained from the input function and the error transfer function have a more direct physical significance than those obtained in the above procedure from the output function and the open-loop transfer function. The determination of these error coefficients usually requires considerable calculation, however, which may be prohibitive.[1]

The following conclusions may be drawn from the preceding discussion:

1. The Type 0 system has a static error proportional to the output displacement for a constant output displacement.

2. The Type 1 system has zero static error for a constant output displacement. For a constant output velocity, the Type 1 system has a constant error which is proportional to the output velocity.

3. The Type 2 system has zero static error for a constant output displacement or a constant output velocity. For a constant acceleration of the output, the Type 2 system has a constant error which is proportional to the output acceleration.

[1] For a general discussion of error coefficients see S. G. Truxal, "Automatic Feedback Control Systems Synthesis," pp. 80–97, McGraw-Hill Book Company, Inc., New York, 1955.

The static error coefficients are equal to the loop gain in each case. The dynamic error is a function of both the loop gain and the time constants of the system.

The error coefficients discussed in this section are based on no torque loading on the output. In a practical system having static torque loading, the exact error cannot be established by the use of these coefficients. If the actuator has been properly chosen so as to satisfy the actual torque losses, the error in utilizing these error coefficients will be small.

19.7c. Stabilization. It is usually found in the design of high-performance servo systems that the maximum system gain determined from stability considerations is insufficient to provide the desired accuracy and that the closed-loop transfer-function time constants do not provide the desired speed of response. Adjustments in the open-loop transfer function to improve the performance characteristics of the system
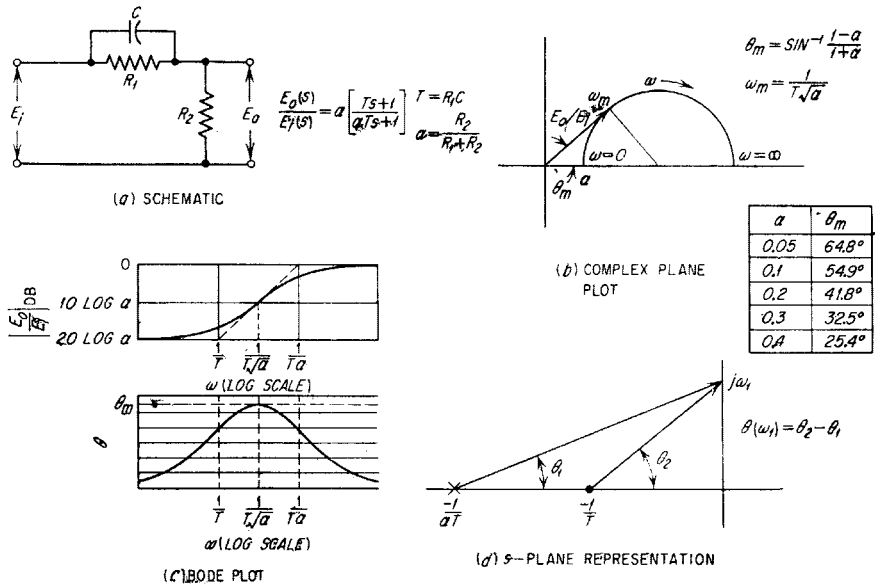
FIG. 19.40. Sample lead network.

thereby permitting an increase in gain can be made by inserting a "stabilizing network" in series with the forward-loop transfer function or in series with the feedback element. Characteristics of the most generally useful series stabilization networks are discussed in the following paragraphs. Phase and amplitude characteristics of a variety of RC networks which may be useful in this application are given in Table 1.4.

1. Lead Network. A system may often be required to respond to frequency components of the input signal higher than can be accomplished for the existing loop gain and system bandwidth. Such a requirement imposes a problem of synthesis involving the addition of a compensating network exhibiting a positive phase shift over portions of the frequency range to increase the cutoff frequency thereby permitting an increase in the loop gain commensurate with the desired degree of stability.

A network that accomplishes this form of compensation is shown in Fig. 19.40a. The phase characteristic of the network produces a leading phase shift only over a limited range of frequencies. The proper choice of $T = R_1C$ is determined by the

frequency region in which it is desirable for the phase lead to be appreciable. The maximum value of phase lead obtainable is determined by the value of α.

The complex plane plot of the transfer function of this network is shown in Fig. 19.40b. The maximum phase lead θ_m obtainable for a given value of α is the angle between the tangent to the locus and the positive real axis as shown on the figure. The table of Fig. 19.40b gives values of θ_m corresponding to various values of α. The amplitude and phase characteristics of this network are shown in Fig. 19.40c. The s-plane location of zeros and poles of the transfer function is shown in 19.40d.

Disadvantages of lead networks are d-c attenuation which must be compensated by increased amplifier gain and amplification of "noise" due to the increased high-frequency response.
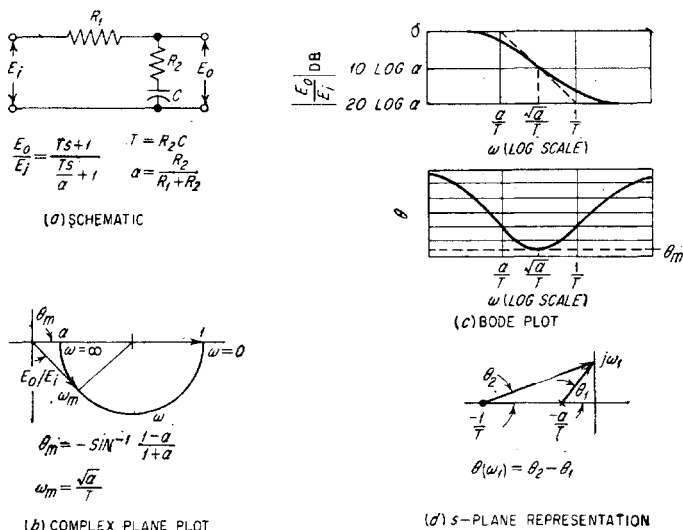
FIG. 19.41. Sample lag or integral network.

2. Lag or Integral Network. It is usually necessary to increase the static error coefficients of an uncompensated system. The lag network shown in Fig. 19.41a permits the gain to be increased at low frequencies without altering the gain characteristic near the cutoff frequency. The complex plane plot, amplitude and phase responses, and s-plane representation are shown in Fig. 19.41b, c, and d, respectively.

In general, the allowable increase in gain of a system after insertion of a lag network is approximately equal to $1/\alpha$. The choice of the time constant T is determined by the frequency region in which it is desired to increase the loop gain. The frequency $\omega = 1/T$ should be at least two octaves below the cutoff frequency.

3. Lag-Lead Network. The advantages of lag and lead networks can be obtained with the single lag-lead network shown in Fig. 19.42. This network may provide the phase lead required to extend the cutoff frequency without decreasing the degree of stability and also provides an increased low-frequency gain, thereby improving the static error coefficient. These characteristics combine to allow a greater increase in loop gain than would be possible with either network used alone. The region between $1/T_1$ and $1/T_b$ adds phase lead near the cutoff frequency. The region between $1/T_a$ and $1/T_2$ provides increased low-frequency gain where required. The frequency $\omega = 1/\sqrt{T_1 T_2}$ must occur below the uncompensated cutoff frequency of the system. The various time constants are interrelated by the circuit constants, and only three

variables can be selected independently, for example, T_1, T_2, and T_3 or T_1, T_2, and A. The general characteristics of a lag-lead network are shown in Fig. 19.42b, and the location of zeros and poles of the transfer function are shown in Fig. 19.42c.
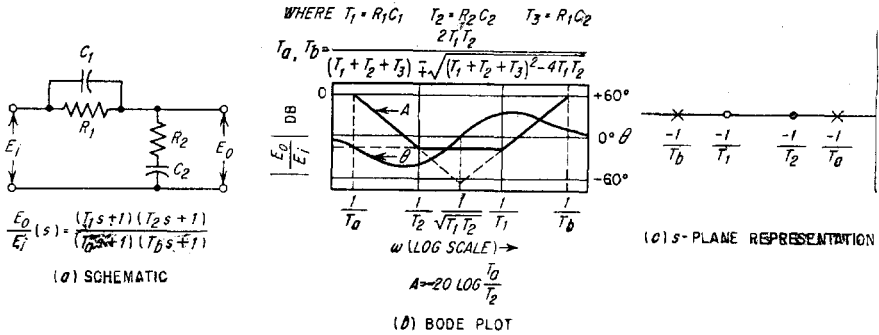
FIG. 19.42. Sample lag-lead network.

The use of these stabilizing networks is illustrated by the following example.

Example 19.5

A Type 1 servo system is shown in Fig. 19.43.   The open-loop transfer function is

$$K_V G(s) \;=\; \frac{K_\epsilon K_N K_A K_G K_m}{N} \;\frac{G_N(s)}{s(T_G s + 1)(T_m s + 1)} \tag{19.111}$$

where K_ϵ = error detector gain, volts per degree error
K_N = equalizer gain, in volts/volt
K_A = amplifier gain, amp/volt
K_G = generator gain, volts/amp
K_m = motor speed constant, deg/sec/volt
$1/N$ = gear reduction between motor shaft and load
$G_N(s)$ = frequency-variant portion of equalizer transfer function
T_G = generator time constant
T_m = motor time constant including effect of load inertia and damping reflected to motor shaft
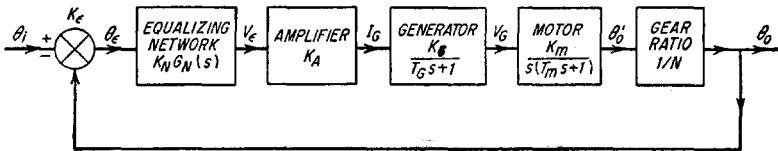
FIG. 19.43. Type 1 servo system.

The velocity error coefficient K_V is

$$K_V \;=\; \frac{K_\epsilon K_N K_A K_G K_m}{N} \qquad \text{degrees per second output per degree error} \tag{19.112}$$

The frequency variant portion of the open-loop transfer function $G(s)$ is

$$G(s) \;=\; \frac{G_N(s)}{s(T_G s + 1)(T_m s + 1)} \tag{19.113}$$

The characteristics of the motor-generator combination are shown in Fig. 19.44.   The slopes of the Bode plots without equalization are determined by the motor-generator combination, as they are the only components with frequency variant transfer functions.   The requirement is to improve the performance of this system by use of a lag network, then a lead network, and finally a lag-lead network.

Lead Network. The uncompensated Bode plot of the system is shown by the solid lines in Fig. 19.45a. The slopes in decibels per octave are indicated by the numbers on the amplitude plot. The cutoff frequency corresponding to a reasonable phase margin is indicated by ω_1.

1. Determine the location of the lead network time constants on the Bode diagram.

The lead network should be introduced in the 12-db-per-octave section between the motor and generator time constants to introduce phase lead at the proper point. This results in the modified amplitude and phase curves indicated by the dotted lines.

2. Determine the new cutoff frequency.

The new cutoff frequency is located on the modified amplitude curve, approximately midway between the break points of the lead network where a satisfactory phase margin is provided. A suitable frequency is shown in the diagram as ω_2.

3. Determine the allowable increase in gain.

The allowable increase in gain is equal to the difference in position of the 0-db line for the compensated and uncompensated cutoff frequencies This is in addition to the increase in gain $1/\alpha$ required to offset the attenuation introduced by the lead network. The effect of

$$\frac{\theta_o}{E_i}(s) = \frac{K_m K_G}{s\left[T_G s+1\right]\left[T_m s+1\right]}$$

(a) BLOCK DIAGRAM   (b) BODE PLOT
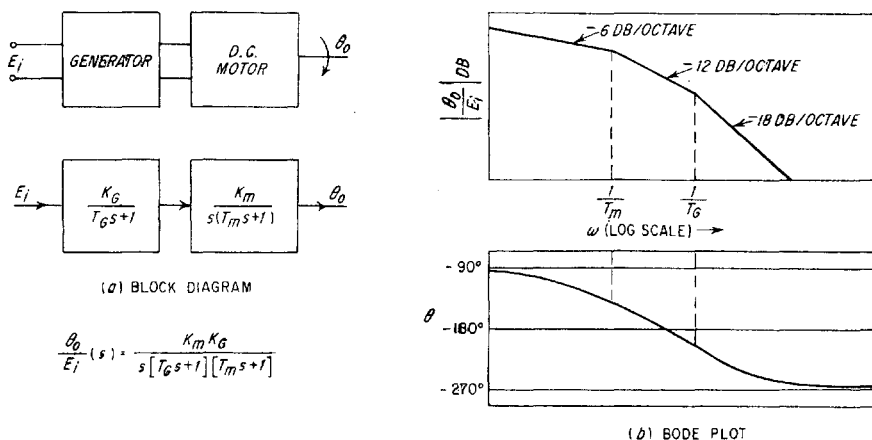
Fig. 19.44. D-c motor and generator.

lead-network compensation on the root-locus diagram of the system is shown in Fig. 19.45b. The solid curve gives the locus of the roots in the uncompensated system. The squares show the location of the roots in the uncompensated system if the gain is increased to the value of the compensated system. It can be seen that such a gain increase without compensation would result in instability. The location of the roots in the compensated system are shown on the dotted curve. The effect of lead compensation in this example has been to increase the cutoff frequency and increase the gain of the system without changing the degree of stability. This results in an increase in the velocity error coefficient and a reduction in the response time.

Lag Network. The uncompensated Bode plot is repeated in Fig. 19.46a. The low-frequency portion has been expanded for the purpose of illustration. The cutoff frequency of the uncompensated system is again indicated by ω_1.

1. Determine the location of the lag-network time constants on the Bode diagram.

The lag-network time constants should be effective at low frequencies to improve the relative gain in this region. They should be as far from the cutoff frequency as practical. Placing the time constants at lower frequencies than are necessary generally results in poorer dynamic accuracy and increases the time required for the system to respond to a step displacement. For the position chosen on the diagram, the amplitude and phase plots are indicated by the dotted lines.

2. Determine the allowable increase in loop gain.

The 0-db gain position for the uncompensated system is shown in the figure. The lag network reduces the gain at frequencies above $1/T_1$ by the factor $20 \log 1/\alpha$. The over-all system gain can therefore be increased by this amount, thereby increasing the low-frequency gain and restoring the cutoff frequency to a frequency slightly below the cutoff frequency before compensation.

3. Determine the new cutoff frequency.

The new cutoff frequency can be established as shown in the figure by determining the frequency corresponding to the original phase margin. This frequency is slightly lower than in the uncompensated case because of the additional lagging phase introduced by the lag network. The effect of this type of compensation on the root-locus plot of the system is shown in Fig. 19.46*b*. The solid lines show the loci of roots of the uncompensated system. The squares on these lines show the location of the roots in the uncompensated system for the desired value of gain. It can be seen that this system is unstable at that
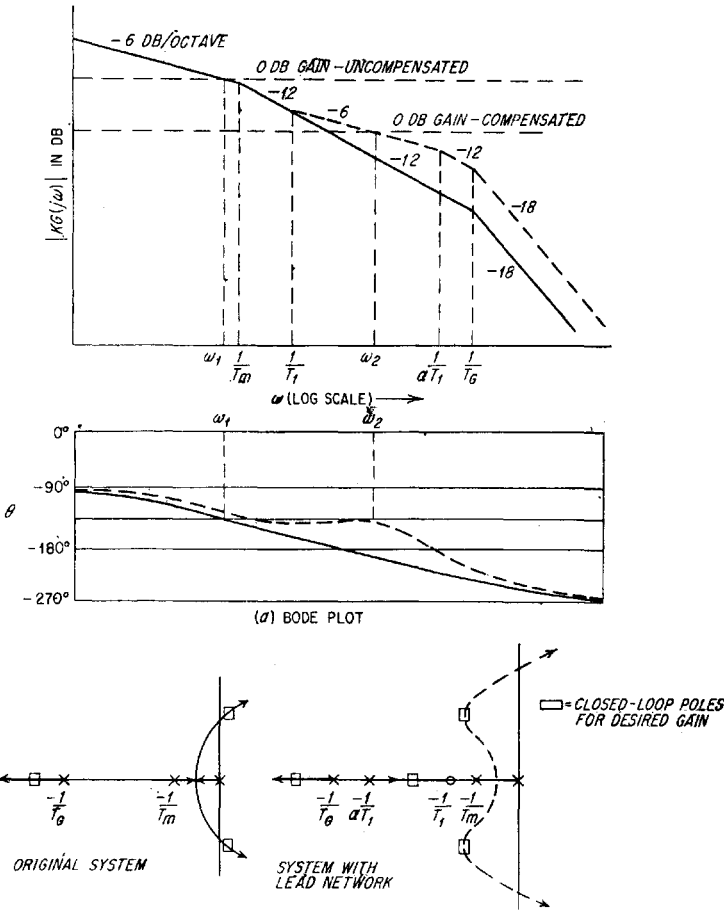
FIG. 19.45. Type 1 system with lead equalization.

value of gain. The dotted curve shows the modified root-locus plot and the position of the roots for the same gain after stabilization.

The effect of lag stabilization in this example is to increase the low-frequency gain, thereby increasing the velocity error coefficient without appreciably affecting the cutoff frequency or the degree of stability.

Lag-Lead Network. The uncompensated Bode plot is again repeated by the solid curves of Fig. 19.47*a*. The cutoff frequency ω_1 is selected for a reasonable phase margin.

1. Determine the location of the lag-lead-network time constants on the Bode diagram.

The previous requirements for lead and lag networks apply to this case with the restriction imposed by the interrelation between time constants as indicated in Fig. 19.42. A representative choice is shown in Fig. 19.47*a*. The amplitude and phase plots for this location of time constants are shown by the dotted lines.

2. Determine the allowable increase in loop gain.

In this case the allowable gain is increased to make up for the reduction in high-frequency gain due to the lag portion of the network in addition to the increase in gain made possible by the use of the lead portion of the network. The new cutoff frequency ω_2 is located on the 6-db-per octave slope between the $1/T_1$ and $1/T_b$ break points at a position corresponding to the original phase margin. The gain change is then the vertical separation between
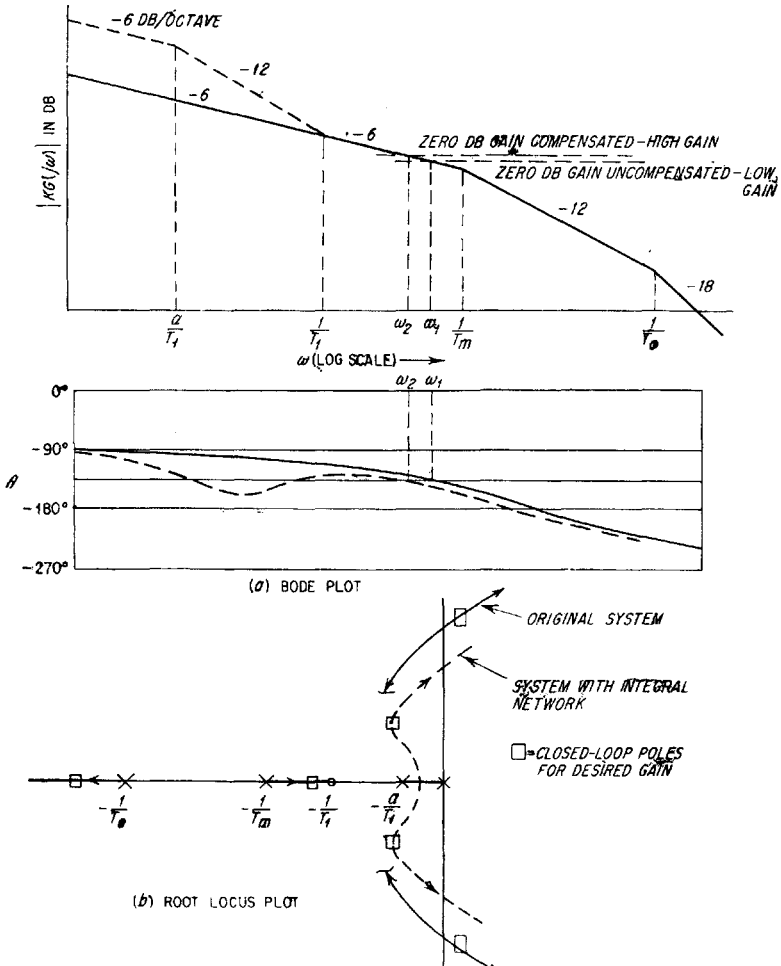
FIG. 19.46. Type 1 system with integral network.

the compensated and uncompensated 0-db lines plus approximately 20 $\log_{10} T_a/T_2$ for the lag portion of the network.

The effect of stabilization on the root-locus plot for this system is illustrated by the curves of Fig. 19.47b. The effect of lag-lead stabilization in this example has been to improve considerably the velocity error coefficient and response time with no decrease in the degree of stability.

Compensation by Tachometer Feedback. An alternate method of stabilization is to modify the forward transfer function of the system by means of an auxiliary feedback loop as shown in Fig. 19.48. This loop can be reduced to a single transfer function which can then be incorporated in the over-all closed loop as shown in Fig. 19.49.

(a) BODE PLOT

(b) ROOT LOCUS PLOT
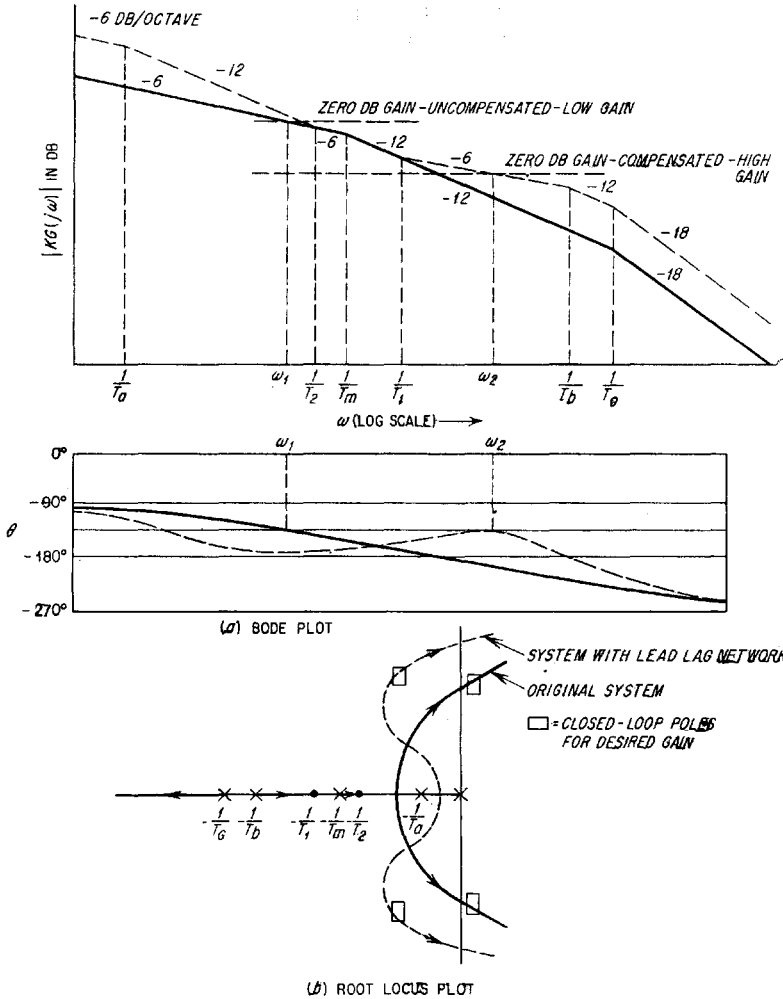
FIG. 19.47. Type 1 system with lag-lead equalization.

$$\omega_n^2 = \frac{K_A' K_G K_m K_T}{T_m T_G} \text{ , FOR } K_A' K_G K_m K_T \gg 1$$

$$2\zeta\omega_n = \frac{T_m + T_G}{T_m T_G}$$

$$r_1 = -\zeta\omega_n + j\omega_n\sqrt{1-\zeta^2}$$

$$r_2 = -\zeta\omega_n - j\omega_n\sqrt{1-\zeta^2}$$

TACHOMETER FEEDBACK LOOP EQUIVALENT TRANSFER FUNCTION
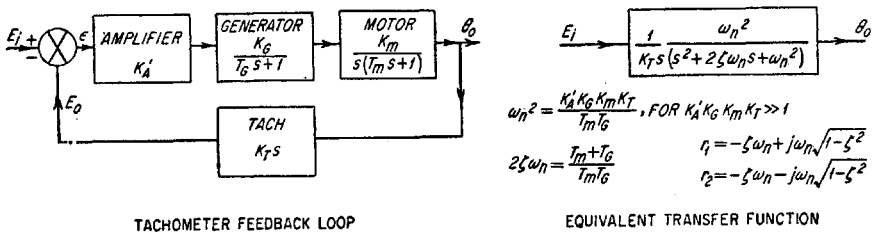
FIG. 19.48. Block diagrams of tachometer feedback loop.

A tachometer providing an output voltage proportional to the rate of rotation of the input shaft of the tachometer is often used as the feedback element in this application. The modified transfer function of the tachometer feedback loop corresponds to a second-order system as discussed in Sec. 19.4b. The characteristics of the tachometer are shown in Fig. 19.50 for reference.

The effect of tachometer feedback on the system performance is illustrated by the Bode plot of Fig. 19.51a. The solid curves show the amplitude and phase of the open-loop transfer function before application of tachometer feedback. The dotted curves show the improvement in amplitude and phase characteristics after application of feedback. In effect, this form of compensation extends the break point of the
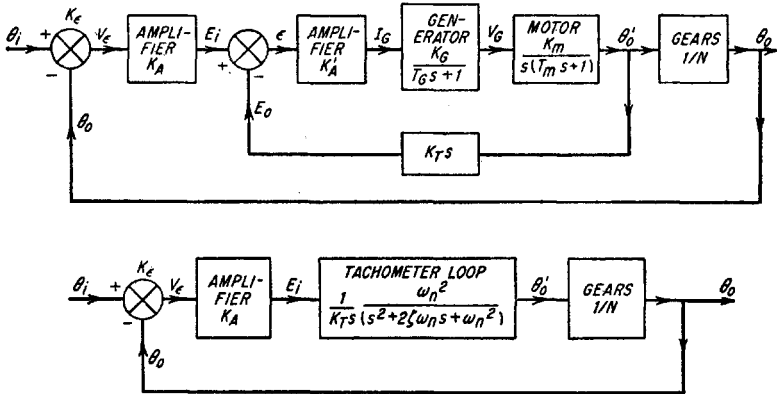
FIG. 19.49. Modification of a servo system by tachometer feedback.

(a) BLOCK DIAGRAM   (b) BODE PLOT   (c) S-PLANE REPRESENTATION
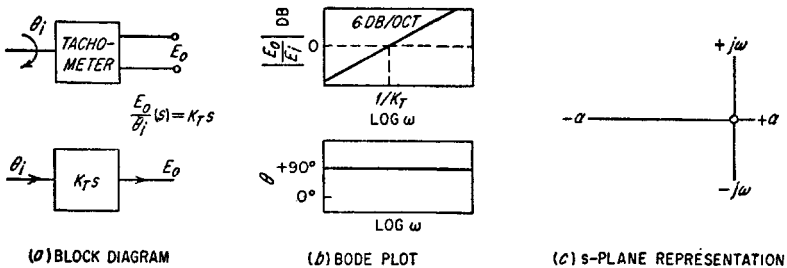
FIG. 19.50. Tachometer characteristics.

6-db-per-octave slope to a point determined by ω_n as given in Fig. 19.48. This allows an increase in cutoff frequency to ω_2 and an increase in system gain as shown in Fig. 19.51a without a reduction in degree of stability.

The effect of tachometer feedback on the root-locus plot of the system is shown in Fig. 19.51b.

When tachometer feedback is employed, due consideration must be given to the performance of the tachometer loop considered as an isolated feedback system.

An advantage of tachometer feedback is the very large system gain available to counteract steady-state and frictional load disturbances. This is because any disturbance which reduces the motion of the output shaft also reduces the feedback signal from the tachometer, causing the over-all system gain to be increased by as much as $K_a' K_G K_m K_T$.

19.7d. Design Procedure. The design of a servomechanism usually involves a compromise between the performance objectives and limitations imposed by the

characteristics of the system elements. It is usually necessary to obtain an approximate equipment configuration based on available or achievable element characteristics, analyze the resultant system performance, correct or adjust the performance by means of gain changes and equalization, reanalyze the modified system, and finally, experimentally verify the indicated performance characteristics. No fixed procedure is applicable for all design problems; however, the following suggested steps are usually helpful in organizing the design effort.

1. Statement of Requirements. This includes all known performance objectives, load characteristics, and physical or environmental conditions.

The specification of objectives includes factors such as the type of servo required (constant output for constant input, constant rate of change of output for constant
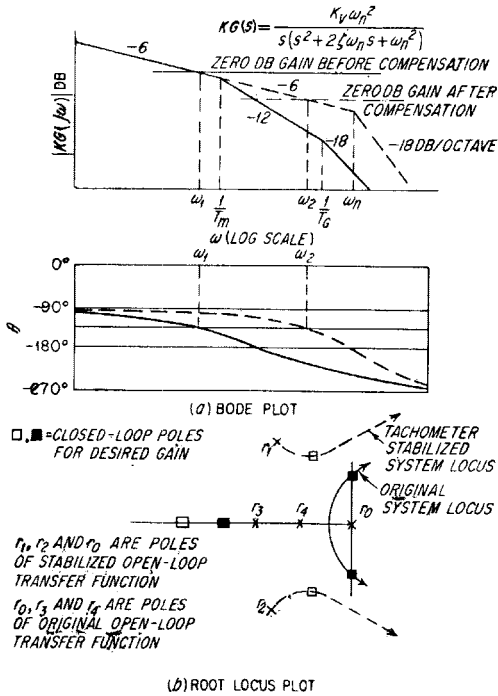
FIG. 19.51. Type 1 system stabilized with tachometer feedback.

input, etc.), the form of input and output data (electrical signals, shaft rotation, etc.), required bandwidth, speed of response, allowable steady-state error, overshoot, time delay, and duration of transient oscillation.

The specification of load characteristics includes the expected load inertia, damping characteristics, and load disturbances; limits of displacement, velocity, and acceleration; and duty cycle of operation.

2. Selection of Elements. The elements discussed in Sec. 19.5b are selected in accordance with the requirements of the system. The servomotor is usually chosen first with characteristics which are compatible with the load and system requirements. The controller is then specified with power characteristics suitable for the chosen motor. Tentative compensation networks are included in the controller when they can be determined in advance. The other elements can usually be chosen in any order, or they may be specified by the original requirements of the system.

3. Analysis of Open- and Closed-loop Characteristics.   Using the methods of Sec. 19.6, degree of stability, overshoot, speed of response, and transient behavior to step and ramp input functions are determined.   This analysis usually results in a need for modification of gain and adjustment of the controller transfer function to provide proper equalization.

4. Stabilization and Improvement of Performance.   This consists of applying the methods of Sec. 19.7 to improve the characteristics analyzed in step 3.

5. Experimental Verification.   After the over-all system transfer function has been adjusted to prove the best compromise between the requirements of the system, the complete servomechanism should be simulated in the laboratory to obtain experimental confirmation of the design.   In many cases, the system design cannot be completed without considerable effort in experimentally determining such factors as transfer functions of elements, evaluating nonlinearities, and observing the effect of disturbances on the system.   It is often necessary to simulate load effects in such tests.

Transmission Lines

20.1. Fundamentals of Transmission Lines. Transmission lines are used to convey energy from a source to a load. They can be of many types depending upon the application and frequency involved. The length of the line can be a fraction of a wavelength as in the case of a 60-cycle power-distribution line where a wavelength is approximately 3,100 miles, or it can be many wavelengths as in the case of a radar transmission line operating at 30,000 Mc where a wavelength is only 0.39 in. At the lower frequencies, open-wire lines are usually employed and conventional concepts of current flowing through the conductors and potential difference between conductors may be used without restriction. At high frequencies, open-wire lines are replaced by coaxial lines, and at very high frequencies (1,000 Mc and up) waveguides, consisting of hollow metal tubing, are used.

Transmission lines contain "distributed constants," i.e., resistance, capacitance, inductance, and conductance spread uniformly along the line. An analysis of transmission-line behavior can be made, however, based on the assumption that the constants of the line are divided into small sections of lumped elements, in which case the accuracy of the assumption and the validity of the results increase as the number of such sections increases for any given length of line.

20.1a. Transmission-line Equations. For purposes of analysis, a transmission line is considered to consist of the following elements:

l, series inductance per unit length
r, series resistance per unit length
c, shunt capacitance per unit length
g, shunt conductance per unit length

The lumped network representation of a unit section of line is shown in Fig. 20.1.

The shunt voltage e along a transmission line is attenuated because of the voltage drop caused by the flow of current i through the series impedance z. The series current along the line is attenuated because of the shunt current through the shunt admittance y.
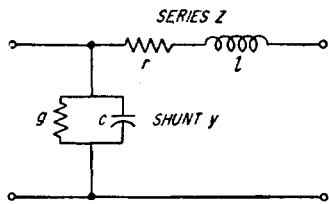
FIG. 20.1. Lumped section of transmission line.

If the load impedance at the receiving end of a transmission line does not have a particular value called the *characteristic impedance* Z_o, a portion of the voltage and current traveling along the line toward the load will be reflected at the load, resulting in voltage and current waves traveling back toward the generator end of the line.

If at the load, $e_o{}^+$ and $i_o{}^+$ are the incident voltage and current and $e_o{}^-$ and $i_o{}^-$ are the reflected voltage and current, the voltage e_x and current i_x at any point a distance x from the receiving end of the line are given by

$$e_x = (e_o{}^+)(\epsilon^{\gamma x}) + (e_o{}^-)(\epsilon^{-\gamma x}) \qquad (20.1)$$
$$i_x = (i_o{}^+)(\epsilon^{\gamma x}) + (i_o{}^-)(\epsilon^{-\gamma x}) \qquad (20.2)$$

20-2

where γ is the *propagation constant* given by

$$\gamma = \sqrt{zy} = \sqrt{(r + j\omega l)(g + j\omega c)} \tag{20.3}$$

The propagation constant can possess real and imaginary parts as follows:

$$\gamma = \alpha + j\beta \tag{20.4}$$

Therefore,

$$\epsilon^{\gamma x} = \epsilon^{\alpha x + j\beta x} = \epsilon^{\alpha x}\epsilon^{j\beta x} \tag{20.5}$$

The quantities $\epsilon^{\alpha x}$ and $\epsilon^{-\alpha x}$ are real numbers representing an increase and decrease in amplitude, respectively, with increasing x. The factors $\epsilon^{j\beta x}$ and $\epsilon^{-j\beta x}$ result in a phase advance and delay, respectively, with increasing x.

20.1b. Line Characteristics.

Characteristic Impedance. The characteristic impedance of a transmission line is defined by

$$Z_o = \sqrt{\frac{z}{y}} = \sqrt{\frac{r + j\omega l}{g + j\omega c}} \tag{20.6}$$

Z_o can also be defined in terms of the voltage and current amplitudes of the incident and reflected waves as follows:

$$Z_o = \frac{e_x{}^+}{i_x{}^+} = \frac{e_x{}^-}{i_x{}^-} \tag{20.7}$$

It can be seen from Eq. (20.6) that for a line with negligible losses ($\omega l \gg r$, $\omega c \gg g$)

$$Z_o = \sqrt{\frac{l}{c}} \tag{20.8}$$

If this assumption is not valid, Eq. (20.6) must be used and Z_o will be complex.

Attenuation. The real part α of the propagation constant is the attenuation constant in nepers per unit length (1 neper equals 8.686 db). The ratio of voltages or currents at points A and B due entirely to a wave going from A to B is given by

$$\frac{e_A}{e_B} = \frac{i_A}{i_B} = \epsilon^{\alpha \Delta x} \tag{20.9}$$

where Δx is the separation between A and B.

The attenuation constant can be expressed in terms of the line constants by expanding Eq. (20.3). The result is, for small values of attenuation,

$$\alpha = \frac{r}{2Z_o} + \frac{gZ_o}{2} \qquad \text{nepers per unit length} \tag{20.10}$$

The first term is due to conductor losses, and the second is due to dielectric losses.

Phase Shift. The imaginary part β of the propagation constant is the phase constant expressed in radians per unit length. In one wavelength, the phase of a propagated signal will shift 2π radians; therefore,

$$\beta = \frac{2\pi}{\lambda} \tag{20.11}$$

where λ = wavelength in transmission line

The phase constant expressed in terms of the line constants is given by Eq. (20.12) for small values of attenuation.

$$\beta = \omega \sqrt{lc} \left[1 + \frac{1}{2} \left(\frac{r}{2\omega l} - \frac{g}{2\omega c} \right)^2 \right] \tag{20.12}$$

For a lossless line this reduces to

$$\beta = \omega \sqrt{lc} \qquad (20.13)$$

Equation (20.13) is also valid for the condition where the conductor and dielectric losses are equal and approximately correct when $\omega l \gg r$ and $\omega c \gg g$.

Group and Phase Velocity.   The amplitude of the voltage or current of steady-state traveling wave at any point a distance x along a transmission line in the direction of wave travel is proportional to $\cos (\omega t - \beta x)$ where $(\omega t - \beta x)$ is the phase angle.   A point of constant phase of the traveling wave is represented by maintaining a constant value for the parenthetical expression, i.e., $\omega t - \beta x = k$.   If this equality is differentiated with respect to time, the following expression for velocity v_p of a point of constant phase of the traveling wave is obtained.

$$v_p = \frac{dx}{dt} = \frac{\omega}{\beta} = f\lambda \qquad (20.14)$$

The velocity represented by v_p is normally called the *phase velocity* of the traveling wave.   If this traveling wave is a carrier signal of amplitude E_c of angular frequency ω_c, the expressions for the instantaneous voltage e at any point on the line, time of transmission t_c, and velocity of phase propagation v_p are as follows:

$$e = E_c \cos (\omega_c t - \beta_c x) \qquad (20.15)$$

$$t_c = \frac{\beta_c x}{\omega_c} \qquad (20.16)$$

$$v_p = \frac{\omega_c}{\beta_c} \qquad (20.17)$$

If this carrier is modulated at a frequency ω_m with an amplitude E_m, upper and lower sideband frequencies will be propagated down the line along with the carrier. If the difference in phase constant between the value β_c at the carrier frequency and the values β_{s1} and β_{s2} at the sideband frequencies is the same magnitude and of opposite sign, i.e., β_{s2} is greater than β_c by the amount that β_c is greater than β_{s1}, and is designated as β_m, the expression for the modulated carrier is

$$e = [E_c + E_m \cos (\omega_m t - \beta_m x)] \cos (\omega_c t - \beta_c x) \qquad (20.18)$$

The expressions enclosed within the brackets is the peak amplitude of the carrier, or the modulation envelope.   From the previous discussion it can be seen that the velocity v_g of the modulation envelope and the time of transmission t_m of the modulation envelope are given by

$$v_g = \frac{\omega_m}{\beta_m} \qquad (20.19)$$

$$t_m = \frac{\beta_m x}{\omega_m} \qquad (20.20)$$

Since ω_m is the difference in frequency between the carrier and either sideband and β_m is the difference in phase constant between the carrier and either sideband, Eq. (20.19) can be expressed as follows:

$$v_g = \frac{\omega_m}{\beta_m} = \frac{d\omega}{d\beta} \qquad (20.21)$$

If the value of β is proportional to ω, then $d\omega/d\beta$ is equal to ω_c/β_c and the velocities expressed by Eqs. (20.14) and (20.21) are equal.   In this case the carrier and modulation envelope travel down the line with the same velocity.   If β is not proportional

to ω, then $d\omega/d\beta$ is not equal to ω_c/β_c and the carrier and modulation envelopes are propagated with different velocities. The velocity as given by Eqs. (20.19) and (20.21) is called the *group velocity*. Intelligence or signal energy is seen to travel at group velocity.

In a *nondispersive* medium the phase velocity is not a function of frequency and the group and phase velocities are equal.

In a *dispersive* medium the phase velocity is a function of frequency causing the group and phase velocities to differ. In communication systems employing dispersive transmission lines, the group velocity is almost always less than the phase velocity.

It can be seen from Eq. (20.12) that, a transmission system will be nondispersive if it is without dissipation, that is, $r = g = 0$, or if $r/\omega l = g/\omega c$, or if the conductor loss is equal to the dielectric loss. Under these conditions, from Eqs. (20.13) and (20.17)

$$v_p = v_g = \frac{1}{\sqrt{lc}} = v \qquad (20.22)$$

where v = velocity of light in a vacuum

In a complicated wave train containing more than one frequency component the signal intelligence is propagated at the group velocity; however, if the spectrum of the modulating signal becomes very wide, the concept of group velocity loses significance because of the large differences in arrival time of the various signal components unless $d\omega/d\beta$ is constant over the frequency range of the entire signal spectrum.

In waveguide transmission systems, the wavelength in the guide λ_g is greater than the velocity of light divided by the frequency of the source. Since the propagated frequency is not altered by the transmission system, it can be seen from Eq. (20.14) that the phase velocity in this case is greater than the velocity of light.

It can be shown that

$$v_p v_g = v^2 \qquad (20.23)$$

and

$$v_g = v_p + \beta \frac{dv_p}{d\beta} \qquad (20.24)$$

In a waveguide transmission system phase velocity can be measured by measurement of λ_g and use of Eq. (20.14). Group velocity can then be determined by use of Eq. (20.23), (20.24), or direct measurement of intelligence transmission time.

20.1c. Reflections and Standing Waves. If a transmission line is infinitely long or is terminated with the characteristic impedance at the frequency of the applied signal, there will be no reflections from the end of the line and the only signal appearing on the line will be the incident wave. If some other termination is used, a portion of the incident wave will be reflected and the total signal appearing on the line will be the sum of the incident and reflected waves. The maximum voltage will be the sum of the maximum magnitudes of the incident and reflected voltages. The minimum voltage will be the difference between the maximum magnitudes of the incident and reflected voltages. Voltage or current maxima or minima will be separated by one-half wavelength measured along the line. Maximum voltage will occur at the point of minimum current, and this will be the point of maximum impedance. Minimum voltage will occur at the point of maximum current, and this will be the point of minimum impedance. Minimum and maximum impedance points will be separated by one-quarter wavelength. If the line has attenuation, the voltage and current standing wave ratios will decrease in amplitude exponentially as the distance from the load is increased. The maximum and minimum values of impedance will decrease in a similar manner and will approach the characteristic impedance of the line as a limiting value.

The *voltage standing wave ratio* (VSWR, or ρ) is the ratio between the maximum and minimum values of voltage or current. The *power standing wave ratio* (SWR, or ρ^2) is the square of the voltage standing wave ratio. Thus,

$$\rho = \frac{e_{max}}{e_{min}} = \frac{i_{max}}{i_{min}} \tag{20.25}$$

$$SWR = (VSWR)^2 = \rho^2 \tag{20.26}$$

If losses are negligible and if the termination is a pure resistance, the voltage standing wave ratio is equal to the ratio of the magnitudes of the terminating resistance to characteristic impedance, or characteristic impedance to terminating resistance, whichever is greater than 1.

The relationship between the incident and reflected voltage waves at any point along a transmission line is expressed by the *reflection coefficient* $\Gamma_x \angle \psi_x$. The reflection coefficient gives the magnitude Γ_x and phase angle ψ_x of the ratio of reflected voltage to incident voltage. If a transmission line has appreciable attenuation, the reflection coefficient will vary in magnitude, decreasing exponentially toward zero as the distance from the load is increased.

The voltage of the incident wave at a point a distance x from the load is

$$e_x^+ = e_o^+ \epsilon^{\gamma x} \tag{20.27}$$

The voltage of the reflected wave at the same point is

$$e_x^- = e_o^- \epsilon^{-\gamma x} \tag{20.28}$$

where e_o^+, e_o^- are incident and reflected voltages at load ($x = 0$)

If the magnitude and phase (in radians), respectively, of the reflection coefficient at the load are denoted by Γ_o and ψ_o, then

$$\Gamma_o \angle \psi_o = \frac{e_o^-}{e_o^+} \tag{20.29}$$

The reflection coefficient, a distance x from the load, can then be expressed as

$$\Gamma_x \angle \psi_x = \frac{e_x^-}{e_x^+} = \frac{e_o^- \epsilon^{-\gamma x}}{e_o^+ \epsilon^{\gamma x}} = \frac{e_o^-}{e_o^+} \epsilon^{-2\gamma x} \tag{20.30}$$

or $\qquad\qquad \Gamma_x \angle \psi_x = \Gamma_o \epsilon^{-2\alpha x} \epsilon^{i(\psi_o - 2\beta x)} \tag{20.31}$

It can be seen from Eq. (20.31) that the magnitude Γ_x of the reflection coefficient decreases exponentially toward zero as the distance from the load increases. Equation (20.31) is shown graphically in Fig. 20.2.

The effect of attenuation on the impedance along the line can be seen from the following.

At any instant the total voltage at any point x along the line is equal to the sum of the incident and reflected voltages. Thus,

$$e_x = e_x^+ + e_x^- \tag{20.32}$$

FIG. 20.2. Variation of reflection coefficient along a transmission line (attenuation exaggerated).

Therefore, from Eq. (20.30)

$$e_x = e_x^+ + e_x^+ \Gamma_x \angle \psi_x = e_x^+ (1 + \Gamma_x \angle \psi_x) \tag{20.33}$$

The reflection coefficient expressing the relationship between reflected and incident current waves is the negative of the reflection coefficient discussed above because the phase difference between incident and reflected voltage differs 180° from the phase

difference between incident and reflected current.   Therefore

$$i_x = i_x{}^+(1 - \Gamma_x \angle \psi_x) \tag{20.34}$$

Dividing Eq. (20.33) by (20.34),

$$Z_x = \frac{e_x{}^+(1 + \Gamma_x \angle \psi_x)}{i_x{}^+(1 - \Gamma_x \angle \psi_x)} \tag{20.35}$$

At any point along a transmission line the absolute magnitude of the ratio of incident voltage to incident current or reflected voltage to reflected current is the characteristic impedance of the line, Z_o.   The ratio of total voltage (incident plus reflected) to total current is the actual impedance at that point of the line, Z_x.   Therefore,

$$Z_x = Z_o \frac{1 + \Gamma_x \angle \psi_x}{1 - \Gamma_x \angle \psi_x} \tag{20.36}$$

and in terms of the reflection coefficient at the load,

$$Z_x = Z_o \frac{1 + \Gamma_o \epsilon^{-2\alpha x} \epsilon^{j(\psi_o - 2\beta x)}}{1 - \Gamma_o \epsilon^{-2\alpha x} \epsilon^{j(\psi_o - 2\beta x)}} \tag{20.37}$$

Neglecting attenuation, the maximum and minimum values of Z_x are obtained when the quantity $\psi_o - 2\beta x$ is equal to zero and π, respectively.   The result is

$$Z_{\max} = Z_o \frac{1 + \Gamma_o}{1 - \Gamma_o} \tag{20.38}$$

$$Z_{\min} = Z_o \frac{1 - \Gamma_o}{1 + \Gamma_o} \tag{20.39}$$

The VSWR has been defined as the ratio of adjacent maximum and minimum voltages on the line.   The maximum voltage is the sum of the magnitudes of the incident and reflected voltage waves, and the minimum voltage is equal to the difference between the incident and reflected waves.   Thus,

$$e_{\max} = |e^+{}_{\max}| + |e^-{}_{\max}| \tag{20.40}$$
$$e_{\min} = |e^+{}_{\max}| - |e^-{}_{\max}| \tag{20.41}$$

From Eq. (20.33)

$$e_{\max} = e^+{}_{\max}(1 + \Gamma_x) \tag{20.42}$$
and
$$e_{\min} = e^+{}_{\max}(1 - \Gamma_x) \tag{20.43}$$

Dividing Eq. (20.42) by (20.43),

$$\rho = \frac{e_{\max}}{e_{\min}} = \frac{1 + \Gamma_x}{1 - \Gamma_x} \tag{20.44}$$

where ρ = VSWR

Conversely
$$\Gamma_x = \frac{\rho - 1}{\rho + 1} \tag{20.45}$$

For Eqs. (20.44) and (20.45) to be valid when the line has appreciable attenuation, Γ_x and ρ must be evaluated at the same point.   Equations (20.38) and (20.39) can also be expressed as

$$Z_{\max} = Z_o \rho \tag{20.46}$$
and
$$Z_{\min} = \frac{Z_o}{\rho} \tag{20.47}$$

The limiting cases of reflection are expressed as follows for various load impedances Z_R:

1. If $Z_R = Z_o$, then $\Gamma_x \angle \psi_x = \Gamma_o \angle \psi_o = 0$, $\rho = 1$, $e_o{}^- = i_o{}^- = 0$, and all of the incident energy is absorbed by the load.

2. If $Z_R = 0$, then $\Gamma_o \angle \psi_o = 1 \angle 180° = -1$, $\rho = \infty$ for a lossless line, $e_o{}^- = -e_o{}^+$, $i_o{}^- = i_o{}^+$, and the incident energy is totally reflected.

3. If $Z_R = \infty$, then $\Gamma_o \angle \psi_o = 1 \angle 0 = 1$, $\rho = \infty$ for a lossless line, $e_o{}^- = e_o{}^+$, $i_o{}^- = -i_o{}^+$, and the incident energy is totally reflected.

The relationships expressed by Eqs. (20.44) and (20.45) are plotted for convenience in Fig. 20.3.

Several additional useful relationships involving standing wave ratio are as follows:[1]

1. Effect of VSWR on Attenuation Constant. The attenuation constant of a transmission line is increased when standing waves are present because of the increased copper loss caused by the higher current densities in the conducting surfaces. The corrected value of attenuation constant α_r is given by

$$\alpha_r = \alpha \left(\frac{\rho^2 + 1}{2\rho} \right) \qquad (20.48)$$

where α = attenuation constant for $\rho = 1$

2. Resultant VSWR when n mismatches combine in the worst phase combination (highest VSWR). When a transmission-line system has more than one discontinuity or point reflecting a portion of the incident power, the resultant standing wave pattern is a combination of the individual standing waves. If there are n points of reflection and voltage standing wave ratios corresponding to each point of reflection are $\rho_1, \rho_2, \ldots, \rho_n$, then the resultant VSWR when the n mismatches combine in the worst manner is

FIG. 20.3. Relationship between reflection coefficient magnitude and voltage standing wave ratio.

$$\rho_{max} = \rho_1 \rho_2 \rho_3 \ldots \rho_n \qquad (20.49)$$

3. Resultant VSWR when n mismatches combine in best phase (lowest VSWR). If the n mismatches combine in the best possible phase, the resultant VSWR is given by

$$\rho_{min} = \frac{\rho_n}{\rho_1 \rho_2 \ldots \rho_{(n-1)}} \qquad (20.50)$$

or is equal to unity, whichever is greater, where

$$\rho_1 < \rho_2 < \rho_3 < \cdots < \rho_n$$

[1] G. L. Ragan, ed., "Microwave Transmission Circuits," Table 2.2, p. 35, Radiation Laboratory Series, vol. 9, McGraw-Hill Book Company, Inc., New York, 1947.

Equations (20.49) and (20.50) are strictly true only for the case where each discontinuity, except the termination at the load end of the line, is a pure reactance (lossless).

4. Reduction of VSWR by means of an Isolating Attenuator.[1] It is often necessary to reduce the VSWR presented to a generator by an unmatched load, e.g., a transmission line with reflections or a short line not terminated in its characteristic impedance. This can be accomplished by inserting an isolating attenuator between the generator and the load. A design chart for determining the required attenuation α to achieve a desired reduction in VSWR is shown in Fig. 20.4.

20.1d. Impedance. The input impedance of a transmission line of length x and termination Z_R is given by

$$Z_x = Z_o \frac{Z_R + Z_o \tanh \gamma x}{Z_o + Z_R \tanh \gamma x} \quad (20.51)$$

FIG. 20.4. VSWR reduction by isolation attenuation.

The use of this equation is somewhat complicated in that the quantity γx is in general a complex number, and charts of the functions of complex hyperbolic angles must be utilized. In a majority of practical cases this difficulty can be overcome by assuming the losses (attenuation) to be negligible. Equation (20.51) then becomes

$$Z_x = Z_o \frac{Z_R + jZ_o \tan \beta x}{Z_o + jZ_R \tan \beta x} \quad (20.52)$$

Equation (20.52) can be evaluated by the use of standard tables of trigonometric functions.

Several special cases are of interest. If the line is sufficiently long that no energy is reflected to the point x during the interval that the incident signal is present, the input impedance Z_x can be considered equal to Z_o. If the terminating impedance Z_R is equal to Z_o, then $Z_x = Z_o$ for all lengths of the line. If the line is short-circuited, Z_x is given by Eq. (20.53). If the line is open-circuited, Z_x is given by Eq. (20.54).

$$Z_x = jZ_o \tan \beta x \quad (20.53)$$
$$Z_x = -jZ_o \cot \beta x \quad (20.54)$$

The input impedances expressed by Eqs. (20.53) and (20.54) are inverse with respect to each other; i.e., the coefficients of Z_o are reciprocal. These expressions are shown graphically in Fig. 20.5. Short- or open-circuited lines of the proper length will exhibit the impedance properties of series and parallel resonant circuits.

The following conditions should be noted:

1. Short-circuit termination

Z_x is inductive for $0 < x < \lambda/4$
Z_x is ∞ for $x = \lambda/4$
Z_x is capacitive for $\lambda/4 < x < \lambda/2$
Z_x is zero for $x = \lambda/2$

---

[1] Adapted from unpublished papers of J. G. McCann.

2. Open-circuit termination

Z_x is capacitive for $0 < x < \lambda/4$
Z_x is zero for $x = \lambda/4$
Z_x is inductive for $\lambda/4 < x < \lambda/2$
Z_x is ∞ for $x = \lambda/2$

3. For a line one-half wavelength long:

$$Z_x = Z_R \qquad (20.55)$$

4. For a line one-quarter wavelength long:

$$Z_x = \frac{Z_o{}^2}{Z_R} \qquad (20.56)$$

Equation (20.56) shows that a quarter-wavelength line can be used to transform impedances from a high Z_R to a low Z_x or from a low Z_R to a high Z_x.
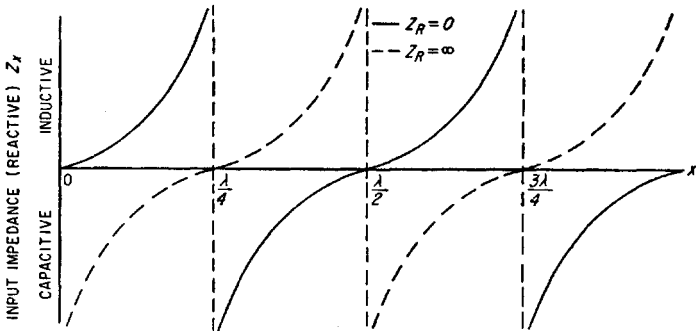
Fig. 20.5. Input impedance of short-circuited and open-circuited lossless transmission lines.

20.1e. Power and Efficiency. In a transmission line containing standing waves, the net power P transmitted down the line is

$$P = \frac{e_{\max}e_{\min}}{Z_o} = p^+ - p^- \qquad (20.57)$$

where p^+ is the incident power, p^- is the reflected power, and $e_{\max}$ and $e_{\min}$ are rms voltages. When no standing waves are present,

$$P = p^+ = \frac{e_x{}^2}{Z_o} \qquad (20.58)$$

The maximum power which can be sent down the line is limited by the line breakdown voltage and standing wave ratio and is given by

$$P = \frac{e_b{}^2}{\rho Z_o} \qquad (20.59)$$

where e_b = rms breakdown voltage

Standing waves decrease the maximum power which can be sent down the line by the factor $1/\rho$. This is illustrated in Fig. 20.6.

The effect of altitude upon maximum power is shown in Fig. 20.7. This is based upon the breakdown voltage being proportional to air density. It is of interest to

note that humidity makes only a 5 per cent reduction in the maximum power which can be sent down a transmission line. The effect of sharp corners or bends, however, is very pronounced in reducing the maximum power. For pulse applications, the breakdown power is approximately inversely proportional to the cube root of the pulse length and to the eighth root of the repetition frequency over a range of pulse lengths from approximately 0.5 to 5 μsec and a range of repetition frequencies from 15 to 2,000 cps.[1]
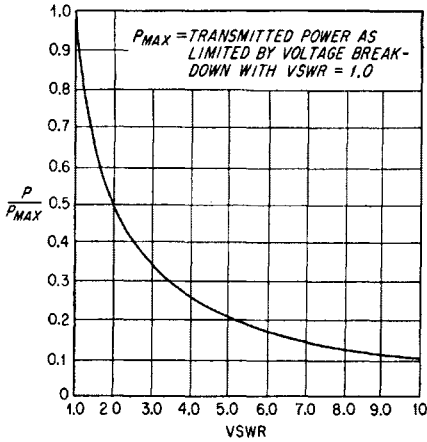
FIG. 20.6. Relationship between power and voltage standing wave ratio.
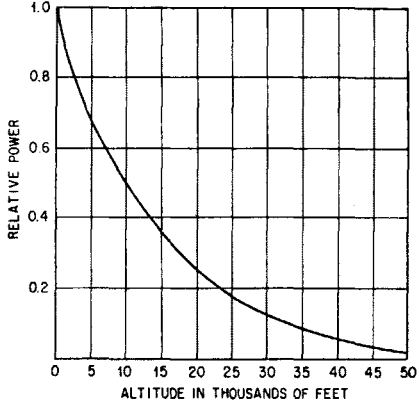
FIG. 20.7. Variation of maximum power with altitude.

20.1f. Losses and Skin Effect. Power loss in a transmission line consists of both conductor loss and dielectric loss. The total loss in any transmission line is always the sum of these two quantities. The attenuation in decibels per unit length because of these losses is 8.68α. The attenuation constant α must be evaluated for each type of transmission line, but in general is given by Eq. (20.10).

$$\alpha = \frac{r}{2Z_o} + \frac{gZ_o}{2} \qquad \text{nepers per unit length} \qquad (20.10)$$

The first term is due to conduction loss, the second to dielectric loss.

In high-frequency or microwave transmission-lines, the resistance per unit length r is greater than the d-c resistance. This is because the current density in the conducting elements decreases exponentially for increasing distance from the surface exposed to the electromagnetic fields. This phenomenon is known as *skin effect*. For purposes of determining power loss, a current density equal to 0.707 of the surface current density can be considered to be distributed uniformly from the surface to the *penetration depth* δ as given by Eq. (20.60).

$$\delta = \frac{1}{\sqrt{\sigma \pi f \mu}} \qquad \text{meters} \qquad (20.60)$$

where f = frequency, cps
   μ = permeability, henrys/m
   σ = conductivity, ohms/m

The permeability μ is equal to $4\pi \times 10^{-7}$ henry/meter for nonmagnetic materials and k times that value for magnetic materials, where k is the ratio of the permeability of the material to the permeability of air.

[1] Ragan, *op. cit.*, pp. 227–242.

For copper, δ reduces to

$$\delta_{\text{copper}} = \frac{6.6}{\sqrt{f}} \quad \text{cm} \tag{20.61}$$

Physically, δ is the depth at which the current density has decreased to $1/\epsilon$ of its surface value. At this depth the current lags the surface current in phase by 1 radian.
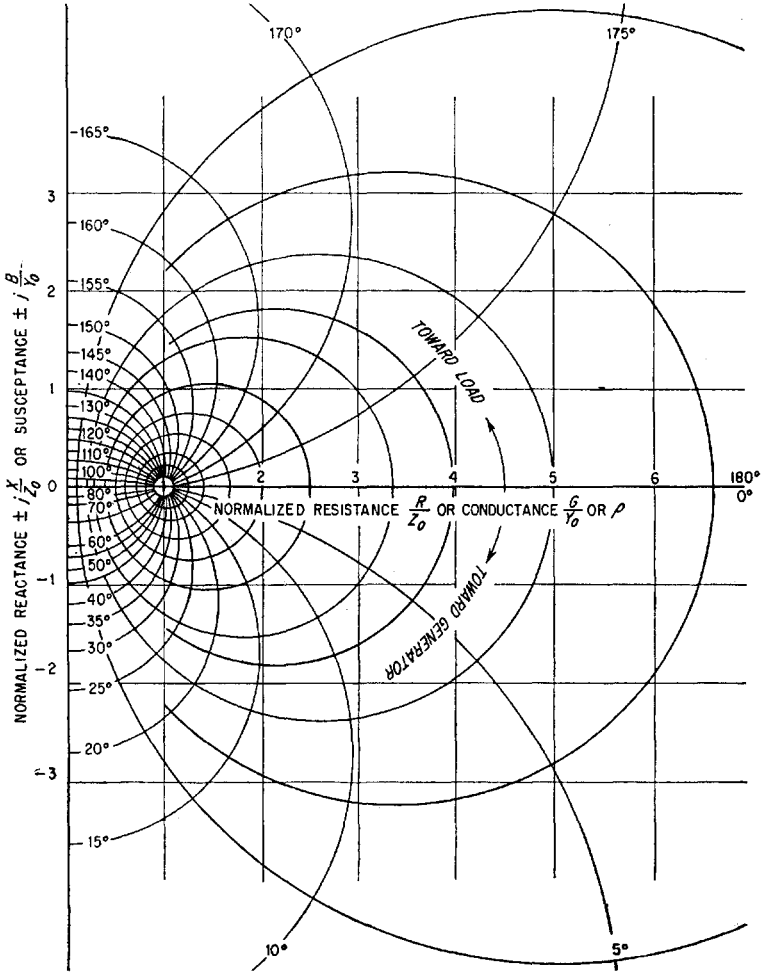
Fig. 20.8. Rectangular impedance or admittance chart.

The dielectric loss expressed by the second portion of Eq. (20.10) is the result of the continual reorientation of the molecules of the dielectric by the electric field. Attenuation due to losses in the dielectric is often difficult to compute from Eq. (20.10) because values for the conductance g per unit length of the dielectric are usually not available; however, a more useful expression is given by Eq. (20.62).

$$\alpha_d = \frac{\pi}{\lambda} \tan \zeta \quad \text{nepers per unit length} \tag{20.62}$$

where α_d = attenuation in nepers per unit length due to dielectric loss

λ = wavelength

tan ζ = loss "tangent" (loss tangent is approximately equal to power factor of dielectric)

Dielectric loss increases approximately proportionately with frequency, while copper loss increases as the square root of frequency. At lower frequencies copper losses usually predominate, and at higher frequencies dielectric losses become increasingly important.

20.1g. Transmission-line Diagrams. Several types of transmission-line diagrams have been devised to solve transmission-line problems graphically. The time and effort required to learn the significance of these diagrams and to master their use is justified in that through using such diagrams a majority of practical transmission-line problems can be solved rapidly without laborious calculations. The following discussion describes two of the most commonly used transmission-line charts.

Rectangular Impedance or Admittance Chart. A rectangular impedance or admittance chart is shown in Fig. 20.8. It is used to portray graphically impedance relationships at points along a transmission line. The following facts are useful in utilizing this chart.

1. The horizontal axis is the normalized resistive component of impedance (real part of Z/Z_o where Z is the impedance being plotted and Z_o is the characteristic impedance of the transmission line).

2. The vertical axis is the normalized reactive component of impedance (imaginary part of Z/Z_o).

3. Each circle corresponds to a constant voltage standing wave ratio ρ.

4. Curves orthogonal to the ρ circles correspond to a constant number of electrical degrees from a voltage minimum.

5. Maximum and minimum impedances occur at the intercepts of the constant ρ circles with the horizontal axis (pure resistances equal to ρZ_o and Z_o/ρ).

6. The point $\rho = 1$ corresponds to $Z = Z_o$.

7. The vertical axis corresponds to the circle $\rho = \infty$.

8. Clockwise rotation around the constant ρ circles is toward the generator end of the line.

Admittance, the reciprocal of impedance, can be obtained by rotating 90 electrical degrees of line length along the appropriate ρ circle from the point Z/Z_o and reading values directly off the chart as normalized conductance along the horizontal axis and normalized susceptance along the vertical axis. Points below the horizontal axis represent negative (inductive) susceptance, while points above represent positive (capacitance) susceptance. This chart does not account for dissipation in the transmission line.

Example 20.1

For a 100-ohm line terminated in $100 + j100$ ohms, determine the impedance one-eighth wavelength away from the termination and the equivalent admittance at this point.

Solution

1. Determine the normalized load impedance.
The normalized load impedance Z_R/Z_o is

$$\frac{Z_R}{Z_o} = \frac{R_R}{Z_o} + \frac{X_R}{Z_o} = 1 + j1$$

2. Obtain the impedance $\lambda/8$ from the termination.
Enter the chart at $R_R/Z_o = +1$, $X_R/Z_o = +1$. Read the standing wave ratio of 2.6 at this point and the phase angle of 148°. Traveling around the $\rho = 2.6$ circle clockwise 45°, read the normalized impedance components at this point of $R_x/Z_o = 2.0$, $X_x/Z_o = -1.0$.

The impedance at this point $\lambda/8$ from the termination is therefore

$$Z_x = Z_o \frac{Z_x}{Z_o} = 200 - j100$$

3. Determine the equivalent admittance.

To find the equivalent admittance rotate an additional 90°. The normalized conductance G_x/Y_o is 0.4, and the normalized susceptance B_x/Y_o is 0.2. The line admittance is therefore

$$Y_x = Y_o \frac{Y_x}{Y_o} = \frac{1}{Z_o} \frac{Y_x}{Y_o} = \frac{1}{100} (0.4 + j0.2) = 0.004 + j0.002$$
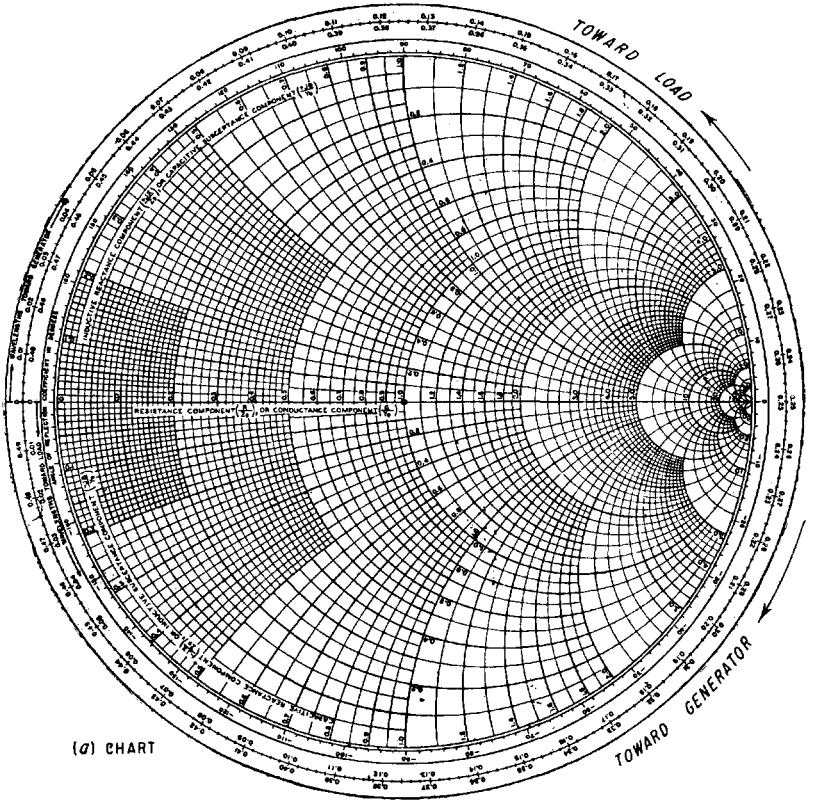
(a) CHART
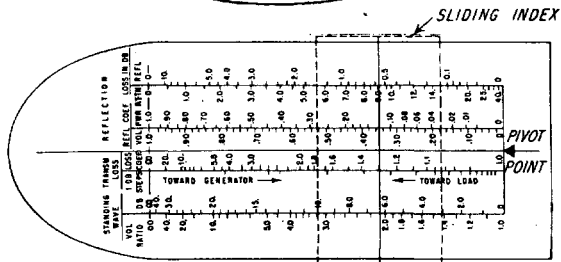
(b) RADIAL ARM TO BE PIVOTED ABOUT CENTER OF CHART

FIG. 20.9. Smith chart.

Reflection Coefficient Chart or Smith Chart.[1] The Smith chart is perhaps the most useful type of transmission-line chart. A conventional Smith chart is shown in Fig. 20.9, while an expanded version is shown in Fig. 20.10. The following paragraphs discuss the factors related by the Smith chart.

Normalized Impedance or Admittance. Normalized impedance Z/Z_o or admittance Y/Y_o is the actual impedance or admittance at any point on the transmission line
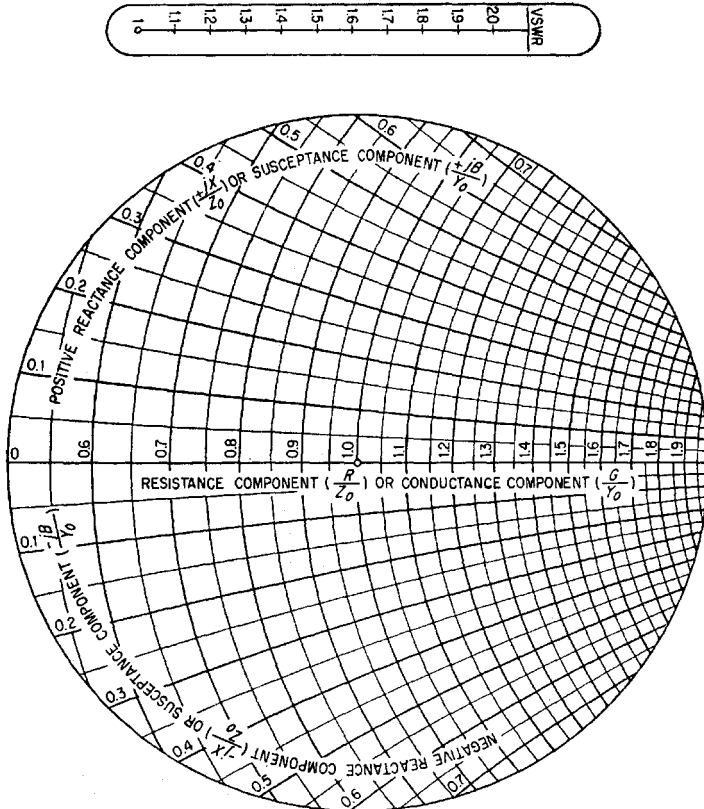
FIG. 20.10. Expanded Smith chart.

divided by the characteristic impedance or admittance of the line. Normalized impedances and admittances are used so that the chart is not limited by the characteristic impedance of the particular type of line under study. Actual impedance or admittance values are obtained by multiplying the normalized values by the characteristic impedance or admittance, respectively, of the line.

Points of normalized impedance are located by the intersection of the circle whose intercept along the horizontal axis is normalized resistance with the circle whose intercept along the outer periphery of the chart is normalized reactance.

Generally, the impedance or admittance at a point along a line is considered to be the impedance or admittance which would be measured looking toward the load end of

[1] Phillips H. Smith, Transmission Line Calculator, *Electronics*, January, 1939, and also An Improved Transmission Line Calculator, *Electronics*, January, 1944. A permanent plastic version or transparent paper versions of this chart similar to Fig. 20.9 can be purchased from the Emeloid Co., Hillside, N.J.

the line if the line were cut at that point. The impedance or admittance at any point completely describes the ratio of the magnitudes of the voltage and current and their relative phase angle. It should be noted that for steady-state conditions the generator impedance and the impedance looking toward the generator from the point at which the line is cut *do not* affect the distribution of voltage and current along the line. The generator impedance will affect the amount of power transferred to the line, however, and consequently will affect the amplitudes of voltage and current but not the ratio of maximum to minimum values.

The outer angular scale of the chart gives the distance between any two points in terms of fractions of a wavelength as measured toward the generator or toward the load. For example, if in Fig. 20.9 the sliding index of the radial arm is set across the coordinates corresponding to the normalized terminating impedance of the line, the normalized impedance (assuming negligible attenuation) one-eighth wavelength toward the generator will be found by rotating the arm clockwise 0.125 wavelength along the outer scale.

If the components of admittance are to be considered instead of impedance, the resistance axis (R/Z_o) becomes the conductance axis (G/Y_o) and the positive reactance scale $(+jX/Z_o)$ becomes the positive susceptance scale $(+jB/Y_o)$ while the negative reactance scale $(-jX/Z_o)$ becomes the negative susceptance scale $(-jB/Y_o)$. Capacitive susceptance is positive, and inductive susceptance is negative.

The normalized admittance corresponding to a given normalized impedance, or in general the reciprocal of any complex number, can be obtained by determining the coordinates of the point on the chart diametrically opposite the original point on the chart an equal distance from the center of the chart.

If the index is adjusted at the correct value of voltage standing wave ratio on the arm, the entire locus of impedances along the line will be obtained as the arm is rotated around the chart. These impedances repeat every 180°, neglecting attenuation, and have a maximum value (corresponding to a current minimum) when the index intersects the horizontal bisecting line (pure resistance) between 1 and ∞. A minimum impedance value (also a pure resistance at a voltage minimum) occurs where the index intersects the horizontal bisecting line between 0 and 1.

Reflection Coefficient. The magnitude of the reflection coefficient is read directly from the radial arm when the index is aligned over the normalized impedance at the point on the line under consideration. The phase angle of the reflection coefficient is read at the intersection of the radial arm and the angular scale along the periphery of the chart. This phase angle is the angle by which the reflected voltage wave lags the incident voltage wave. It should be noted that all normalized impedances along the radial arm correspond to the same phase angle of reflection coefficient. All normalized impedances around the chart, a given radial distance along the arm, correspond to a given amplitude of reflection coefficient, assuming the line has negligible attenuation.

Attenuation. The attenuation due to leakage, copper, and dielectric losses is plotted in 1-db steps along the radial arm. Normalized impedance values must be corrected when the attenuation is significant. This is accomplished by first calculating the total attenuation over the length of line under consideration from a knowledge of the attenuation per unit length (or wavelength) and the separation between points under consideration. The index is then moved outward or inward the required number of decibels as the arm is rotated, depending upon whether the direction is toward the load or toward the generator, respectively.

The scale along the radial arm labeled transmission *loss coefficient* gives a factor by which the calculated loss in decibels should be multiplied to account for the increased loss due to higher voltages and currents along the line when standing waves are present. This does not affect the standing wave ratio or line impedance. The transmis-

sion loss coefficient should be averaged over the length of the line under consideration if the normal attenuation is sufficient to cause the standing wave ratio to change appreciably.

The total reflection loss L_r is defined as

$$L_r = -10 \log_{10} \frac{P_{\text{absorbed}}}{P_{\text{incident}}} \tag{20.63}$$

or in terms of reflection coefficient,

$$L_r = 10 \log_{10} (1 - \Gamma_x{}^2) \tag{20.64}$$

This represents a real loss when attenuation is present to dissipate the energy of the reflected wave. The reflection loss is obtained from the reading on the arm when the index is positioned at the proper value of reflection coefficient. Equations (20.63) and (20.64) assume that there is no secondary reflection from the generator end of the line and that power not absorbed in the load is completely dissipated by the attenuation of the transmission line.

Example 20.2

A transmission line has a characteristic impedance of $50 + j0$ ohms and a voltage standing wave ratio of 2.0 measured near the load end of the line. A current minimum exists 0.35 wavelength from the load. The line is 9.5 wavelengths long and has an attenuation of 0.1 db per wavelength. Find (1) the load impedance, (2) the input impedance, and (3) the total attenuation in the line.

Solution

1. Find the load impedance:

a. Set the slide on the radial arm to a VSWR of 2.0.

b. Rotate the arm until the index intersects the R/Z_o axis at a current minimum (between 1 and ∞).

c. Rotate the radial arm 0.35λ counterclockwise (toward the load) and read the series components of impedance as $R/Z_o = 0.67$ and $jX/Z_o = -0.48$. As $Z_o = 50 + j0$, $Z_R = 34 - j24$.

2. Find the input impedance:

a. The line is 9.5 wavelengths long; therefore the line is an exact multiple of $\frac{1}{2}\lambda$ and the sending end impedance equals the terminating impedance, neglecting attenuation.

The total attenuation $= 0.1$ db per wavelength $\times 9.5$ wavelength $= 0.95$ db. With the slide aligned over Z_R/Z_o, move the slide 0.95 db "toward the generator" and read the corrected coordinates of sending end impedance as $R/Z_o = 0.75$, $X/Z_o = -j0.41$; therefore, $Z_S = 37.5 - j20.5$.

3. Find total insertion loss:

a. To determine the increased attenuation due to standing waves average the two values of "transmission-loss coefficient" obtained at the load end and the generator end. These are 1.26 and 1.16, respectively, therefore the total attenuation L_S corrected for standing wave ratio is

$$L_S = 0.95 \text{ db} \times \frac{1.26 + 1.16}{2}$$
$$= 1.15 \text{ db}$$

20.1h. Transmission Lines as Circuit Elements. Transmission lines are often used as circuit elements. For example, a re-examination of Fig. 20.5 will show that neglecting loss the input reactance of a line terminated in a short circuit varies from zero to plus infinity (inductive) as the line length varies from zero to $\lambda/4$. As the line is further lengthened, the input reactance varies from minus infinity (capacitive) to zero at $\lambda/2$. A similar situation occurs in the case of an open-ended line; however, the regions of inductive and capacitive reactance are reversed. The equations relating the reactance to the line length are repeated here for reference.

For a short-circuit termination,

$$Z_x = jZ_o \tan \beta x \qquad (20.53)$$

For an open-circuit termination,

$$Z_x = -jZ_o \cot \beta x \qquad (20.54)$$

Example 20.3

Provide an inductive reactance of $+j100$ ohms using both a short-circuited transmission line and an open-circuit transmission line having a characteristic impedance equal to 100 ohms and negligible attenuation.   The wavelength is 0.4 m.

Solution

$$\beta = 2\pi/\lambda = 2\pi/0.4 = 5\pi \text{ radians/meter}$$

From Eq. (20.53), for the short-circuit line

$$+j100 = j100 \tan (5\pi)x$$

therefore,

$$x = \frac{1}{5\pi} \tan^{-1} 1$$

$$x = \frac{1}{5\pi} \frac{\pi}{4} = \frac{1}{20} \text{ meter}$$

NOTE: Since $\lambda = \frac{2}{5}$ m, this corresponds to a line $\lambda/8$ long.
From Eq. (20.54), for the open-circuit line,

$$+j100 = -j100 \cot (5\pi)x$$

$$x = \frac{1}{5\pi} \cot^{-1} (-1)$$

$$x = \frac{1}{5\pi} \frac{3\pi}{4} = \frac{3}{20} \text{ meter} = \frac{3}{8} \lambda$$

There is exactly $\lambda/4$ difference between the length of the two lines for the same reactance. A $\lambda/8$ line or odd multiple thereof will give a reactance equal in magnitude to the characteristic impedance when the lines are terminated in a short or open circuit.

Transmission lines can also be used to simulate series or parallel resonant circuits. The short-circuited line presents an infinite impedance (assuming no losses) for lengths which are odd multiples of $\lambda/4$, while an open-circuited line presents zero impedance (again assuming no losses) for lengths which are odd multiples of $\lambda/4$.   If the termination is neither an open- nor a short-circuited, but is a pure reactance, the line length for zero or infinite impedance can be calculated by means of Eq. (20.52).

Example 20.4

Assume the line of the previous problem to be terminated in an inductive reactance of $+j100$.   The line length for an infinite input impedance is calculated as follows:
From Eq. (20.52)

$$Z_x = Z_o \frac{Z_R + Z_o j \tan \beta x}{Z_o + Z_R j \tan \beta x} = \infty$$

therefore

$$Z_o + Z_R j \tan \beta x = 0$$

or

$$j \tan \beta x = -\frac{Z_o}{Z_r} = +j$$

$$x = \frac{1}{5\pi} \tan^{-1} (1) = \frac{1}{20} \text{ meter} = \lambda/8$$

20.2. Open-wire Transmission Lines.   Open-wire lines are usually employed for power transmission at frequencies where radiation does not present a serious problem. The field configurations for open-wire lines are relatively simple, and the conventional concepts of voltage and current distribution as discussed in Sec. 20.1 can be applied.

To prevent considerable complication in the calculation of transmission-line constants, the following simplifying assumptions are made.

1. Skin effect is present.

2. The proximity effect is negligible; i.e., the spacing between conductors is at least 10 times the conductor diameter, yet small compared to the wavelength.

3. The conductor diameter is very small compared to the wavelength.

4. The height of the conductors above the ground is great compared to the separation.

5. The length of the line is long compared to the spacing.

Some general relationships applying to parallel-wire transmission lines are as follows:

$$Z_o = \sqrt{\frac{l}{c}} \quad \text{ohms} \tag{20.65}$$

$$\alpha = \frac{r}{2Z_o} \quad \text{nepers per unit length} \tag{20.66}$$

$$\beta = \omega \sqrt{lc} \quad \text{radians per unit length} \tag{20.67}$$

where Z_o is the characteristic impedance, l is in henrys per unit length, c is in farads per unit length, α is the attenuation constant in nepers per unit length, β is the phase constant in radians per unit length, and r is in ohms per unit length.
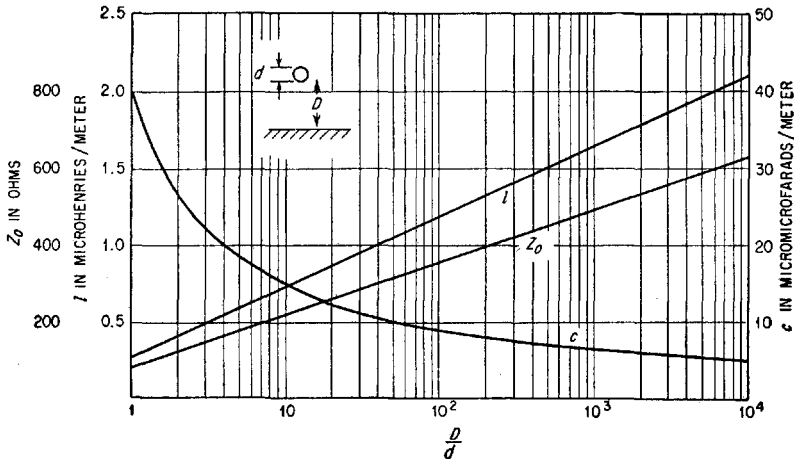
FIG. 20.11. Single-wire transmission line—characteristic impedance, inductance per unit length, and capacitance per unit length as a function of D/d.

20.2a. Single Wire Above Ground Plane. A single-wire line is shown in Fig. 20.11. The constants for this type of line are:

$$Z_o = 138 \log_{10} \frac{4D}{d} \quad \text{ohms} \tag{20.68}$$

$$l = 0.460 \log_{10} \frac{4D}{d} \quad \mu\text{h/m} \tag{20.69}$$

$$c = \frac{24.12}{\log_{10}(4D/d)} \quad \mu\mu\text{f/m} \tag{20.70}$$

$$r = \frac{8.3 \sqrt{f}}{d} \quad \mu\text{ohms/m for copper} \tag{20.71}$$

where d is in centimeters for Eq. (20.71), D,d in consistent units for Eqs. (20.68) to

(20.70), and f in cycles per second. Z_o, l, and c are shown in Fig. 20.11 as a function of D/d.

20.2b. Two-wire Transmission Lines.   A two-wire balanced transmission line is shown in Fig. 20.12a.   This type of line is perhaps the most conventional.   There is

(a) CONVENTIONAL   (b) PARALLEL LINES (MICROSTRIP) FOR
    TWO WIRE LINE   MICROWAVES
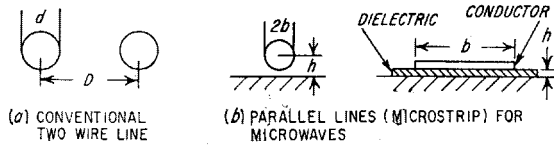
FIG. 20.12. Parallel-wire transmission lines.
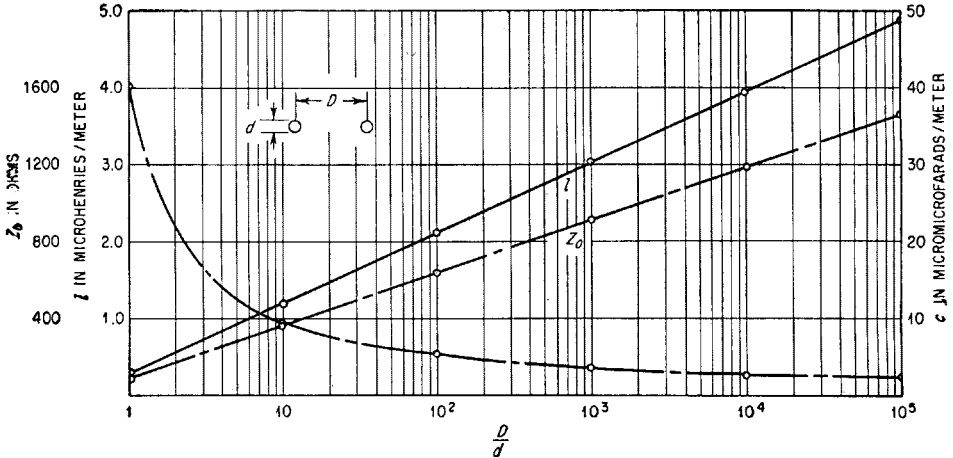
FIG. 20.13. Characteristic impedance, inductance per unit length, and capacitance per unit length for a parallel-wire transmission line as a function of D/d.

less radiation[1] than with the single wire line, and in addition, the ground currents are essentially zero.   The constants for this type of line assuming a well-developed skin effect are:

$$Z_o = 276 \log_{10} \frac{2D}{d} \qquad \text{ohms} \tag{20.72}$$

$$l = 0.921 \log_{10} \frac{2D}{d} \qquad \mu\text{h/m} \tag{20.73}$$

$$c = \frac{12.06}{\log_{10} (2D/d)} \qquad \mu\mu\text{f/m} \tag{20.74}$$

$$r = \frac{8.3 \sqrt{f}}{d/2} \qquad \mu\text{ohms/m for copper} \tag{20.75}$$

where d is in centimeters in Eq. (20.75). Z_o, l, and c are shown in Fig. 20.13 as a function of D/d.

The two-wire transmission line shown in Fig. 20.12a is usually restricted to frequencies at which the losses from radiation are negligible.   It has been recently shown[2] that an extension of parallel-wire transmission systems can be used up to

[1] For a treatment of radiation on open-wire lines, see E. J. Sterba and C. B. Feldman, "Transmission Lines for Short Wave Radio Systems," *Proc. IRE*, vol. 20, p. 1163, July, 1932.

[2] *Proc. IRE*, vol. 40, No. 12, 1952: Microstrip—A New Transmission Technique for the Kilomegacycle Range by D. D. Grieg and H. F. Englemann, Simplified Theory of Microstrip Transmission Systems by F. Assadourian and E. Rimai, Microstrip Components by J. A. Kostriza.

microwave frequencies. To reduce the radiation loss and provide usable line constants, the configurations shown in Fig. 20.12b are used. For large values of b/h essentially all of the energy is confined to the space immediately surrounding the line. The attenuation is greater than with waveguide but less than coaxial line using the same dielectric material. For small values of h/λ the radiated power per wavelength is a small fraction of the transmitted power. This type of line has reduced power capabilities and greater cross coupling between adjacent lines than other microwave structures. However, it is possible to greatly simplify microwave components by the use of this technique.[1]

20.2c. Four-wire Balanced Line. This type of line provides greater shielding and less radiation than the balanced pair. The constants for this type of line are given by the following:

$$Z_o = 138 \log_{10} \frac{D \sqrt{2}}{d} \qquad \text{ohms} \qquad (20.76)$$

$$l = 0.460 \log_{10} \frac{D \sqrt{2}}{d} \qquad \mu\text{h/m} \qquad (20.77)$$

$$c = \frac{24.1}{\log_{10} \dfrac{D \sqrt{2}}{d}} \qquad \mu\mu\text{f/m} \qquad (20.78)$$

$$r = \frac{8.3 \sqrt{f}}{d} \qquad \mu\text{ohms/m} \qquad (20.79)$$

where d is in centimeters in Eq. (20.79). Z_o, l, and c are shown in Fig. 20.14 as a function of D/d.
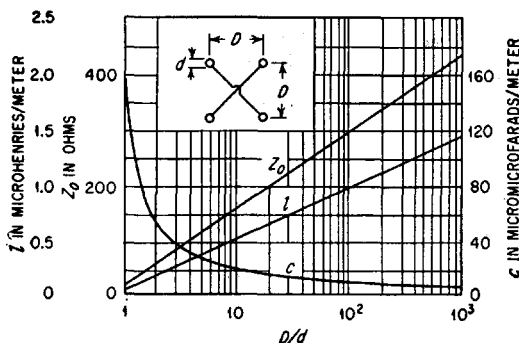
Fig. 20.14. Four-wire transmission line characteristics.

20.3. Coaxial Transmission Lines. Coaxial lines are used for signal and power transmission from low frequencies up to thousands of megacycles. The line currents can be considered to flow along the outside surface of the inner conductor and along the inside surface of the larger coaxial outer conductor, since the thickness of the conductors is almost always large compared to the depth of penetration of current into the conductors because of skin effect. Consequently, external radiation from coaxial transmission lines having a solid outer conductor is virtually nonexistent, and insulated structures supporting the outer conductor are not required. Coaxial lines can be either flexible with one or more layers of conducting braid for the outer conductor, or they can be rigid, using solid metallic tubing for the outer conductor. The inner conductor can be supported by a continuous dielectric, by spaced dielectric beads, or by "stubs" consisting of short sections of coaxial line in parallel with the main line.

[1] Licensed for use through the Internationl Telephone and Telegraph Company. See Section 20.5g.

The primary considerations in the design or use of a coaxial line are:

1. The conditions under which the line must operate such as temperature, humidity, shock, vibration, etc.

2. The "mode" of propagation to be utilized (this is usually the "dominant mode" or mode having the lowest cutoff frequency as is discussed in Sec. 20.3g)

3. The constants of the line such as characteristic impedance, series inductance, series resistance, shunt capacitance, shunt conductance, attenuation, and phase shift

4. The power carrying capacity

The most significant parameter in the design and use of coaxial lines is the ratio of the inside diameter D of the outer conductor to the outside diameter d of the inside conductor. As all of the above factors are not optimized at the same value of D/d, a compromise is necessary. For high-frequency applications (1,000 Mc and above), a characteristic impedance of 50 ohms is customarily used. This corresponds to a value of D/d equal to 2.3, assuming air dielectric between the conductors. A discussion of the manner in which the significant parameters of coaxial line operation vary with D/d is given in the following paragraphs. Unless noted otherwise, data are given for the dominant (TEM) mode.

20.3a. Coaxial-line Constants. The *characteristic impedance Z_o*, as discussed in Sec. 20.1b, is given by the general expression

$$Z_o = \sqrt{\frac{r + j\omega l}{g + j\omega c}} \qquad (20.6)$$

where r = resistance, ohms per unit length
g = conductance, mhos per unit length
c = capacitance, farads per unit length
l = inductance, henrys per unit length

The conductance g is almost always small compared to ωc, and at the frequencies usually employed r is small compared to ωl. Z_o then becomes essentially a pure resistance given by

$$Z_o = \sqrt{\frac{l}{c}} \qquad (20.8)$$

or in terms of dimensions of the line,

$$Z_o = 138 \log_{10} \frac{D}{d} \qquad \text{ohms} \qquad (20.80)$$

The above expression assumes the medium between the inner and outer conductor is air or another gas with an effective dielectric constant of unity. If the medium has a dielectric constant different from unity, Z_o is reduced by the factor $1/(\sqrt{\epsilon_r})$ where ϵ_r is the dielectric constant of the medium relative to air.

If the dielectric medium is not continuous, but is in the form of sections or beads supporting the inner conductor, the effective dielectric constant ϵ_r' is given by

$$\epsilon_r' = \frac{\epsilon_r t + s}{t + s} \qquad (20.81)$$

where t = length of dielectric section
s = spacing between sections

The *inductance per unit length l*, assuming a well-developed skin effect, is given by

$$l = 0.46 \log_{10} \frac{D}{d} \qquad \mu h/m \qquad (20.82)$$

The *capacitance per unit length* c with a homogeneous medium between the conductors is given by

$$c = \frac{24.1\epsilon_r}{\log_{10}(D/d)} \qquad \mu\mu f/m \qquad (20.83)$$

The loop *resistance per unit length* r is given by the sum of the resistances of the inner and outer conductors.

Inner conductor $\qquad\qquad r_d = \dfrac{1}{\pi\delta d\sigma}$ $\qquad\qquad\qquad$ (20.84)

Outer conductor $\qquad\qquad r_D = \dfrac{1}{\pi\delta D\sigma}$ $\qquad\qquad\qquad$ (20.85)

where δ = skin depth
$\quad\sigma$ = conductivity
If each conductor is copper

$$r = 8.3 \sqrt{f}\left(\frac{1}{d} + \frac{1}{D}\right) \qquad \mu ohms/m \qquad (20.86)$$

where f is in cycles per second and D and d are in centimeters. Z_o, l, and c are shown in Fig. 20.15 for an air-filled line as a function of D/d.
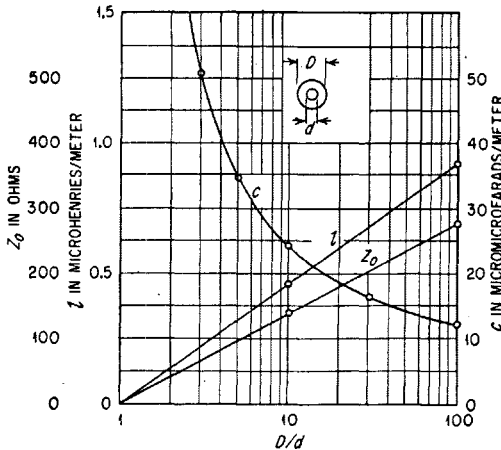
FIG. 20.15. Characteristic impedance, inductance, and capacitance as a function of the ratio D/d for a coaxial line.

20.3b. Coaxial Line Supports.[1] *Dielectric Supports.* The inner conductor of a coaxial line can be supported with dielectric material by several different methods. Where the dielectric loss is not prohibitive, solid dielectric lines are used. Such lines are free from internal reflections and, except for temperature effects upon the dielectric at extreme ambients, are very rugged mechanically. A lower characteristic impedance is obtained from a dielectric-filled line compared to an air-filled line with the same conductor dimensions.

The dielectric loss can be reduced through the use of dielectric beads spaced along the line, thereby decreasing the total amount of dielectric contributing to line loss. A simple bead as shown in Fig. 20.16a will produce a reflection in the line at the surface of the bead. Such reflections can be theoretically eliminated at one frequency by utilizing beads $\lambda/2$ long as shown in Fig. 20.16b. The input impedance of any half-wave section is equal to the terminating impedance of the half-wave section; con-

[1] Ragan, *op. cit.*, p. 155.

sequently, no mismatch occurs and there are no reflections.   It should be remembered the actual length must be corrected because of the decrease in wavelength within the dielectric.   The corrected length x' is given by

$$x' = \frac{\lambda_o}{2\sqrt{\epsilon_r}} \tag{20.87}$$

where λ_o = wavelength in air

Beads can also be spaced approximately $\lambda_o/4$, causing the reflections to add out of phase and cancel.   Transmission lines utilizing these two methods are frequency-sensitive, however, and large standing wave ratios can occur at frequencies other than the design frequency.

(a) SPACED DIELECTRIC BEADS   (b) HALF-WAVELENGTH BEAD   (c) UNDERCUT DIELECTRIC BEAD
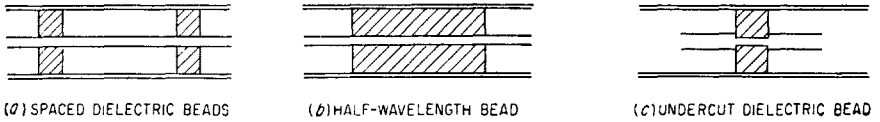
FIG. 20.16. Dielectric beads.

Undercut beads as shown in Fig. 20.16c can be used to avoid this frequency dependence.   The impedance of the dielectric section is adjusted by means of the D/d ratio to be equal to the characteristic impedance of the air-filled portion of the line.   This type of bead support provides relatively broadband operation.   The corrected value of inner conductor diameter d' is given approximately by

$$d' = \frac{d}{\sqrt{\epsilon_r}} \tag{20.88}$$

There are some constructional difficulties in assembling this type of coaxial line. In addition, the power-handling capacity of a beaded line is reduced because of high electric-field gradients produced at the sharp corners of the undercut section, and there is increased leakage over the face of the dielectric bead.

(a) SIMPLE STUB   (b) BROAD BAND STUB
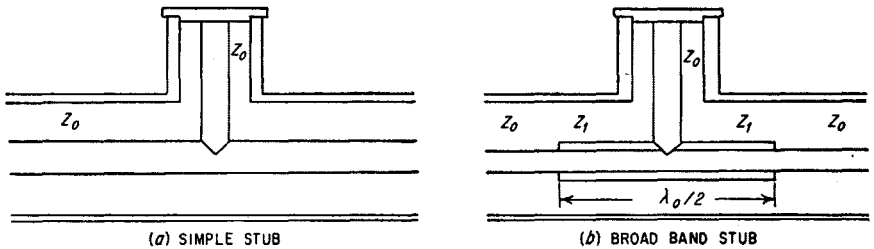
FIG. 20.17. Stub supports.

Because of these limitations, stub supports are frequently used in coaxial lines having rigid walls in preference to bead supports.   A stub is a section of coaxial line $\lambda/4$ long-shorted at one end as shown in Fig. 20.17a.   The input admittance of a $\lambda/4$ shorted line with no losses is zero, and since the reflection at such a junction is proportional to the stub admittance, the stub has no effect upon the operation of the line. This will only be true, however, at one frequency.   At frequencies different from the design frequency, the stub will no longer be a $\lambda/4$ section and reflection will occur at the junction.

A broadband stub support overcoming this difficulty is shown in Fig. 20.17b.   A transformer is added to the transmission line changing the characteristic impedance from Z_o to Z_1 for a distance of a quarter wavelength each side of the stub.   At the

design wavelength λ_o the transformer is $\lambda/2$ long, the stub is effectively $\lambda/4$ long, and no reflection occurs. At frequencies less than the design frequency, the transformer has a capacitive susceptance; however, the stub has an approximately equal inductive susceptance, and the parallel combination has a zero admittance at the junction, resulting in no reflection. In a similar manner, for frequencies greater than the design frequency the transformer has an inductive susceptance approximately canceled by the capacitive susceptance of the stub. Design equations applicable to wideband stubs are as follows:[1]

$$\tan^2 \left(\frac{\pi}{2} \frac{\lambda_o}{\lambda_1} \right) = \tan^2 \left(\frac{\pi}{2} \frac{\lambda_o}{\lambda_2} \right) = A^2 \tag{20.89}$$

$$\left(\frac{Z_1}{Z_o} \right)^3 + 2 \left(\frac{Z_1}{Z_o} \right)^2 + \frac{1}{A^2} \left(\frac{Z_1}{Z_o} \right) - 2 = 0 \tag{20.90}$$

where A = a constant determined from values of λ_o, λ_1, and λ_2

The frequencies corresponding to the wavelengths λ_1 and λ_2 are spaced the same percentage each side of the frequency corresponding to the design wavelength λ_o and are the frequencies at which minimum reflection will occur. Broadbanding over a wavelength variation of ± 10 per cent is possible by this method.

Because of fringe effects at the junction of the stub and the transformer section, the optimum length of the stub is not exactly $\lambda/4$ and consequently must be adjusted experimentally.

20.3c. Losses, Attenuation, and Phase Shift. From Sec. 20.1 the following general expression for the propagation constant γ is given by Eq. (20.3).

$$\gamma = \alpha + j\beta = \sqrt{(r + j\omega l)(g + j\omega c)} \tag{20.3}$$

The minimum possible value of the real part α of Eq. (20.3) is given by Eq. (20.91)

$$\alpha_{\min} = \sqrt{rg} \qquad \text{nepers per unit length} \tag{20.91}$$

and this is the d-c value of attenuation. At higher frequencies where $r \ll \omega l$ and $g \ll \omega c$ the following expression for α applies from Sec. 20.1b.

$$\alpha = \frac{r}{2Z_o} + \frac{gZ_o}{2} \qquad \text{nepers per unit length} \tag{20.10}$$

If the range of frequencies is such that $r \approx \omega l$ or $g \approx \omega c$, then the more cumbersome Eq. (20.3) must be used.

The conductor loss is

$$\alpha_c = \frac{4.34r}{Z_o} \qquad \text{db per unit length} \tag{20.92}$$

The expression for conductor attenuation α_c in terms of the line constants is, for a copper transmission line,

$$\alpha_c = \frac{\sqrt{f\epsilon_r}\,[(D/d) + 1]}{3.8D\,\log_{10}(D/d)}\,10^{-6} \qquad \text{db/m} \tag{20.93}$$

where f is in cycles per second and D and d are in centimeters.

For a fixed outer diameter, α_c has a minimum value at $D/d = 3.6$ which corresponds to $Z_o = 77$ ohms for an air-filled line. This minimum value of α_c is given by

$$\alpha_c = \frac{2.2\,\sqrt{f\epsilon_r}}{D}\,10^{-6} \qquad \text{db/m} \tag{20.94}$$

[1] From Radiation Laboratory Report 53-2 by R. V. Pound.

When the outer conductor is not solid but is braided to give flexibility, the attenuation in decibels per unit length due to the resistance of the outer conductor is multiplied by a factor of approximately 2.75.

To compute the dielectric loss, it is assumed the dielectric material can be described by a complex dielectric constant:

$$\epsilon_r = \epsilon' - j\epsilon''$$

and
$$\tan \zeta = \epsilon''/\epsilon' \qquad\qquad (20.95)$$

Tan ζ is commonly known as the "loss tangent." The conventional power factor of the dielectric is cos θ, where $\theta = 90° - \zeta$. Dielectric attenuation is then equal to

$$\alpha_d = \frac{27.3 \sqrt{\epsilon_r}}{\lambda_{\text{air}}} \tan \zeta \qquad \text{db per unit length} \qquad (20.96)$$

Phase Shift. The expression for the imaginary part β of the propagation constant γ is given approximately by Eq. (20.12).

$$\beta = \omega \sqrt{lc} \left[1 + \frac{1}{2}\left(\frac{r}{2\omega l} - \frac{g}{2\omega c} \right)^2 \right] \qquad \text{radians per unit length} \quad (20.12)$$

which reduces to the following expression for a lossless line

$$\beta = \omega \sqrt{lc} \qquad \text{radians per unit length} \qquad (20.13)$$

Equation (20.13) is generally applicable in high-frequency transmission lines. Where the losses are significant, Eq. (20.12) must be used.

20.3d. Maximum Power and Voltage Gradient. Assuming a fixed outer diameter and a matched line, maximum power can be transmitted on a coaxial line when $D/d = 1.65$, which corresponds to a characteristic impedance of 30 ohms for an air-filled line. Maximum voltage can be impressed on the line when D/d equals 2.718 corresponding to a characteristic impedance of 60 ohms.

A transmission line experiences its maximum voltage gradient at the inner conductor. Breakdown gradients are normally assumed to be between 20 to 30 kv/in., but irregularities, an appreciable standing wave ratio, dielectric structures, or other factors can cause accelerated breakdown. The gradient at the surface of the inner conductor is

$$\text{Gradient} = \frac{2E}{d \ln (D/d)} \qquad (20.97)$$

where E = potential between conductors

ln = logarithm to base ϵ

20.3e. Line Wavelength. If a dielectric other than air is used in a coaxial line, the wavelength in the dielectric relative to the wavelength in air is given by

$$\lambda_1 = \frac{\lambda_o}{\sqrt{\mu \epsilon_r}} \qquad (20.98)$$

where λ_1 = wavelength in dielectric

λ_o = wavelength in air

μ = permeability of medium relative to air (almost always = 1)

ϵ_r = dielectric constant of medium relative to air

20.3f. Q Considerations. A transmission line with small attenuation per wavelength and with a nondissipative termination can be considered to have a Q given by

$$\frac{1}{Q} = \frac{2\alpha}{\beta} = \frac{r}{\omega l} + \frac{g}{\omega c} \qquad (20.99)$$

Since $r \ll \omega l$ and $g \ll \omega c$, transmission lines can have Q values relatively high compared to lumped constant circuit elements.

The Q of a resonant section can also be approximated by Eq. (20.99).

20.3g. Coaxial-line Modes. The *modes* of a transmission line designate the distribution of the electric and magnetic fields. Specific modes are indicated by symbols

ELECTRIC FIELD ───── MAGNETIC FIELD
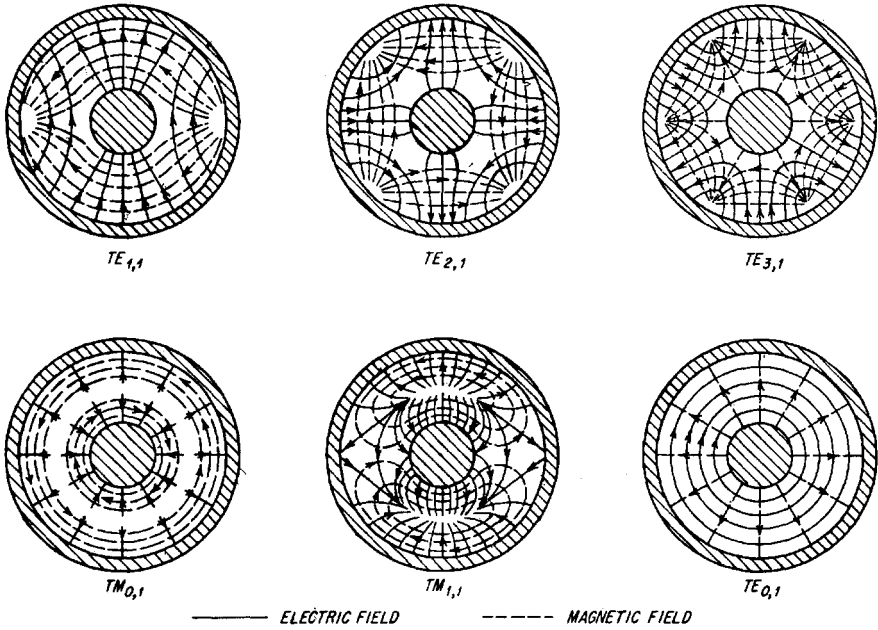
Fig. 20.18. *TE* and *TM* coaxial modes.

such as TEM, TE_{mn}, and TM_{mn}. The symbol TEM indicates that this mode has both electric and magnetic fields transverse to the axis of the line only. TE_{mn} indicates that modes of this classification have only the electric field transverse to the axis of the line. TM_{mn} indicates that modes of this classification have only the magnetic field transverse to the axis of the line.

The subscript m denotes the number of full period variations of the radial component of field in the angular direction. The subscript n denotes the number of half period variations of the angular component of field in the radial direction. Pictorial diagrams of the electric- and magnetic-field distribution transverse to the axis of a coaxial line are shown in Fig. 20.18 for the important TE and TM

── ELECTRIC FIELD ─── MAGNETIC FIELD

ELECTRIC AND MAGNETIC FIELDS DECREASE WITH RADIAL DISTANCE FROM CENTER CONDUCTOR

Fig. 20.19. *TEM* coaxial mode.

modes. These modes degenerate to similar modes for circular waveguide as the diameter of the inner conductor becomes vanishingly small. The dominant mode is the mode with the lowest cutoff frequency, i.e., the frequency below which the transmitted energy is rapidly attenuated and no useful transmission takes place. The dominant mode is almost always the only desirable mode, and except for special applications, an effort is made to prevent higher-order modes from being propagated. The dominant mode for coaxial lines is the TEM mode shown in

Fig. 20.19. It should be noted that the electric and magnetic fields are mutually perpendicular. It is also true that the current in the walls of the transmission line is perpendicular to the magnetic field.

The *TEM* mode has no lower cutoff frequency. The first higher-order mode is the TE_{11}. It can be propagated when the frequency is such that one wavelength in the medium between the conductors is approximately equal to the mean circumference C of the line. The cutoff frequency is equal to $C/\lambda_c \sqrt{\epsilon_r}$ where λ_c is determined from Fig. 20.20. The order in which higher-order modes will also be propagated as the frequency increases depends upon the D/d ratio. It follows that the frequency separation between modes is a function of D/d. This effect can be important in applications where it is required to separate
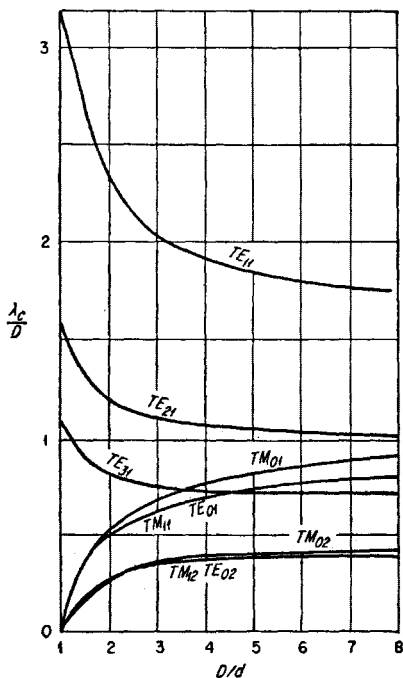
FIG. 20.20. Relationship between cutoff wavelength and line dimensions for higher-order coaxial modes. (*Adapted from Moreno, "Microwave Transmission Design Data," McGraw-Hill Book Company, Inc., New York, 1948.*)
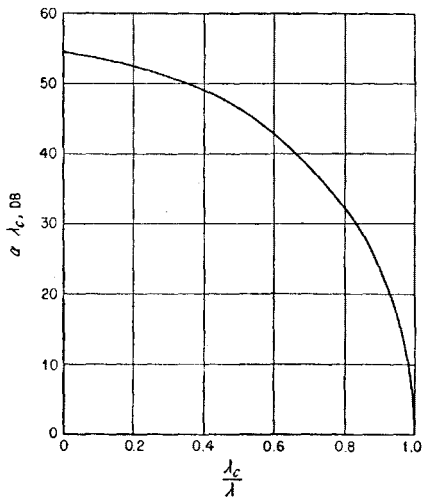
FIG. 20.21. Relationship between attenuation, cutoff wavelength, and operating wavelength in the cutoff region.

two higher-order modes. Figure 20.20 shows the order in which higher-order coaxial modes are propagated as a function of line dimension. The attenuation of any mode in the cutoff region is given by Eq. (20.100).

$$\alpha = \frac{54.5}{\lambda_c} \sqrt{1 - \left(\frac{\lambda_c}{\lambda}\right)^2} \quad \text{db per unit length} \tag{20.100}$$

where λ_c = cutoff wavelength of the mode

λ = actual wavelength

Equation (20.100) is shown graphically in Fig. 20.21.

The transmission-line constants such as characteristic impedance, attenuation, and phase velocity differ in the higher-order modes from the conventional values associated with the dominant mode.[1]

---

[1] For details see Radiation Laboratories Series, vol. 9, G. L. Ragan, "Microwave Transmission Circuits," sec. 2.10, McGraw-Hill Book Company, Inc., 1947

20.4. Coaxial Circuit Elements. Transmission lines are often called upon to perform many specialized circuit functions in addition to their conventional use as a means of conveying signal power at high frequencies. The following paragraphs discuss the specific coaxial configurations designed to provide the functions of impedance matching, impedance transformation, mode excitation, and mode filtering.[1]

20.4a. Impedance Elements. Impedance elements can be added to a coaxial transmission line in series with the line or in shunt with the line. Series elements are added in such a manner as to interrupt the circuit along the axis of the line and consequently are placed in series with either or both the inner and outer conductors. Shunt elements are added in such a manner as to cause a common electric field to exist between the element and the line. Consequently, shunt elements do not interrupt the entire line but branch the inner and outer conductors at one side of the line.

(*a*) SERIES ELEMENTS   (*b*) SHUNT ELEMENTS
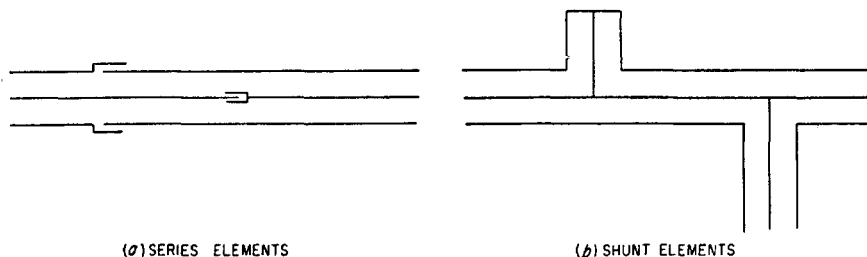
FIG. 20.22. Coaxial series and shunt elements.

Figure 20.22 shows both series and shunt elements added to a coaxial transmission line.

Stubs. Because of its versatility, the shorted stub is the most conventional form of shunt element. The shunt impedance presented to the transmission line can be inductive, capacitive, resonant, or antiresonant, depending upon the wavelength and the length of the stub. A plot of the relationship between reactance and stub length is shown in Fig. 20.5. Figure 20.5 assumes that the stub has no attenuation. If this is not the case, the minimum and maximum impedances corresponding to the resonant and antiresonant conditions are given by Eqs. (20.101) and (20.102).

$$Z_{min} = Z_o \coth \alpha x \simeq Z_o \alpha x \qquad \text{ohms} \qquad (20.101)$$

$$Z_{max} = Z_o \tanh \alpha x \simeq \frac{Z_o}{\alpha x} \qquad \text{ohms} \qquad (20.102)$$

where Z_o = characteristic impedance, ohms

    α = attenuation constant, nepers per unit length

    x = stub length in units consistent with α

Impedance Matching. One of the most useful functions of a coaxial stub is to match impedances. More specifically, stubs are used to cancel the effect of a mismatch, thereby eliminating the standing wave which would otherwise exist. It was previously shown that a stub can have either an inductive or capacitive input reactance depending upon its length. As the admittances of two elements connected in shunt add algebraically to form a resultant admittance, it follows that a stub in parallel with a coaxial line can be caused to cancel any reactive component of line impedance resulting from mismatch, etc. The stub susceptance must be equal in magnitude and opposite in sign to the line susceptance at the point of connection. The only additional requirement is to choose the point of attachment so that the real part of the

[1] For additional design data on these and other coaxial circuit elements see *ibid.* and vol. 10, N. Marcuvitz, "Waveguide Handbook," 1951.

impedance or admittance looking toward the load is equal to the characteristic impedance or admittance of the line.

Example 20.5

It is required to eliminate the standing wave on a lossless transmission line which has a characteristic impedance of 100 ohms and is terminated in a pure resistance of 50 ohms.
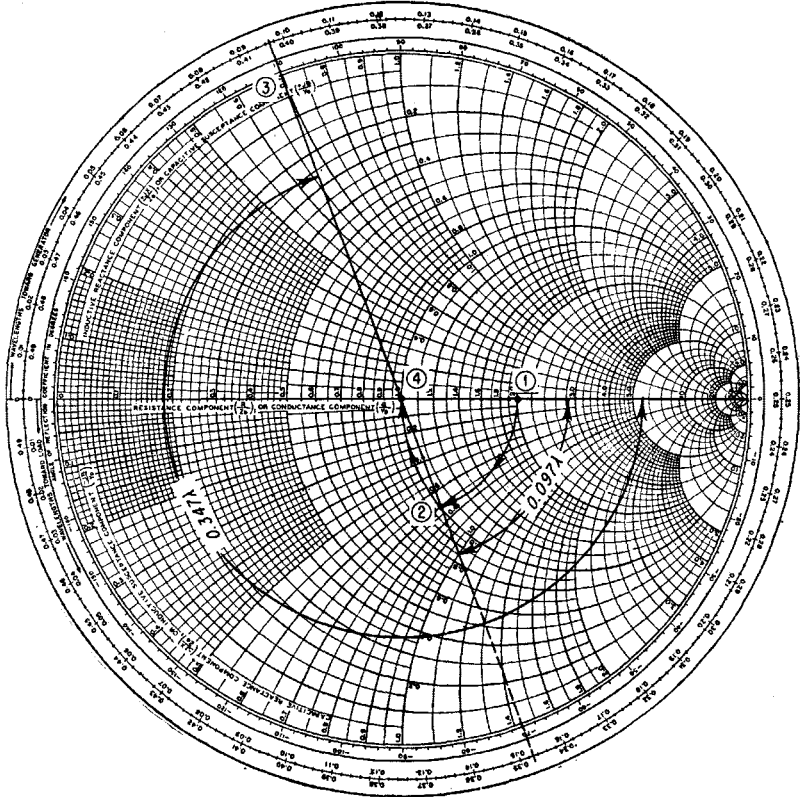
FIG. 20.23. Single-stub tuner design example.

Solution

1. Determine by the use of a Smith chart as shown in Fig. 20.23 the distance x from the load at which the real part of the normalized line admittance is unity.

$$Z_o = 100 \text{ ohms}$$

therefore

$$Y_o = 0.01 \text{ mhos}$$

The normalized receiving-end impedance is

$$\frac{Z_R}{Z_o} = \frac{50 + j0}{100} = 0.5 + j0 \text{ ohms}$$

and the normalized receiving-end admittance is

$$\frac{Y_R}{Y_o} = \frac{Z_o}{Z_R} = 2.0 + j0 \text{ ohms}'$$

The Smith chart is entered at this value of admittance which is point 1 in Fig. 20.23. The radial arm is rotated toward the generator until the index intersects the circle correspond-

ing to unity normalized conductance which is point 2. At this point the line length is 0.097λ, and the normalized susceptance is $1.0 - j0.7$ mhos.

2. Determine the length of shorted stub to give a positive (capacitive) susceptance of $j0.7$ mhos.

The normalized receiving-end (i.e., shorted-end) impedance of the stub is

$$\frac{Z_R}{Z_o} = \frac{0 + j0}{Z_o} = 0 \text{ ohms}$$

and the normalized receiving-end admittance is

$$\frac{Y_R}{Y_o} = \frac{Z_o}{Z_R} = \infty \text{ mhos}$$

Entering the chart at the extreme right end of the horizontal axis, which corresponds to infinite normalized conductance, the radial arm is rotated until the index intersects the $+j0.7$ intercept along the outer circumference at 0.347λ. This is point 3 on Fig. 20.25. The final design, therefore, is for a stub of length 0.347λ located 0.097λ from the termination. It can be seen that in this case an open-circuited stub a $\lambda/4$ shorter than the short-circuited stub could have also been used. As the coordinates of point 2 are $1 - j0.7$, it can be seen that adding the stub susceptance of $+j0.7$ will shift the Smith chart plot from point 2 to point 4 with coordinates $1 + j0$ for a perfect match.

It is often not convenient to allow the stub to slide along the line. This limitation can be overcome by the use of a double-stub section consisting of two stubs with fixed spacing and of fixed position along the line. The operation of a double-stub tuner is similar to a single stub in that the first stub, nearest the generator, is used to add susceptance of equal magnitude and opposite sign to the line susceptance at that point. The stub nearest the load, however, is used to cause the normalized conductance to equal unity at the first stub. The spacing between stubs is usually taken to be an odd multiple of an eighth wavelength, for example, 3$\lambda/8$. In terms of the Smith chart, the function of the stub nearest the load is to add a susceptance to the line so as to place the line admittance at the first stub somewhere on the unit conductance circle. The problem then is the same as for a single stub. It can be shown that two stubs with an odd-multiple eighth-wavelength spacing can only match loads to the characteristic admittance Y_o if the conductance of the load at the second stub is less than twice Y_o. As the stub spacing approaches $\lambda/2$, the range of conductances that can be matched increases; however, practical limitations make the odd $\lambda/8$ spacing a good compromise. It should be noted that either a $\lambda/4$ section of line between the second stub and the load when required or the addition of a third stub will allow any load to be matched, except a pure reactance. When three stubs are used they are spaced $\lambda/4$ apart.

The following example illustrates the design of a double-stub tuner.

Example 20.6

Determine the constants of a double-stub tuner as shown in Fig. 20.24 to eliminate the standing wave on a transmission line with Z_o of $100 + j0$ ohms which is terminated in an impedance of $100 - j100$ ohms. The stub spacing is 3/8λ (135°), and one stub is located $\lambda/2$ from the termination of the line. Neglect attenuation.

Solution

1. To simplify calculation, change impedance values to normalized admittance values. The characteristic admittance is

$$Y_o = \frac{1}{Z_o} = 0.01 \text{ mho}$$

The terminating admittance becomes

$$Y_R = \frac{1}{Z_R} = \frac{1}{100 - j100} = 0.005 + j0.005 \text{ mho}$$

The normalized terminating admittance becomes

$$y_r = \frac{Y_R}{Y_o} = 0.5 + j0.5$$

Because the termination is $\lambda/2$ from stub 1, the value of y_r referred to the position of stub 1 is also $0.5 + j0.5$.
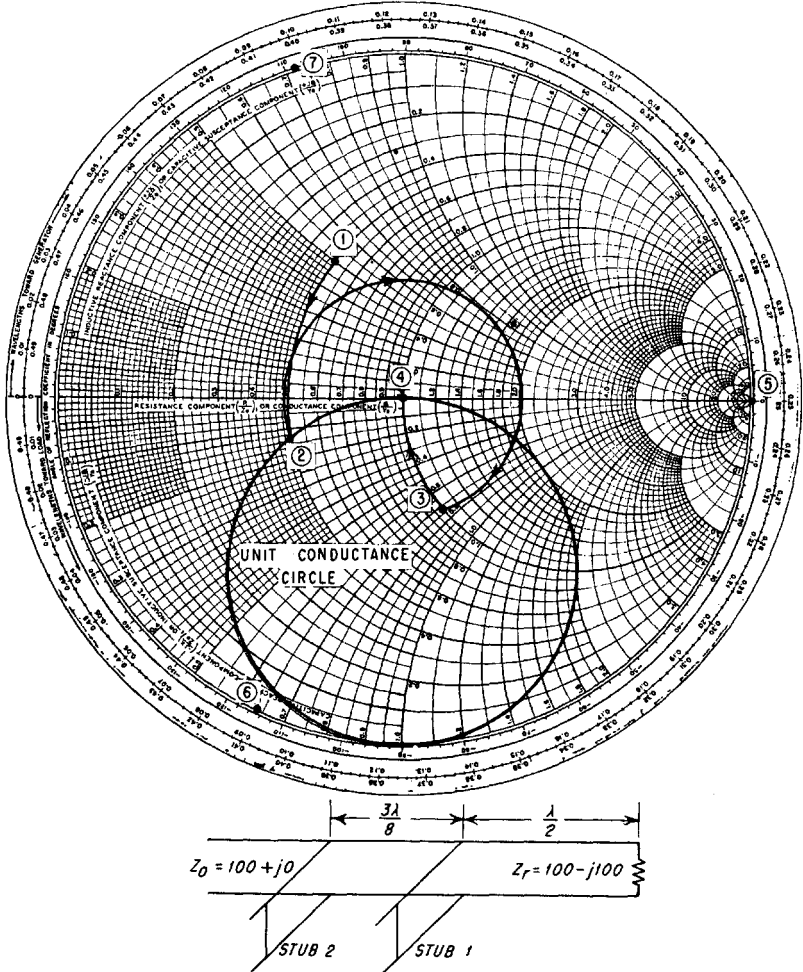
FIG. 20.24. Double-stub tuner design example.

2. To eliminate standing waves the normalized admittance at the junction of the line and the stub away from the load must equal $1 + j0$. This means the admittance at this point must be on the unit conductance circle of a Smith chart (refer to Fig. 20.24). The admittance at the junction of the stub nearest the load must be adjusted, therefore, so that it intersects a circle of diameter equal to the unit conductance circle which is rotated $3\lambda/8$ toward the load from the normal position of the unit conductance circle (see Fig. 20.24). This can be accomplished as follows:

a. Locate the normalized terminating admittance $0.5 + j0.5$ on the Smith chart.

b. Add the required normalized susceptance to move the point $0.5 + j0.5$ around the 0.5 conductance circle until it intersects the displaced unit conductance circle. This

requires an additional normalized susceptance of $-j0.63$. This is the required normalized susceptance of the first stub.

  c. Rotate the point located above $(0.5 - j0.13)$ $3\lambda/8$ away from the load. The point so located will be on the unit conductance circle $(1 - j0.72)$.

  d. The normalized admittance of the second stub must therefore be $+j0.72$ to bring the point along the unit conductance circle to the point $1 + j0$ required for perfect match.

The lengths of the stubs can be determined from Eq. (20.53) or by using the scales on the Smith chart as follows:

  (1) The reflection coefficient for a short-circuit termination is $1\underline{|180°}$. Therefore, enter the chart at point 5 (infinite conductance).

  (2) Rotate the radial scale "toward the generator" until it is aligned with the intersection of the required normalized susceptance line at the periphery of the chart.

  (3) The stub length in wavelengths can then be read directly from the outer scale of the Smith chart.

In this example, the stub lengths are 0.17λ for a susceptance of $-j0.63$ (point 6) and 0.349λ for a susceptance of $j0.72$ (point 7).

  20.4b. Impedance Transformation.[1] In addition to matching impedances, it is often necessary to transform from one impedance to another as in the case of a junction between two coaxial lines having different characteristic impedances. In general, transmission line impedance transformers do not have the frequency sensitivity inherent in lumped matching devices. The conventional types of transmission line impedance transformers are discussed in the following paragraphs.

  Tapered Lines. A tapered line is one in which the diameter of one or both of the conductors is varied so as to cause the characteristic impedance of a section of transmission line to vary continuously along its length. The taper is usually extended over several wavelengths to minimize reflections and frequency sensitivity. Tapered lines are constructed so that the inductance and capacitance per unit length vary logarithmically, one increasing and the other decreasing, causing the characteristic impedance also to vary logarithmically.

  Tapered lines are also constructed so that the ratio D/d varies linearly with length. The impedance change in this case is logarithmic, and again the line should be as long as possible to minimize undesirable reflection effects.[2]

  Quarter-wave Section. A quarter-wave line transforms impedances by the relationship $Z_o{}^2 = Z_1 Z_2$ where Z_1 and Z_2 are the impedances at either end of a $\lambda/4$ section of characteristic impedance Z_o. The section will not remain a quarter wavelength long if the frequency is altered and consequently is frequency-sensitive.[3] This frequency dependence can be greatly minimized by cascading two or more $\lambda/4$ sections having different characteristic impedances. The required impedance relationships for two sections are given in Eqs. (20.103) and (20.104).

$$(Z_o{}')^4 = (Z_1)^3(Z_2) \qquad\qquad (20.103)$$
$$(Z_o{}'')^4 = (Z_1)(Z_2)^3 \qquad\qquad (20.104)$$

where Z_1, Z_2 = impedances to be matched

  $Z_o{}'$ = characteristic impedance of $\lambda/4$ section adjacent to Z_1

  $Z_o{}''$ = characteristic impedance of $\lambda/4$ section adjacent to Z_2

A two-section transformer of this type is shown in Fig. 20.25.

  20.4c. Mode Excitation and Filtering. Higher-order modes are usually undesirable in coaxial lines. In some instances, however, higher-order modes are deliberately excited at one point in the transmission system and separated from the dominant

---

[1] For additional design data on the items discussed in this section see *ibid.*

[2] See H. J. Reich, P. F. Ordung, H. L. Krauss, and J. G. Skalnik, "Microwave Theory and Techniques," pp. 183–191, D. Van Nostrand Company, Inc., Princeton, N.J., 1953, for additional information.

[3] The frequency sensitivity is dependent on the ratio Z_1/Z_2, increasing as this ratio departs from unity.

mode at another point in the system. When any of these modes exist simultaneously with the dominant (TEM) mode, the actual fields existing in the line are the linear superposition of the fields corresponding to each mode. In general, the composite field will resemble neither mode separately.
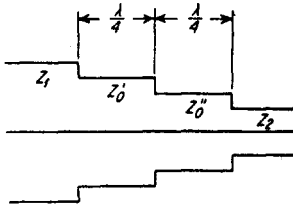
Fig. 20.25. Two-stage quarter-wave transformer.

In general, either the electric or magnetic field can be used to excite a given mode. If the electric field is to be excited, one or more probes must be oriented so as to be aligned with the direction of each maximum of the electric field as shown in Fig. 20.26a. If the magnetic field is to be excited, a loop must be oriented so as to align the plane of the loop normal to the direction of the magnetic field as shown in Fig. 20.26b. When higher-order modes are to be excited, the line dimensions must be such that the line is not in the cutoff region for these modes.

Mode filtering or suppression can be accomplished in a variety of ways. One method is to attenuate higher-order modes by restricting the line dimensions so that the line is below cutoff for these modes. Obstacles placed on the line will also serve to separate modes. In general, a metal obstacle should be placed so that the metal surface is parallel to the maximum electric field of the mode to be eliminated and perpendicular to the maximum electric field of the mode to be passed. Fig. 20.26c and d show methods for filtering higher-order modes.
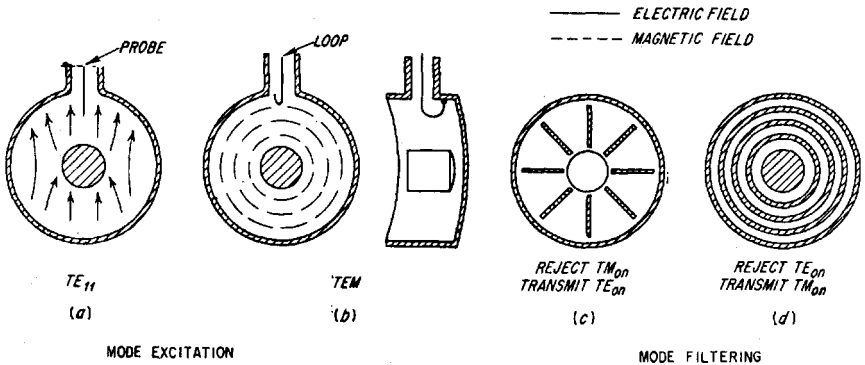
Fig. 20.26. Coaxial mode excitation and filtering.

20.4d. Baluns.[1] Occasionally, it is necessary to convert from a coaxial line (unbalanced line) to a two-wire transmission line (balanced line). A *balun* is used for the purpose as shown in Fig. 20.27. Configurations *a* and *b* in this figure are narrow-band baluns. Configuration *c* has a wider bandwidth than *a* or *b*. Configuration *d* is a folded version of *c* with the same bandwidth, but only half as long.

20.4e. Coaxial Attenuators and Terminations.[2] Coaxial attenuators are used for isolation between components or to obtain a reduction of power level on a transmission line. Coaxial terminations are used to provide a nonreflecting load at the end of a section of coaxial lines. Representative attenuators and terminations are shown in Fig. 20.28. The "below cutoff" attenuators shown in Fig. 20.28a and b are sections of circular waveguides with dimensions such that the wavelength in the guide is

[1] For design data on Baluns, see "Very High Frequency Techniques," vol. 1, pp. 86-92 and 382-390, McGraw-Hill Book Company, Inc., New York, 1947.

[2] C. G. Montgomery, "Technique of Microwave Measurements," pp. 679-803, Radiation Laboratory Series, vol. 11, McGraw-Hill Book Company, Inc., New York, 1947.

below the cutoff wavelength. For this condition, the attenuation is given by Eq. (20.105).

$$\alpha = \frac{2\pi}{\lambda_c} \sqrt{1 - \left(\frac{\lambda_c}{\lambda}\right)^2} \qquad \text{neper per unit length} \qquad (20.105)$$

Equation (20.105) is valid for TE and TM modes in either circular or rectangular waveguides.
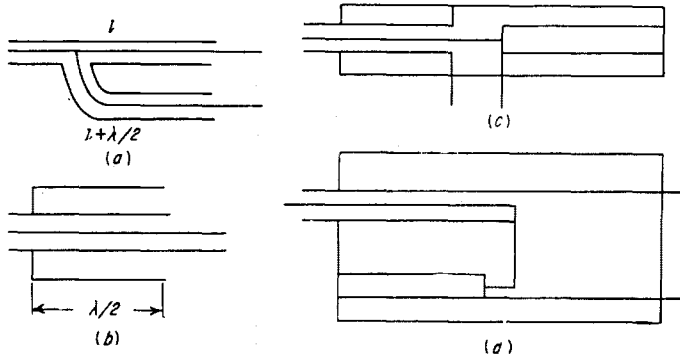
FIG. 20.27. Balun configurations.
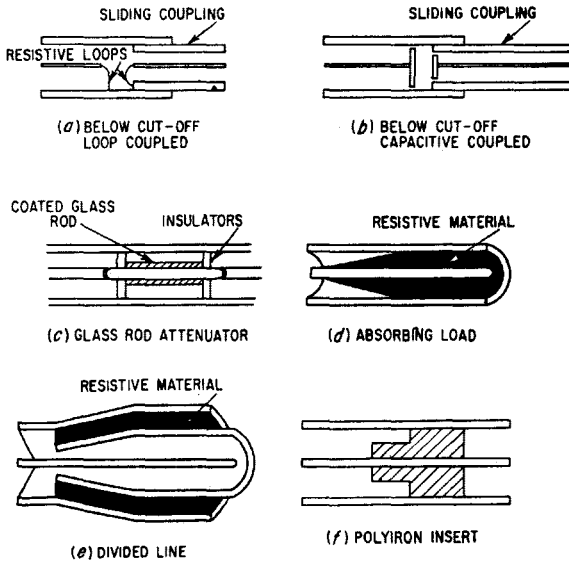
FIG. 20.28. Coaxial line attenuators and terminations.

The attenuator can be made variable by changing the length of the waveguide sections.

The coupling between the waveguide and the coaxial line can be a resistive loop as shown in Fig. 20.28a or a capacitive disk as shown in Fig. 20.28b. The loop is made resistive to terminate properly the coaxial line and can be formed from either a very small fixed resistor or a strip of resistive material. The metallic capacitive disk in Fig. 20.28b is often preceded in the coaxial line by a second disk of resistive material to terminate properly the coaxial-line impedance.

Resistive material can be used directly as shown in the glass-rod attenuator of Fig. 20.28c and the load of Fig. 20.28d. Glass rods can be metalized or coated with a resistive material such as Aquadag. Terminations can be made from mixtures of sand with lossy material. Polyiron can be used as an attenuating insert as shown in Fig. 20.28f. The step aides in matching the polyiron section to the input coaxial line. Polyiron has the advantage of possessing high attenuation at microwaves, and therefore only a small insert is required. In addition, it is rugged mechanically. It has the disadvantage of being hard to machine, it often varies in characteristics, and the dimensions of the insert are usually critical if broadband operation is desired.

Another approach is shown in the divided line attenuator of Fig. 20.28e. In this case the power is divided between the concentric coaxial lines and absorbed in the outer coaxial line by the use of resistive material.

Additional variations of these types are also used; for example, the glass-rod attenuator can be made variable by sliding a section of the inner conductor of the coaxial line over the resistive coating on the glass rod.

A limitation of attenuators and terminations which must be carefully considered is the power-handling capacity of the resistive material, which is often quite low.

20.5. Waveguide Transmission Lines. Waveguides are usually the most desirable form of transmission line for applications at wavelengths shorter than approximately 10 cm. At longer wavelengths the size becomes increasingly prohibitive. At wavelengths shorter than about 1 cm the small size of the waveguide becomes a practical limitation. Basically, the advantage of waveguides over coaxial lines of comparable size is their greater power-handling capacity and lower attenuation. Along with waveguide transmission lines there have been developed a large number of waveguide circuit elements and special devices making it possible to duplicate at microwaves most of the circuit configurations used at lower frequencies.

Although the lower-frequency concepts of voltage and current are usually replaced by electric- and magnetic-field concepts when considering waveguides, quantities such as standing waves, reflection coefficient, propagation constant, etc., are also applicable to waveguide transmission lines.

20.5a. Waveguide Modes. The modes of a waveguide transmission line designate the distribution of the electric and magnetic fields. Specific modes are indicated by symbols such as TE_{mn} and TM_{mn}. The symbol TE indicates the electric field is everywhere transverse to the axis of the transmission line. The symbol TM indicates the magnetic field is everywhere transverse to the axis of the transmission line. In addition, the electric and magnetic fields are orthogonal.

—— ELECTRIC FIELD   – – MAGNETIC FIELD

Fig. 20.29. TE and TM rectangular waveguide modes.

For rectangular waveguide, the subscript m denotes the number of maxima of electric field along the wide or a dimension as shown in Fig. 20.29. The subscript n denotes the number of maxima of electric field along the narrow or b dimension.

Figure 20.29a shows the dominant mode for rectangular waveguide. This is the mode with the longest cutoff wavelength. This subscript notation for rectangular waveguides applies for both TE and TM modes.

For circular waveguide, the subscript designations are similar to those used for coaxial lines. The subscript m denotes the number of full period variations of radial component of field in the angular direction. The subscript n denotes the number of half period variations of angular component of field in the radial direction. This subscript notation applies for either TE or TM modes. This is illustrated in Fig. 20.30.

$TM_{0,1}$
$\lambda_0 = \dfrac{c}{2.405}$
(a)

$TE_{2,1}$
$\lambda_c = \dfrac{c}{3.054}$
(b)

$TE_{1,1}$ (DOMINANT)
$\lambda_c = \dfrac{c}{1.841}$
(c)

$TM_{1,1}$
$\lambda_c = \dfrac{c}{3.832}$
(d)

$TE_{0,1}$
$\lambda_c = \dfrac{c}{3.832}$
(e)

——— ELECTRIC FIELD
- - - - MAGNETIC FIELD
c = CIRCUMFERENCE OF
       CIRCULAR WAVE GUIDE
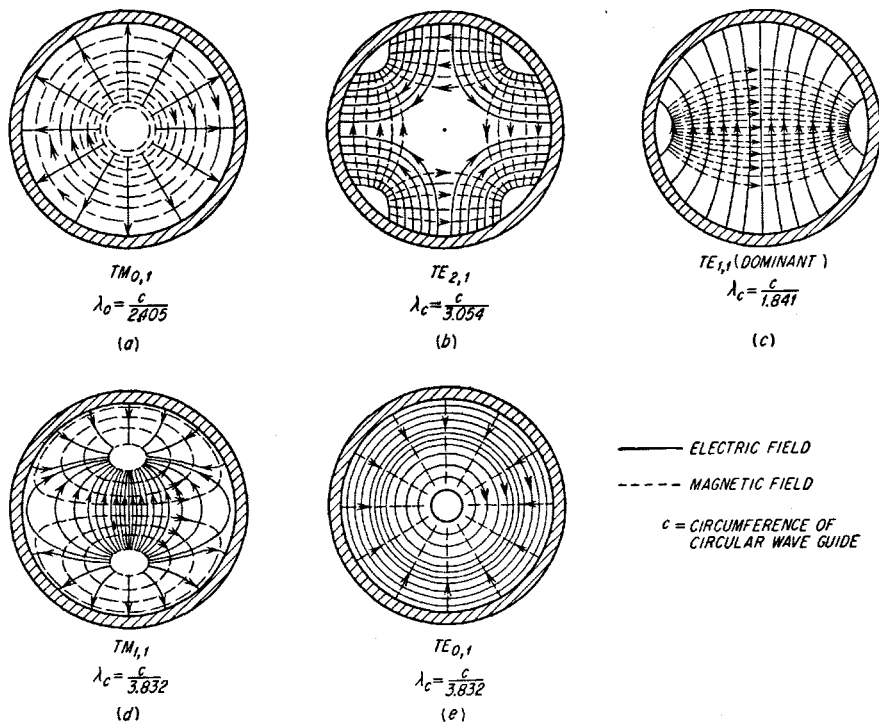
FIG. 20.30. TE and TM circular waveguide modes.

Field configurations for several higher-order modes of rectangular and circular waveguides are shown in Figs. 20.29 and 20.30. For convenience, the cutoff wavelengths are also given.

The general expressions for the cutoff wavelength λ_c of rectangular and circular waveguides are given by Eqs. (20.106) and (20.107), respectively.

$$\lambda_c = \frac{2}{\sqrt{\left(\dfrac{m}{a}\right)^2 + \left(\dfrac{n}{b}\right)^2}} \qquad (20.106)$$

where m, n are the mode subscripts and a, b are the rectangular waveguide dimensions from Fig. 20.29.

$$\lambda_c = KD \qquad (20.107)$$

where D is the circular waveguide diameter. KD is given in Table 20.1.

TABLE 20.1. CIRCULAR WAVEGUIDE CUTOFF WAVELENGTHS

TE		m			m			TM	
		0	1	2	2	1	0		
n	1	0.82D	1.7D	1.03D	0.61D	0.82D	1.3D	1	n
	2	0.45D	0.59D	0.47D	0.38D	0.45D	0.57D	2	

At wavelengths longer than the cutoff wavelength for each mode, the mode will be rapidly attenuated. The attenuation below cutoff for rectangular and circular guides is given by

$$\alpha_c = 8.69 \sqrt{\left(\frac{2\pi}{\lambda_c}\right)^2 - \epsilon_r \left(\frac{2\pi}{\lambda}\right)^2} \quad \text{db per unit length} \quad (20.108)$$

or $\qquad \alpha_c \approx \dfrac{54.6}{\lambda_c} \quad$ db per unit length for $\lambda \gg \lambda_c$

Normally, waveguides are not operated near cutoff for any desirable mode. The group velocity v_g and phase velocity v_p are rapidly changing functions of frequency in this region as shown in Fig. 20.31. The analytical relationships for the curves shown in the figure are as follows:

$$v_g = v \sqrt{1 - \left(\frac{f_c}{f}\right)^2} = \frac{v^2}{v_p} \quad (20.109)$$

where v = velocity of light in unbounded medium

f_c = frequency of cutoff

In general the guide wavelength is greater than the free-space wavelength. The dependence of guide wavelength λ_g upon cutoff wavelength λ_c is given in Eq. (20.110).

$$\lambda_g = \frac{\lambda}{\sqrt{\epsilon_r - \left(\frac{\lambda}{\lambda_c}\right)^2}} \quad (20.110)$$

where ϵ_r = dielectric constant of medium relative to air

λ = free-space wavelength
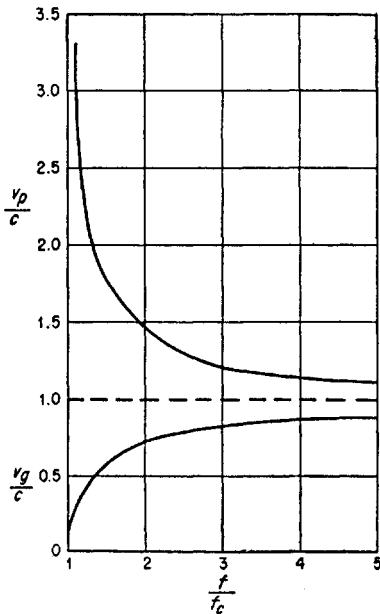
FIG. 20.31. Group and phase velocity as a function of the ratio of operating to cutoff frequency.

20.5b. *Attenuation.* The attenuation caused by conductor losses in a waveguide transmission system depends upon the distribution of current in the walls of the guide and consequently is different for each mode. In general, the attenuation decreases from a very high value at frequencies below the cutoff frequency to a broad minimum for frequencies several times the cutoff frequency. For frequencies considerably above the cutoff frequency, the attenuation increases as the frequency is increased. Representative values of attenuation of rectangular waveguides are given in Table 20.2 along with other pertinent data.

TABLE 20.2. ARMED SERVICES STANDARD LIST OF RIGID RECTANGULAR WAVEGUIDES

Jan type	Material	Dimensions (inches) Inside Width	Height	Tolerance	Outside Width	Height	Tolerance	Wall thickness Nominal	Deviation from mean[1]	Maximum inner corner radius[2]	Recommended operating range for TE_{10} mode Frequency, kmc/sec	Wavelength, cm	Cutoff for TE_{10} mode Frequency, kmc/sec	Wavelength, cm	Range in $2\lambda/\lambda_c$	Range in λ_g/λ	Theoretical attenuation lowest to highest frequency,[3] db/100 ft	Theoretical cw power rating lowest to highest frequency,[4] megawatts
RG-69/U... Brass RG-103/U.. Aluminum		6.500	3.250	±0.005	6.660	3.410	±0.005	0.080	±0.008	3/64	1.12–1.70	26.79–17.65	0.908	33.04	1.62–1.07	1.70–1.18	0.317–0.212 0.269–0.178	11.9–17.2
RG-104/U.. Brass RG-105/U.. Aluminum		4.300	2.150	±0.005	4.460	2.310	±0.005	0.080	±0.008	3/64	1.70–2.60	17.65–11.54	1.375	21.82	1.62–1.06	1.70–1.18	0.588–0.385 0.501–0.330	5.2–7.5
RG-48/U... Brass RG-75/U... Aluminum		2.840	1.340	±0.005	3.000	1.500	±0.005	0.080	±0.006	3/64	2.60–3.95	11.54–7.60	2.080	14.42	1.60–1.05	1.67–1.17	1.102–0.752 0.940–0.641	2.2–3.2
RG-49/U... Brass RG-95/U... Aluminum		1.872	0.872	±0.005	2.000	1.000	±0.005	0.064	±0.004	1/32	3.95–5.85	7.60–5.13	3.155	9.51	1.60–1.08	1.67–1.19	2.08 –1.44 1.77 –1.22	1.4–2.0
RG-50/U... Brass RG-106/U.. Aluminum		1.372	0.622	±0.004	1.500	0.750	±0.004	0.064	±0.004	1/32	5.85–8.20	5.13–3.66	4.285	7.00	1.47–1.05	1.48–1.17	2.87–2.30 2.45–1.94	0.56–0.71
RG-51/U... Brass RG-68/U... Aluminum		1.122	0.497	±0.004	1.250	0.625	±0.004	0.064	±0.004	1/32	7.05–10.00	4.25–3.00	5.260	5.70	1.49–1.05	1.51–1.17	4.12–3.21 3.50–2.74	0.35–0.46
RG-52/U... Brass RG-67/U... Aluminum		0.900	0.400	±0.003	1.000	0.500	±0.003	0.050	±0.004	1/32	8.20–12.40	3.66–2.42	6.560	4.57	1.60–1.06	1.68–1.18	6.45–4.48 5.49–3.83	0.20–0.29
RG-91/U... Brass RG-107/U.. Silver		0.622	0.311	±0.002	0.702	0.391	±0.003	0.040	±0.003	1/64	12.40–18.00	2.42–1.67	9.490	3.16	1.53–1.06	1.55–1.18	9.51–8.31 6.14–5.36	0.12–0.16
RG-53/U... Brass RG-66/U... Silver		0.420	0.170	±0.002	0.500	0.250	±0.003	0.040	±0.003	1/64	18.00–26.50	1.67–1.13	14.080	2.13	1.57–1.06	1.60–1.18	20.7–14.8 13.3–9.5	0.043–0.058
RG-96/U...	Silver	0.280	0.140	±0.0015	0.360	0.220	±0.002	0.040	±0.002	1/64	26.50–40.00	1.13–0.750	21.100	1.423	1.59–1.05	1.65–1.17	21.9–15.0	0.022–0.031
RG-97/U...	Silver	0.224	0.112	±0.0010	0.304	0.192	±0.002	0.040	±0.002	0.010	33.00–50.00	0.909–0.600	26.350	1.138	1.60–1.05	1.67–1.17	31.0–20.9	0.014–0.020
RG-98/U...	Silver	0.148	0.074	±0.0010	0.228	0.154	±0.002	0.040	±0.002	0.008	50.00–75.00	0.600–0.400	39.900	0.752	1.60–1.06	1.67–1.17	52.9–39.1	0.0063–0.0090
RG-99/U...	Silver	0.122	0.061	±0.0005	0.202	0.141	±0.002	0.040	±0.002	0.006	60.00–90.00	0.500–0.330	48.400	0.620	1.61–1.06	1.68–1.18	93.3–52.2	0.0042–0.0060

[1] The mean wall thickness is defined as one-half the difference between corresponding inside and outside dimensions as measured at any cross section perpendicular to the longitudinal axis.
[2] Minimum outer corner radius—1/64 in., maximum outer corner radius—1/32 in.
[3] Resistivity of brass—7.0 × 10⁻⁶ ohm-cm, silver—1.62 × 10⁻⁶ ohm-cm, aluminum—2.83 × 10⁻⁶ ohm-cm.
[4] Breakdown of air—15,000 volts/cm (safety factor of approximately 2 at sea level).

TRANSMISSION LINES

If a material other than copper is used, the attenuation values in the table will be multiplied by the square root of the ratio of the resistivity of the new material to the resistivity of copper.

If a waveguide is filled with dielectric material other than air, the attenuation due to conductor losses is increased.

In addition to the effect on conductor loss, the presence of a dielectric material other than air inside a waveguide introduces an additional attenuation given by Eq. (20.111) for TE and TM modes.

$$\alpha_D = \frac{8.38\pi\lambda_g}{\lambda_o^2} \tan \zeta \qquad \text{db/m} \qquad (20.111)$$

where λ_g = guide wavelength in dielectric material

ζ = loss tangent of dielectric material

λ_o = wavelength in unbounded dielectric material

The increased attenuation due to the dielectric loss will be much larger than the increased conductor losses caused by the presence of the dielectric.

20.5c. Phase Constant.[1] The numerical value of the phase constant of a transmission line is dependent upon the mode of propagation, the dimensions of the line, and the frequency. The following expressions give the phase constant of rectangular and circular waveguides.

For rectangular guide

TE_{mn} and TM_{mn},
$$\beta_{mn}^2 = \frac{\omega^2}{c^2} - \pi^2 \left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right) \qquad (20.112)$$

For circular guide

TE_{11},
$$\beta^2 = \frac{\omega^2}{v^2} - \left(\frac{3.682}{D}\right)^2 \qquad (20.113)$$

TE_{01},
$$\beta^2 = \frac{\omega^2}{v^2} - \left(\frac{7.664}{D}\right)^2 \qquad (20.114)$$

TM_{01},
$$\beta^2 = \frac{\omega^2}{v^2} - \left(\frac{4.81}{D}\right)^2 \qquad (20.115)$$

where v = velocity of light in an unbounded medium

It can be seen from these expressions that at very high frequencies the phase constant becomes almost proportional to frequency and from Sec. 20.2d the group and phase velocities approach the velocity of light. At lower frequencies where the second factor in each expression is significant, group and phase velocities differ considerably.

20.5d. Characteristic Impedance. The value of the characteristic impedance of a waveguide transmission line cannot be uniquely determined in terms of conventional voltage, current, and power relationships. This is partially because the currents and voltages are nonuniform and do not follow the distribution normally assumed in transmission-line theory. Several different bases are used in practice for determining waveguide impedance; however, they yield different numerical values. A factor common in each expression for impedance is the wave impedance Z_w of free space given by Eq. (20.116).

$$Z_w = \sqrt{\frac{\mu_o}{\epsilon_o}} = 377 \text{ ohms} \qquad (20.116)$$

where μ_o and ϵ_o are the permeability and dielectric constant (permittivity), respectively, of free space and Z_w can also be regarded as the ratio of electric- and magnetic-field intensities.

---

[1] See T. Moreno, "Microwave Transmission Design Data," McGraw-Hill Book Company, Inc., New York, 1948.

If the transmission system involves a medium with a permeability relative to air μ_r and a dielectric constant relative to air ϵ_r, the wave impedance Z_w for the *TEM* mode becomes

$$Z_w' = 377 \sqrt{\frac{\mu_r}{\epsilon_r}} \qquad \text{ohms} \qquad (20.117)$$

and the wave impedances for *TE* and *TM* modes are

TE, $$Z_w = Z_w' \frac{\lambda_g}{\lambda} \qquad \text{ohms} \qquad (20.118)$$

TM, $$Z_w = Z_w' \frac{\lambda}{\lambda_g} \qquad \text{ohms} \qquad (20.119)$$

Equations (20.120) to (20.122) give several of the commonly used expressions for the characteristic impedance of a rectangular waveguide.

VI, $$Z_o = Z_w \frac{\pi b}{2a} \qquad \text{ohms} \qquad (20.120)$$

WV, $$Z_o = Z_w \frac{2b}{a} \qquad \text{ohms} \qquad (20.121)$$

WI, $$Z_o = Z_w \frac{\pi b}{4a} \qquad \text{ohms} \qquad (20.122)$$

The symbols V, I, and W denote the basis upon which the determination is made, i.e., voltage and current, power and voltage, or power and current.

For circular guide with the dominant TE_{11} mode

WV, $$Z_o = \frac{754}{\sqrt{1 - (\lambda/\lambda_c)^2}} \qquad \text{ohms} \qquad (20.123)$$

where λ = wavelength in an unbounded medium
λ_c = cutoff wavelength

20.5e. Ridge Waveguide.[1] Circular and rectangular ridge waveguides are shown in Fig. 20.32. The general properties of ridge waveguide are a lower cutoff frequency,

(a) RIDGE IN CIRCULAR GUIDE   (b) RIDGE IN RECTANGULAR GUIDE   (c) WAVEGUIDE TO COAXIAL TRANSITION
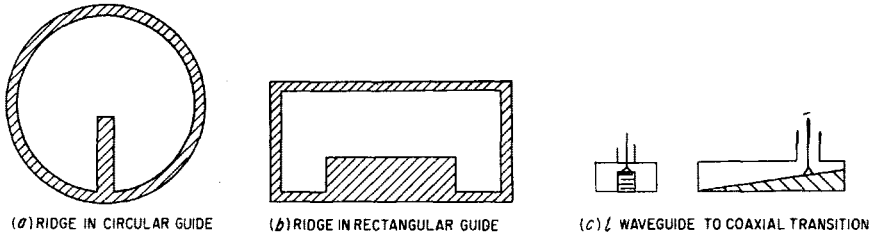
FIG. 20.32. Ridge waveguides.

greater mode separation, higher attenuation, lower characteristic impedance, and lower phase velocity. Ridge waveguides are primarily used where the bandwidth of the transmission system is to be increased by lowering the cutoff frequency. Bandwidth improvement of approximately 4:1 is achievable. Matching between regular and ridge waveguides can be accomplished by a tapered ridge at least one wavelength long. A tapered ridge can also be used for a wideband match between waveguide and a coaxial line as shown in Fig. 20.32c.

[1] For further information see S. B. Cohn, Properties of Ridge Waveguide, *Proc. IRE,* August, 1947.

20.5f. Parallel Plate Waveguide.   A parallel plate waveguide[1] is shown in Fig. 20.33. If the separation a is small compared to the width b of the strip, the characteristic impedance is given by

$$Z_o = \frac{120\pi}{\sqrt{\epsilon_r}} \frac{a}{b} \quad \text{ohms} \tag{20.124}$$

By making the spacing small, low values of characteristic impedance are achievable with this type of waveguide.   The dominant mode (TEM) has an infinite cutoff wavelength, and the guide wavelength λ_g for this mode is identical to the free-space wavelength λ.   The next highest modes $(TE_{01}$ and $TM_{01})$ are propagated at wavelengths shorter than $\lambda_g = 2b$.

20.5g. Microwave Strip Transmission Lines.[2]   Coaxial and two-wire transmission lines can be modified in form to produce the equivalent transmission line configurations shown in Fig. 20.34a. These *microwave strip* transmission lines are composed of a thin rectangular strip which is small compared to the single or double ground plane conductor as shown in the figure.   The conductors can be separated and mechanically supported by a solid dielectric material as shown in the first two types of strip transmission line, or the center conductor can be etched on one or both sides of a dielectric sheet which is supported by posts between the ground plane conductors.   The center strip can also be supported by thin sections of dielectric between the ground plane conductors.   The solid dielectric lines will have the disadvantages of

Fig. 20.34. Microwave strip circuits.

Fig. 20.33.   Parallel   plate waveguide.

greater weight, higher attenuation, and lower Q as compared to the line wherein a majority of the space between conductors is air-filled.   Strip transmission lines have the over-all advantages of lower weight and (usually) smaller size and lower fabrication

[1] See   N.   Marcuvitz,   "Waveguide   Handbook,"   p.   62,   Radiation   Laboratory   Series, vol. 10, McGraw-Hill Book Company, Inc., New York, 1951, or G. C. Southworth, "Principles and Applications of Waveguide Transmissions," p. 93, D. Van Nostrand Company, Inc., Princeton, N.J., 1950.

[2] For   additional   data   see   *IRE Transactions on Microwave Theory and Techniques*, vol. MTT-3, March 1955, No. 2, "Symposium on Microwave Strip Circuits."

cost than their coaxial or waveguide counterparts. They may exhibit increased attenuation and/or radiation loss characteristics, however. As dielectric loss is particularly important in strip lines, low loss dielectrics such as Teflon or Teflon-impregnated Fiberglas are usually used. Additional losses are introduced by discontinuities or bends in the line. Microwave strip transmission lines are usually designed to operate in the TEM mode. This requires that two ground planes be employed and that the spacing between ground planes be less than $\lambda/2$ if higher-order modes are not to be supported. For this case and provided the strip width is small compared to $\lambda/2$, the attenuation of energy radiated transverse to the line axis will be $\alpha = 27/d$ db per unit of distance, where d is the distance between ground plane conductors. The coupling between adjacent strip lines can therefore be made very small. The power-handling capability of a strip line is comparable to a coaxial line with the same conductor separation. Despite the problems associated with strip transmission lines, they have important application in very lightweight equipment or in equipment where "subminiaturization" of the microwave portions of the equipment is important.

Typical microwave strip components are shown in Fig. 20.34. In each case the dielectric material is shown crosshatched. The couplings shown in Fig. 20.34b are used to convert from strip to coaxial or waveguide transmission lines. The lumped elements shown in Fig. 20.34c are used to insert series or shunt impedances in the line. The open-circuit stub will add a shunt capacitance to the line if the stub length is less than $\lambda/4$. The metallic post will add a shunt inductance to the line, the magnitude of the inductance increasing as the post diameter is increased. The offset line adds a series inductance to the line. An abrupt change of cross-sectional area of the small strip conductor produces the same effect. The gap adds a series capacitance to the line.

These lumped elements can be combined to form resonant sections, as shown in Fig. 20.34d.

Typical directional coupler configurations are shown in Fig. 20.34e. The side-coupled coupler takes advantage of the fact that energy traveling from terminal 1 to terminal 2 will be capacitively coupled into terminals 3 and 4 with the signal at terminal 4 exceeding that at terminal 3. The two- and three-branch couplers utilize transmission lines approximately $\lambda/4$ in length and with approximately $\lambda/4$ separation. The shunt conductor widths are adjusted in accordance with the magnitudes of the binomial coefficients as shown in the figure for the two- and three-branch cases. Branched line couplers permit close coupling without introducing serious reflections. The use of multibranch couplers increases the frequency band over which a desired directivity or coupling can be maintained. The coupling is adjusted by varying the length and width of the shunt conductors. The directivity is a function of the spacing between branches.

Hybrid junctions are shown in Fig. 20.34f. The branch junction is constructed so that the branch lengths and separation are each approximately $\lambda/4$. The characteristic impedance (inversely proportional to width) of the branches is made approximately the same as that of the main transmission line. The series section between branches has a characteristic impedance approximately 0.707 times the main transmission line impedance. The spacing between the branches of the hybrid ring is approximately $\lambda/4$. The width of the strip line between branches is decreased in comparison with the main line such that the characteristic impedance of the ring is 1.4 times the line impedance.

The filters shown in Fig. 20.34g can be designed with Q's of several thousand. The left-hand configuration is a low-pass filter. The filter elements are shunt capacitance formed by the broad stubs transverse to the line axis and series inductance formed by the very narrow series branches. The gap filter is a bandpass section with the

filter characteristic depending upon the gap spacings and separations. The stub filter is a high-pass section. The shunt inductances are formed by the stubs transverse to the line. The stubs can be either short-circuited at the end with a length less than $\lambda/4$ or open-circuited at the end with a length greater than $\lambda/4$. The series capacitance is formed by separating the stub sections from the line by a thin dielectric material, forming a small capacitor between the stub termination and the line.

The elements shown in the figure are a few of the many circuit configurations possible with microwave strip fabrication methods.

20.5h. Surface-wave Transmission.[1] A single conductor, when properly excited, can propagate a surface wave with sufficiently low attenuation and radiation loss to be a practical transmission line. A practical difficulty is that the energy is contained in the space surrounding the wire, and objects in close proximity to the wire can disturb the field causing reflection and loss. As the wire diameter is reduced,
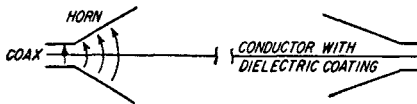
FIG. 20.35. Surface-wave transmission line.

the radius required to encircle a given percentage of the total power decreases, but there is an accompanying increase in the attenuation, so a compromise is necessary. A thin coat of dielectric material will confine the field to a smaller space around the wire with a slight increase in attenuation. Another practical difficulty for long transmission distances is the loss caused by radiation at bends. In spite of these disadvantages, the great simplicity of this type of "waveguide" may justify its use in some applications.

A surface-wave transmission system is shown in Fig. 20.35. The conical horn coupler between the line and the coaxial termination is convenient, but introduces additional coupling loss. To minimize this, the horn apertures should be as large as practical.

Attenuations comparable to conventional waveguides are possible using this technique. As an example, tests at 3,000 Mc on a wire of radius 0.1 cm coated to a thickness of 5×10^{-8} cm with a dielectric of relative dielectric constant compared to air of 3.0 yielded an attenuation neglecting the termination of approximately 0.025 db/m and contain 90 per cent of the power within a radius of 10 cm.

20.6. Waveguide Circuit Elements. Waveguides are often used to perform many specialized circuit functions in addition to their conventional use as a means of conveying signal power at high frequencies. The following paragraphs discuss waveguide configurations used to obtain impedance matching devices, waveguide transitions, mode excitation, and filtering elements, polarization changes, and waveguide hybrids.

20.6a. Impedance Elements. Waveguide impedance elements can be located in series or shunt with a waveguide transmission system. The field configuration, and consequently the mode, is important in determining whether a given element is in series or shunt with the line.

A branch entering one side of a waveguide is in series if a signal fed into the branch divides at the line and is in phase opposition at points an equal distance from the junction (electric field in series with branch and main line). A branch is in parallel if a signal fed into the branch divides at the line and is in phase agreement at points an equal distance from the junction (electric field is parallel with branch and main line). Examples of series and shunt branches for rectangular and circular waveguides operating in the dominant mode are shown in Fig. 20.36.

The impedance or admittance connected to a waveguide by a series or shunt branch-

---

[1] George Gobau, Surface Waves and Their Application to Transmission Lines, *J. Appl. Phys.*, November, 1950, p. 1119; Single Conductor Surface—Wave Transmission Lines, *Proc. IRE*, June, 1951, pp. 619–624.

ing arm depends upon the length of the arm and its termination, as discussed in previous sections.

Purely reactive or susceptive elements can be added in shunt with a waveguide system through the use of obstacles placed within the guide. The type of reactance or susceptance, i.e., inductive or capacitive, depends upon the field configuration relative to the orientation of the obstacle. Examples of inductive and capacitive

RECTANGULAR WAVEGUIDE BRANCHES   CIRCULAR WAVEGUIDE BRANCHES
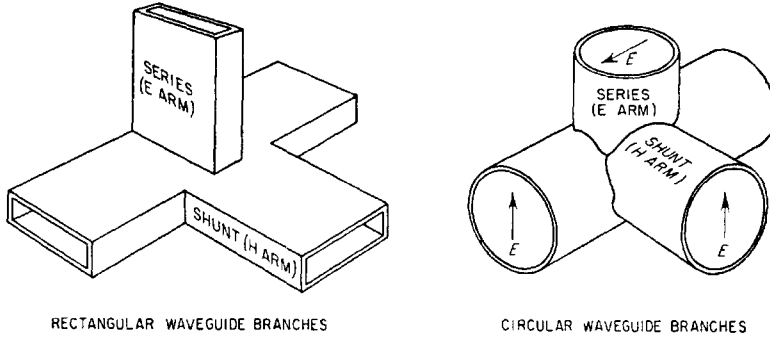
FIG. 20.36. Series and shunt branches for rectangular and circular waveguides operating in the dominant mode.
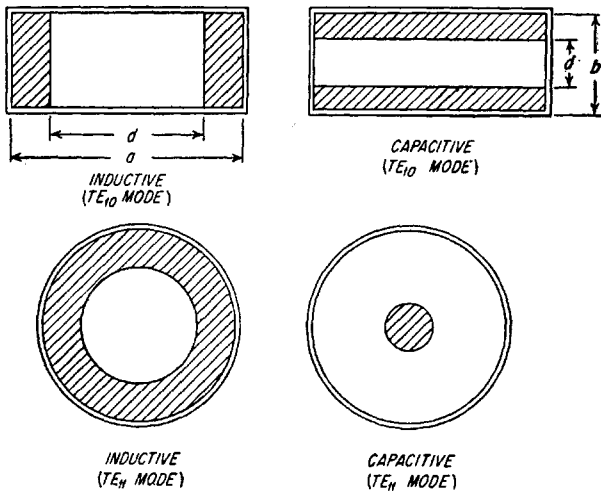
FIG. 20.37. Inductive and capacitive diaphragms in rectangular and circular waveguides.

obstacles (also called irises or diaphragms) in rectangular and circular waveguides operating in the dominant mode are shown in Fig. 20.37.

Irises in rectangular guide can be symmetrical as shown in the figure, or they can be asymmetrical with all of the obstruction placed on one side of the guide. The relationships between the dimensions of the iris and the normalized susceptance B/Y_o for symmetrical and asymmetrical irises in rectangular guides are shown graphically in Figs. 20.38 to 20.41.[1] These figures neglect the small effect on susceptance of the finite iris thickness.

[1] Adapted from N. Marcuvitz, "Waveguide Handbook," Radiation Laboratory Series, vol. 10, 1951, G. L. Ragan, "Microwave Transmission Circuits," Radiation Laboratory Series, vol. 9, McGraw-Hill Book Company, Inc., New York, 1948.

In general, irises should be as thin as mechanical considerations will allow.   Inductive irises are preferred to capacitive irises in rectangular waveguides because of voltage breakdown considerations.   Asymmetrical irises give slightly higher values of susceptance than symmetrical irises for the same ratio of open to closed area.
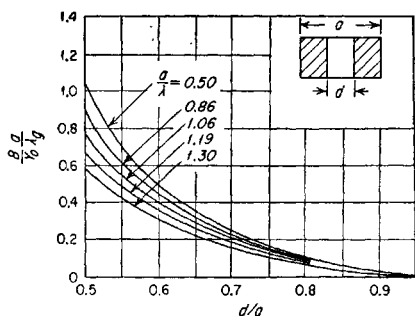
Fig. 20.38. Normalized susceptance of a symmetrical inductive diaphragm.
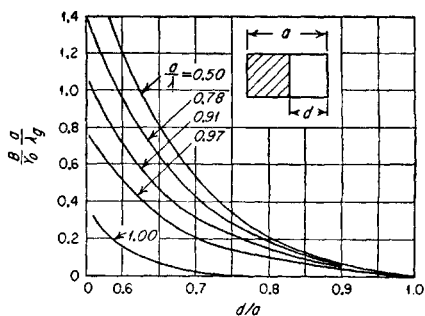
Fig. 20.39. Normalized susceptance of an asymmetrical inductive diaphragm.
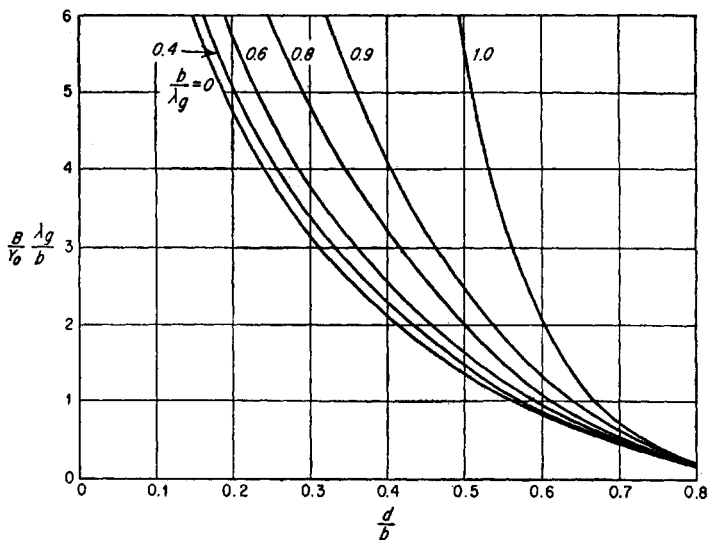
Fig. 20.40. Normalized susceptance of a symmetrical capacitive diaphragm.   (*Adapted from Marcuvitz, "Waveguide Handbook," McGraw-Hill Book Company, Inc., New York, 1951.*)

Resonant Irises.[1]   By the proper combination of reactive elements, resonant sections can be formed in waveguide structures.   A combination of inductive and capacitive diaphragms can form a parallel resonant section in shunt with the guide and consequently transmit the resonant frequency but reflect other frequencies. Resonant sections are shown for rectangular and circular guide in Fig. 20.42.   The dimensions of a resonant rectangular slit can be determined from Eq. (20.125).

$$\frac{a}{b}\sqrt{1 - \left(\frac{\lambda}{2a}\right)^2} = \frac{a'}{b'}\sqrt{1 - \left(\frac{\lambda}{2a'}\right)^2} \qquad (20.125)$$

[1] See Southworth, *op. cit.*, or A. T. Starr, "Radio and Radar Technique," Sir Isaac Pitman & Sons, Ltd., London, 1953.

where a, b, a', and b' are as shown in the figure and λ is the free-space resonant wavelength.
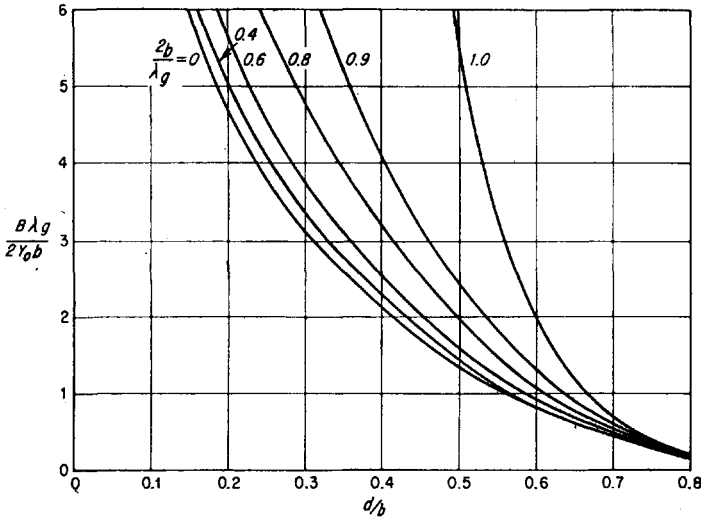
FIG. 20.41. Normalized susceptance of an asymmetrical capacitance diaphragm. (*Adapted from Marcuvitz, "Waveguide Handbook," McGraw-Hill Book Company, Inc., New York, 1951.*)

It should be noted that an interchange of open and solid areas interchanges the transmission and rejection functions of the section.

Values of Q between 10 and 100 are obtained with such sections.

Impedance Matching and Impedance Transformation. Use of Susceptance Diaphragms. The inductive and capacitive diaphragms discussed in the preceding paragraphs are useful in eliminating at one frequency the standing wave resulting from a discontinuity or load mismatch. The susceptance acts in such a manner as to add a standing wave to the line equal in magnitude but opposite in phase to the undesired standing wave.

The relationship between normalized susceptance and standing wave ratio can be determined from a Smith chart by determining the radial distance from the center of the chart (unity standing wave ratio) to the intersection of the normalized susceptance and the unit conductance circle. This distance, measured on the radial arm, gives the standing wave ratio associated with a particular normalized susceptance. This relationship is also given[1] in Eq. (20.126) and Fig. 20.43 for convenience.

FIG. 20.42. Resonant sections in rectangular and circular waveguides.

[1] From Microwave Transmission Design Data, p. 97, Sperry Gyroscope Co., May, 1944, and Ragan, *op. cit.*, p. 210.

$$\frac{B}{Y_o} = \frac{\rho - 1}{\sqrt{\rho}} \tag{20.126}$$

where B/Y_o = normalized susceptance required to correct a voltage standing wave ratio ρ.

The distance between the required susceptance and the nearest voltage minimum is the distance in wavelengths, as measured around the Smith chart, between a radial line through the intersection of the required value of susceptance and the unit conductance circle and the normalized resistance axis of the chart in the direction of decreasing values of normalized resistance.
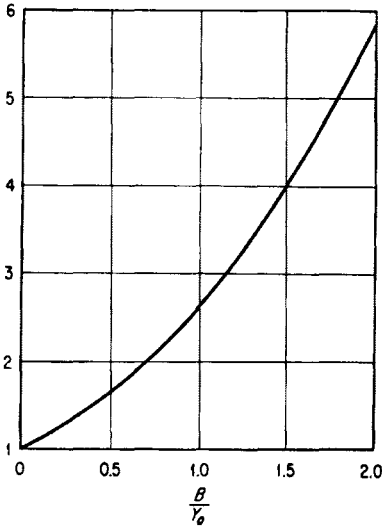
FIG. 20.43. Relationship between normalized susceptance and voltage standing wave ratio.
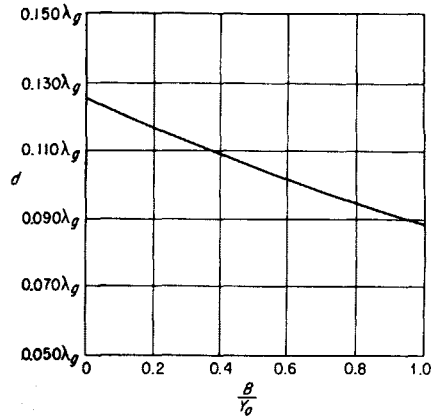
FIG. 20.44. Relationship between the normalized susceptance of a matching element and its distance from a voltage minimum.

This relationship is also given[1] in Eq. (20.127) and Fig. 20.44 for convenience.

$$d = \frac{90° - \tan^{-1}|B/2Y_o|}{720°}\,\lambda_g \tag{20.127}$$

where $B/2Y_o$ = one-half normalized susceptance
  λ_g = wavelength in the guide
  d = distance between susceptance and the nearest voltage minimum

If an inductive iris is used, it is placed on the load side of the voltage minimum. A capacitive iris is placed on the generator side of a voltage minimum.

These relationships are illustrated by the following example.

Example 20.7

Design a symmetrical inductive diaphragm to eliminate a voltage standing wave ratio of 2:1 on a RG-52u rectangular-waveguide transmission line. There is a minimum of the standing wave pattern 1.3 cm from the load. The free-space wavelength is 3.0 cm.

Solution

1. Determine the required inductive susceptance. From Fig. 20.43 or Eq. (20.126), the normalized susceptance required to match a 2:1 VSWR is found to be 0.707.

─────────

[1] From Microwave Transmission Design Data, p. 97, and Ragan, *op. cit.*, p. 210.

2. Determine the proper location of the diaphragm from the Smith chart. From Fig. 20.44 or Eq. (20.127), the separation between the matching iris and the voltage minimum is 0.098 λ_g.

3. Determine the guide wavelength.

From Eq. (20.110),

$$\lambda_g = \frac{\lambda}{\sqrt{\epsilon_r - (\lambda/\lambda_c)^2}}$$

From Table 20.2, $\lambda_c = 4.57$ cm.

$$\therefore \lambda_g = \frac{3}{\sqrt{1 - (3/4.57)^2}} = 3.98 \text{ cm}$$

4. Determine the actual position of the diaphragm. The spacing between the VSWR minimum and the diaphragm is $0.098\lambda_g$.

$$d = 0.098 \times 3.98 = 0.390 \text{ cm}$$

The distance between the VSWR minimum and the load is 1.3 cm. Therefore, the diaphragm is spaced $1.3 - 0.390$, or 0.910 cm, from the load.

5. Determine the dimensions of the diaphragm.

$$\frac{B}{Y_o} \frac{a}{\lambda_g} = 0.7 \times \frac{0.9 \times 2.54}{3.98} = 0.402$$

$$\frac{a}{\lambda} = \frac{0.9 \times 2.54}{3.0} = 0.765$$

Therefore, from Fig. 20.38, $d/a = 0.62$. Each wing of the diaphragm is

$$\frac{1}{2}[0.9 \times 2.54 \times (1 - 0.62)] = 0.43 \text{ cm}$$

A capacitive diaphragm designed to eliminate the same standing wave ratio would have the same normalized susceptance as the inductive diaphragm and would be located toward the generator from the VSWR minimum.

Quarter-wavelength Transformers. The considerations discussed in Sec. 20.4b apply in general to quarter-wave transformers used to match two waveguides having different characteristic impedances.

There will be a reflection at the input end of a single or multiple section $\lambda/4$ transformer, because of the susceptance added to the line by the change in cross section. The magnitude of this reflection can be estimated by referring to the data for susceptance diaphragms given in the preceding paragraphs. Quarter-wave transformers can be either symmetrically or asymmetrically positioned with respect to the connecting waveguide.

Stub Tuners. The considerations discussed in Sec. 20.4a apply in general to stub tuners used in waveguide transmission lines. Stubs added in shunt with a waveguide (joined to the narrow side of the waveguide) are identical in action to coaxial stubs. Stubs joined to the wide side of a waveguide are in series with the waveguide. Although stub tuners are usually avoided because of their size, they possess the advantage over other methods of impedance matching of having power-handling capabilities comparable to the main transmission line.

Screw Tuners. A post inserted into a waveguide with the axis of the post in line with the electric field will add a capacitive susceptance if the post length is less than about $\lambda_o/4$. If the post is longer than $\lambda_o/4$, the susceptance will be inductive. For a post length of approximately $\lambda_o/4$, the shunt admittance of the post will be essentially infinite, except as limited by the losses in the post. The Q of the post at resonance increases inversely with the diameter of the post. The adjustment of resonant length of the post is difficult because the frequency of resonance is a rapidly changing function of post length.

Tapered Lines. Tapered waveguides can be used to match two waveguides having different impedances. The taper should be extended a distance of several wavelengths to minimize reflections.

20.6b. Waveguide Twists and Bends. The most convenient method of changing the orientation of a rectangular-waveguide system is to twist the waveguide around its axis until it is properly aligned. To minimize reflections, the cross section of the twisted section should remain constant and the twists should be as long as possible. The direction of polarization cannot be altered by twisting a circular waveguide. A rectangular-waveguide twist is shown in Fig. 20.45a.

A bend in a circular- or rectangular-waveguide system will result in an impedance mismatch, and reflections will occur. Such reflections can be minimized by maintaining a uniform cross section. A bend in circular waveguide may tend to produce elliptical polarization. A circular-waveguide bend is shown in Fig. 20.45b.

(a) TWIST IN RECTANGULAR GUIDE   (b) CIRCULAR WAVEGUIDE BEND

(c) *E* PLANE BEND   (d) *H* PLANE BEND

(e) SHARP CORNER   (f) MITERED CORNER   (g) DOUBLE CORNER
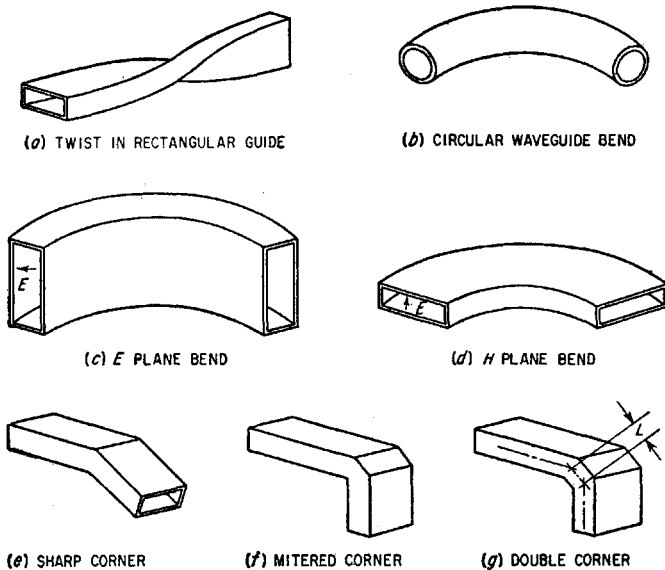
FIG. 20.45. Waveguide twists and bends.

Bends in rectangular waveguide are of the *E*- and *H*-plane types, as shown in Fig. 20.45c and d, respectively.

Waveguide corners, as shown in Fig. 20.45e, f, and g, are often used instead of bends. The sharp corner is seldom used, except for small changes in direction. The mitered corner can be used for angles up to 90°. The double corner gives the best over-all result. For *E*-plane double-corner bends, the mean length *L* should be approximately $\lambda_g/4$. For *H*-plane double-corner bends, the mean length should be slightly longer.

20.6c. Coaxial-to-waveguide Transitions. Several methods of coupling the dominant coaxial line mode into the dominant waveguide mode are shown in Fig. 20.46. The probes shown in Fig. 20.46a and c are narrow-band low-power transitions. The magnetic loop shown in Fig. 20.46b is also a narrow-band low-power transition. The cross-bar coupling shown in Fig. 20.46d is capable of higher-power operation than the previous transitions and has somewhat greater bandwidth. The "door-knob" transition shown in Fig. 20.46e is a high-power transition. Where unusually large bandwidth is required, a coaxial to ridge waveguide transition can be used.

20.6d. Mode Excitation. The same principles discussed in Sec. 20.4c apply to the problem of exciting modes in rectangular waveguides. Mode excitation in circular waveguide is essentially the same as shown in Fig. 20.26 for coaxial lines. Representative methods of mode excitation in rectangular waveguides are shown in Fig. 20.47. In general, one electric probe is used for each maximum of electric field, or one magnetic loop is used for each maximum of magnetic field.

(a)   (b)   (c)   (d)   (e)
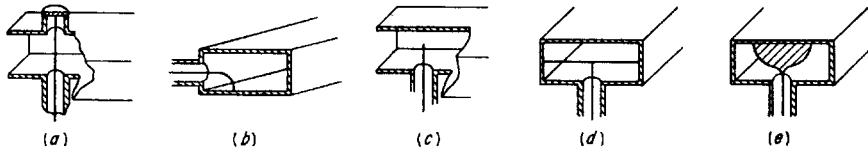
FIG. 20.46. Coaxial-to-waveguide couplings.
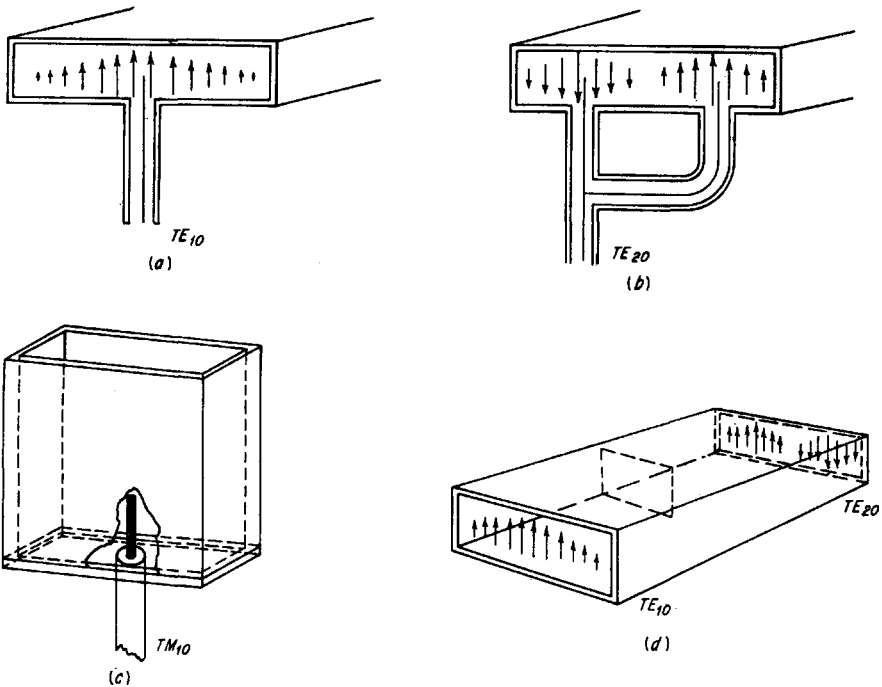
FIG. 20.47. Mode excitation.

Mode transitions are used to convert one mode into a different mode or to transfer identical modes between different types of waveguides. An example is shown in Fig. 20.47d.

Mode filtering in waveguide follows the same principles outlined in Sec. 20.4c for coaxial lines.

20.6e. Biconjugate Networks. A biconjugate network is defined[1] as a linear network with four, and only four, resistances conjugate in pairs. Conjugate resistances are defined as resistances such that a voltage in series with one resistance causes no current to flow in the other. Familiar forms of biconjugate networks are directional couplers, hybrid junctions (magic T's), and hybrid rings.

[1] L. J. Cutrona, Theory of Biconjugate Networks, *Proc. IRE*, July, 1951.

Directional Couplers.[1]   A simplified directional coupler is shown in Fig. 20.48a. Incident power entering arm A will be transmitted to arm B, and a small amount will also be transmitted through the two holes into the waveguide containing arms C and D.   Energy will travel in both directions from each hole.   The two signals will add a phase in the forward direction, and power can be coupled from arm D.   The two signals will cancel in the reverse direction, and no energy will be transmitted into arm C.   Conversely, power incident upon arm B will be coupled to arms A

(a) BASIC TWO HOLE COUPLER

(b) DUAL TWO HOLE COUPLER

(c) CROSS GUIDE COUPLER

(d) LONG SLOT COUPLER

(e) MULTI HOLE COUPLER
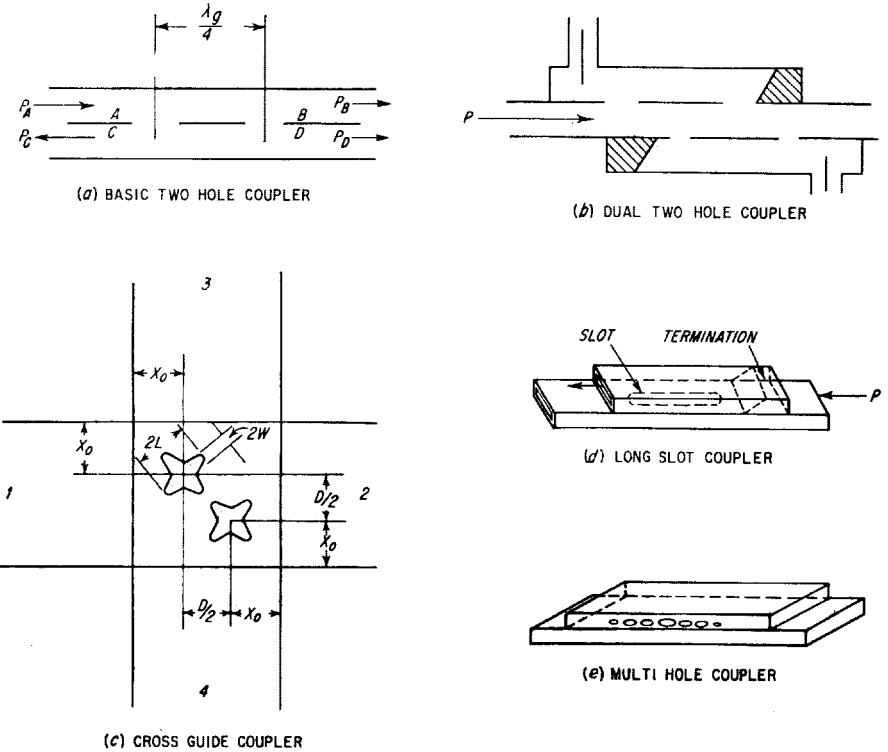
FIG. 20.48. Directional couplers.

and C but will not be coupled to arm D.   This simplified explanation assumes there are no reflections or standing waves and that the hole size is small so that essentially the same power is coupled through each hole.

A directional coupler is usually specified by its coupling C and directivity D. Referring to Fig. 20.48a, these two parameters are

$$C = 10 \log_{10} \frac{P_D}{P_A} \tag{20.128}$$

$$D = 10 \log_{10} \frac{P_D}{P_C} \tag{20.129}$$

$$10 \log_{10} \frac{P_C}{P_A} = C - D$$

[1] For design details see vol. 11, Radiation Laboratory Series, "Technique of Microwave Measurements" by C. G. Montgomery, McGraw-Hill Book Company, Inc., New York, 1947.

If, for example, a directional coupler has a coupling of -30 db and a directivity of 20 db, then the ratio P_C/P_A is -50 db.

Because of their property of sampling power traveling in only one direction, directional couplers are useful in measuring separately the transmitted and received powers of a communication system.

In practical use, absorbing material is used in the zero signal arm to prevent the reflection of any uncanceled signal. Two separate couplers are, therefore, required if power is to be measured in two directions. A dual coupler of this type is shown in Fig. 20.48b.

Two-hole couplers are frequency-sensitive in that the spacing between holes will be one-quarter guide wavelength at only one frequency. This frequency dependence can be reduced by using a cross-guide coupler as shown in Fig. 20.48c. It consists of two rectangular waveguides, joined on the broad faces with their longitudinal axes at right angles. In the coupling area, the two waveguides share a common wall. The coupling mechanism consists of two identical pairs of crossed slots as shown in Fig. 20.48c. With power incident on arm 1, a wave is set up in the secondary arm in the direction of arm 3, and very little power is coupled to arm 4. The chief advantage of this type of coupler is that both the coupling and directivity are fairly independent of frequency. A semiempirical design equation that is satisfactory for couplings as tight as -20 db is[1]

$$C = 4e^{-\pi t k/L} \frac{\pi^4}{9a^4b^2} \frac{L^6}{[\ln{(4L/W)} - 1]^2} \frac{1}{[(4fL/11.8)^2 - 1]^2} \sin{\frac{2\pi D}{\lambda g}} \quad (20.130)$$

where t = wall thickness in coupling area
  k = 1.275 for -20 to -30 db coupling and decreases to 1.0 for -50 to -60 db
  L = half length of coupling slot
  a = wide dimension of waveguide
  b = narrow dimension of waveguide
  W = half width of coupling slot
  f = frequency, kmc
  $D/2$ = path length between two slots
  λ_g = guide wavelength
All lengths are in the same units.

Directivities in the order of 20 db are achievable with cross-guide couplers.

Where large amounts of coupling and high discrimination against coupling in the opposite direction are required, the "long slot" coupler shown in Fig. 20.48d is useful.

Where high directivity and wide bandwidth are required, the multihole coupler shown in Fig. 20.48e is used. The hole sizes decrease symmetrically from the center in accordance with the binomial coefficients. A directivity of 40 db is obtainable with a four-hole coupler of this type, the hole areas being in the ratio of 1:3:3:1.

Hybrid Junctions (Magic T's). A rectangular-waveguide magic T is shown in Fig. 20.49. A signal fed into the H arm will divide between the side arms, and no signal will be coupled into the E arm. The side-arm signals will be in phase equal distances from the H arm. If a signal is fed into the E arm, it will divide between the side arms and no signal will be fed into the H arm. The side-arm signals in this case will be out of phase equal distances from the E arm. Signals fed simultaneously into both the E and H arms will add in one side arm and subtract in the other. In a similar manner, in-phase signals sent into each side arm will not be coupled to the other side arm but will add in the H arm and subtract in the E arm. The previous statements apply only for the case of perfect match at the junction and no reflections

---

[1] From an unpublished paper of M. Ingalsbe.

from the terminations of each arm. When reflections are present, the isolation between conjugate arms is reduced. To match the E and H arms to the junction, matching devices are necessary. These decrease the maximum power-handling capacity of the magic T and are frequency-sensitive.
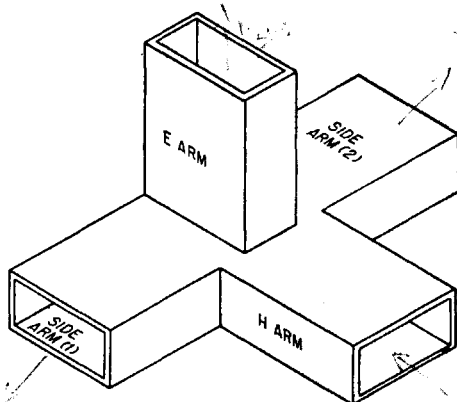
Fig. 20.49. Magic T—rectangular waveguide.

Hybrid Rings. A hybrid ring is shown in Fig. 20.50. The properties of a hybrid ring are similar to a magic T in that arms A and C are conjugate and arms B and D are conjugate. The branch arms can be either series or shunt elements.

In general, a hybrid ring will have a greater power-handling capacity than a magic T because of the absence of matching elements in the guide. For theoretically optimum results with series arms, the branch guides should be larger than the ring guide such that the impedance ratio is $\sqrt{2}$. With shunt arms the impedance ratio is reversed.
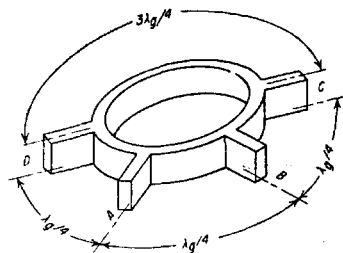
Fig. 20.50. Series hybrid ring.
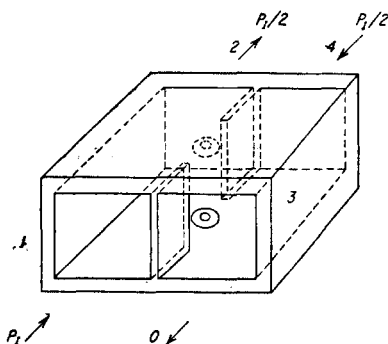
Fig. 20.51. "Short-slot" hybrid junction.

Other hybrid ring configurations can also be evolved. If the physical spacing between arms must be increased, this can be accomplished by inserting a section an integral number of wavelengths long. Coaxial hybrid rings can also be constructed.

The "short-slot" hybrid junction[1] shown in Fig. 20.51 is another form of hybrid junction with many useful applications. It is sometimes referred to as a "3-db coupler" because power incident to branch 1 will be evenly divided in branches 2 and

[1] Henry J. Riblett, The Short Slot Hybrid Junction, *Proc. IRE*, vol. 40, no. 2, pp. 180–184, Feb., 1952.

4. Any signal component crossing the junction will be shifted in phase by 90°, so the signals out of branches 2 and 4 have a relative phase difference of 90°. A hybrid junction of this type operating in the band of 8,500 to 9,500 Mc can divide power equally to within ±0.25 db, provide isolation greater than 30 db, and have a standing wave ratio below 1.07.

20.6f. Waveguide Attenuators and Terminations. Waveguide attenuators and terminations employ the same materials discussed in Sec. 20.4e for coaxial attenuators. Their method of use is somewhat different, however, as shown in Fig. 20.52. The variable attenuators in Fig. 20.52a and b are composed of a movable resistive element where the attenuation increases as the element approaches the position of maximum electric field (the center of the guide in Fig. 20.52a) or as the insertion of the element is increased (Fig. 20.52b). The shape of the resistive element is usually the single taper shown for the fixed attenuator in Fig. 20.52e or the double taper shown for the fixed attenuator in Fig. 20.52f. Montgomery[1] has shown that the attenuation of such

(a)PARALLEL STRIP   (b)RESISTIVE PLUNGER   (c)BIFURCATED LINE   (d)ABSORBING LOAD

(e)SINGLE TAPER   (f)DOUBLE TAPER
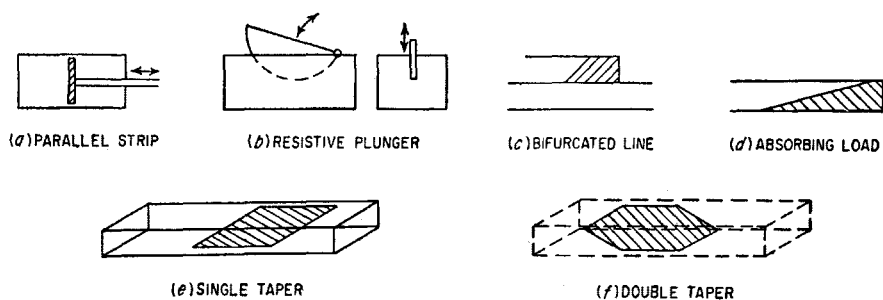
FIG. 20.52. Waveguide attenuators and terminations.

strips is given by Eq. (20.131) for 1 by ½ in. waveguide with 0.050 in. wall at a wavelength of 3.3 cm and Eq. (20.132) for 3 by 1½ in. waveguide with 0.080 in. wall at a wavelength of 10 cm.

$$\alpha_{3.3} = 15.7 - 1.5 \frac{\Omega}{100} \quad \text{db/in.} \tag{20.131}$$

$$\alpha_{10.0} = 5 - 0.38 \frac{\Omega}{100} \quad \text{db/in.} \tag{20.132}$$

where Ω = resistance per square of material

These equations apply for the TE_{10} mode with the electric-field maximum in the center of the guide. Attenuators for the TE_{10} mode would have two resistive strips located at the two maxima of electric field.

The bifurcated line and absorbing load separate the incident power into two portions. The portion sent into the upper waveguide with the absorbing termination is dissipated in this termination. The remainder of the power is sent down the bottom waveguide.

A variation is the absorbing-wall termination. This termination utilizes lossy material in the walls of the waveguide and does not require the tapered form shown in Fig. 20.52d.

20.7. Microwave Cavities and Filters. Resonant cavities are devices whose physical dimensions are in the order of a wavelength of the energizing source. Their principal use, therefore, is at frequencies where the dimensions are convenient and practical in the construction of physical apparatus. Cavities provide a means of

[1] Montgomery, *op. cit.*, pp. 729–803.

signal storage, frequency measurement, and circuit synthesis for applications requiring resonant elements.

Resonant structures such as tuned transmission lines and waveguide irises have been discussed in previous sections. In practice these are often not satisfactory resonators because of radiation loss, difficulty of tuning, or insufficient Q. Cavity resonators can usually be used to overcome these advantages. Resonators are often used in conjunction with other apparatus such as a tube, transmission line, or other load, and the final performance of the cavity can only be determined by tests involving all the connecting structures.

20.7a. Microwave Cavities. Cavity resonators in general consist of simple geometrical shapes containing an internal hollow cavity, the dimensions of which are related to the resonant wavelengths and the modes excited within the cavity.

Resonant Wavelength. If the input impedance to a two-terminal network consisting only of reactive elements is measured over a large range of frequencies, it will be found that the impedance alternates between maximum and minimum values. This is also true for cavity resonators. The values of the resonant wavelengths as a function of cavity shape and mode are given in Eqs. (20.133) and (20.134) for the commonly

(a) RECTANGULAR   (b) CYLINDRICAL   (c) COAXIAL
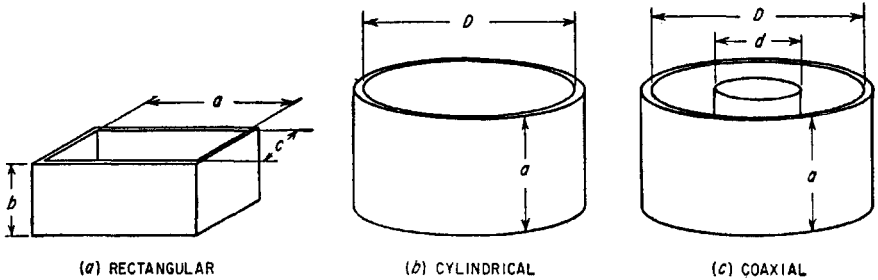
FIG. 20.53. Simple cavity resonator shapes.

used resonator sections shown in Fig. 20.53. For a rectangular cavity, TE_{lmn} or TM_{lmn} modes,

$$\lambda_k = \frac{2}{\sqrt{(l/a)^2 + (m/b)^2 + (n/c)^2}} \tag{20.133}$$

where λ_k = resonant wavelength
l, m, n = number of half-cycle variations of electric field along the a, b, and c dimensions, respectively (the electric field variation can be zero along only one axis).

For cylindrical cavities, TE or TM modes,

$$\lambda_k = \frac{1}{\sqrt{\left(\dfrac{l}{2a}\right)^2 + \left(\dfrac{1}{kD}\right)^2}} \tag{20.134}$$

where a and D are as shown in Fig. 20.53 and k is obtained from Table 20.1 for the corresponding circular waveguide mode. l is the number of half-cycle variations of electric field along the cylinder axis. For TE modes l cannot equal zero. For TM modes l can equal zero.

For coaxial cavities, TEM mode,

$$\lambda_k = \frac{2a}{l} \tag{20.135}$$

Cavity Q.[1]   The Q of a cavity resonator can be defined as

$$Q_o = \frac{f_o}{\Delta_f} \qquad (20.136)$$

where Δ_f = bandwidth between those frequencies at which impedance is 70.7 per cent of impedance at f_o or frequencies at which cavity reactance = cavity resistance, f_o = resonant frequency

The value of Q_o is also proportional to the ratio of the volume of the cavity to the surface of the area enclosing the cavity.

As defined above, Q_o is the *unloaded* Q of the cavity. In an unloaded cavity, all the power delivered to the cavity through the input terminals is dissipated within the cavity. If an external load is coupled into the cavity, the Q will be decreased. The loaded Q, Q_L, of the cavity is defined as the Q with an external load coupled to the cavity. Fundamental relationships for Q_o and Q_L are given in Eqs. (20.137) and (20.138):

$$Q_o = \frac{2\pi f_o \times \text{energy stored in cavity}}{\text{power dissipated in cavity}} \qquad (20.137)$$

$$Q_L = \frac{2\pi f_o \times \text{energy stored in cavity}}{\text{power dissipated in cavity plus load}} \qquad (20.138)$$

It can be seen from Eq. (20.137) that Q_o will depend upon the mode and the resistance losses in the cavity walls. Q for a coaxial resonator can also be defined as $\beta/2\alpha$
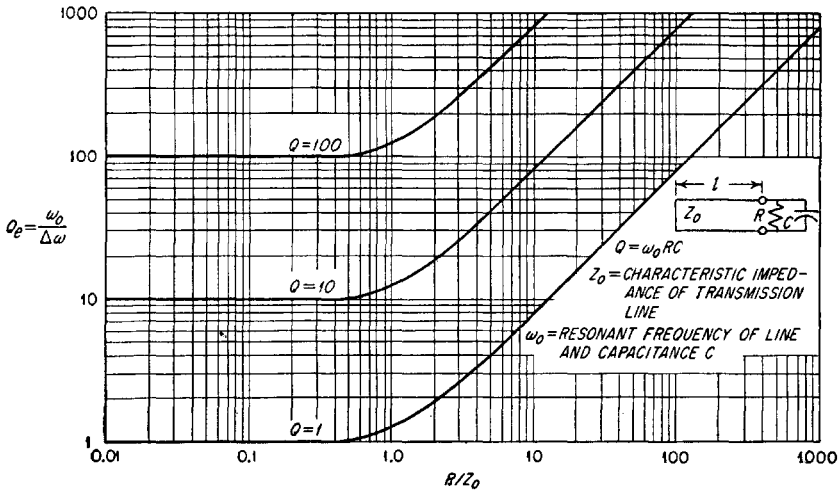
FIG. 20.54. Effective Q of a resonant circuit employing a lossless shorted transmission line as the inductive element.

as discussed in Sec. 20.3f. Q values of 10^4 are achievable. Where considerably lower Q values are required, for example, 10 to 50, resonant irises are employed instead of cavity resonators.

In practice, a section of transmission line is sometimes resonated with lumped circuit elements to form the equivalent of a resonant cavity. For example, in the design of high-frequency oscillators the inductance of the resonant circuit is often supplied more conveniently by a section of shorted transmission line less than $\lambda/4$

[1] For a discussion of methods of measurement of cavity characteristics, see Montgomery, *op. cit.*, and H. J. Reich, P. F. Ordung, H. J. Kraus, and J. G. Skalnik, "Microwave Theory and Techniques," pp. 451–455, D. Van Nostrand Company, Inc., Princeton, N. J., 1953.

in length.[1] Figure 20.54 gives the effective Q of such a combination in terms of the Q of the lumped circuit elements and the ratio of lumped resistance to characteristic impedance of the line.

Cavity Tuning. The addition of an external load will alter the resonant frequency of a cavity. Tuning devices are usually incorporated in the cavity to compensate for unwanted frequency changes. One such device consists of a flat metallic paddle oriented so that the plane of the paddle can be turned normal to the magnetic field. Tuning can also be accomplished by screws parallel to the electric field. Gross changes in resonant frequency can be accomplished by altering one of the cavity dimensions.

Multimode Cavities. It is sometimes desirable to excite more than one mode in a cavity resonator. In general, this can be accomplished by maintaining the fields of each mode orthogonal. Each mode can then be separately coupled and tuned. If internal coupling between modes is accomplished in a controlled manner, over-all transfer characteristics comparable to double-tuned or triple-tuned circuits can be obtained.

20.7b. Microwave Filters. Filter theory has been developed in great detail for lumped constant networks.[2] No basically new theory has evolved for distributed structures used at microwaves. In general, design procedures consist of utilizing lumped constant theory and interpreting the results in terms of transmission-line components. An explanation of the details of design for low-pass, high-pass, and bandpass microwave filters is beyond the intent of this section;[3] however, in view of the widespread use of microwave filters, a brief description will be given of the several design approaches.

Transmission-line Cutoff Filters. In general, a waveguide transmission line can be regarded as a high-pass filter. This does not apply to coaxial lines operating in the *TEM* mode, where the cutoff frequency is zero. By proper combination of waveguide and coaxial lines, low-pass, bandpass, and band-reject characteristics can be obtained.

Equivalent Circuit Method. This is perhaps the most straightforward method of analyzing an existing filter structure. The filter configuration must be reducible to a T or π equivalent configuration. This can often be directly visualized from a knowledge of series and shunt transmission-line elements.[4] The T or π parameters are evaluated, and the filter performance is calculated from lumped constant theory. In a similar manner, filter design is possible once the image constants and insertion loss characteristics are specified through the use of lumped constant relationships.

The lumped constant elements are in general formed by open-circuited or short-circuited lines added in series or shunt as discussed in previous sections. In addition, dielectric material can be used to insert series or shunt capacitive elements.

The synthesis of lumped element bandpass filters often results in series and parallel resonant filter branches. These can be produced in microwave structures by the addition of resonant lines in series or shunt or by the use of resonant irises.

Coupled Resonant Chambers. The previous methods become impractical when narrow band filters are required. Instead, resonant chambers in waveguides are coupled together to form either wide or narrow band filters in a manner analogous to the coupling of lumped tuned circuits at lower frequencies.

[1] See Sec. 6.

[2] Discussion of basic filter theory is given in "Communication Networks," vol. II, by E. A. Guilleman, John Wiley & Sons, Inc., New York, 1935. A concise treatment of the practical design of lumped constant filters is given in Secs. 18.1 to 18.9. Much of this material is applicable to microwave networks.

[3] Two excellent references for detailed design data are "Very High Frequency Techniques," chaps. 26 and 27, and Ragan, *op. cit.*

[4] See Secs. 20.4 and 20.6.

One form of coupled filter consists of resonant chambers separated by irises.[1] Tuning screws are usually included in each chamber for alignment purposes. For inductive coupling each cavity length must be between $\lambda_g/4$ and $\lambda_g/2$. For capacitive coupling each cavity length must be between 0 and $\lambda_g/4$ or between $\lambda_g/2$ and $3\lambda_g/4$.

Other types of coupled cavities are useful in forming microwave filters. When the required passband becomes very small, it becomes impractical to directly couple resonant elements. This problem is circumvented by the use of $\lambda/4$ coupling sections between cavities.

20.8. Delay Lines. Delay lines are a specialized and important type of transmission line in which the line parameters are adjusted to decrease the velocity of signal transmission. Delay lines are of three primary types: lumped constant, distributed constant, and supersonic. Delay lines are useful in applications requiring the time delay of electrical signals from a fraction of a microsecond to several thousand microseconds. The characteristic impedance of practical delay lines normally varies from several hundred to several thousand ohms. Specific characteristics of each type of line are discussed in the following paragraphs.
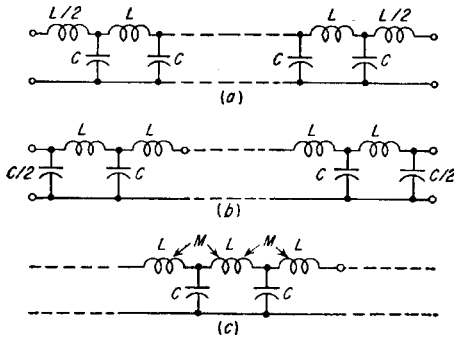
FIG. 20.55. Lumped-constant delay lines.

20.8a. Lumped-constant Delay Lines. A lumped-constant delay line is one in which the normally distributed inductance and capacitance per unit length of a transmission line occur in a single section of the delay line as a lumped inductance and a lumped capacitance.

The degree to which the lumped-constant line approximates a distributed constant line depends upon the number of sections used to form the complete line. The larger the number of sections the better the approximation. Several types of lumped-constant lines are shown schematically in Fig. 20.55. Methods of matching into and out of a lumped-constant delay line with no mutual coupling are given in Sec. 16.

To minimize phase distortion, the important harmonic components of a signal to be transmitted through a delay line should be below one-half the *cutoff* frequency f_c which is given by

$$f_c = \frac{1}{\pi \sqrt{lc}} \qquad (20.139)$$

where l = inductance per section, henrys
c = capacitance per section, farads

A signal not meeting this requirement will, in general, be distorted because of the increased time delay to signal components above $f_c/2$. A signal having frequency components above f_c will, in addition, be distorted because of high attenuation of these components. It can be seen that a delay line is a low-pass filter. The time

[1] For further data see W. W. Mumford, "Maximally Flat Filters in Waveguides," *Bell Telephone System Tech. J.*, vol. 27, no. 4, pp. 684–713, October, 1948.

delay per section can be determined from Fig. 20.56. The characteristic impedance Z_o of the delay line for very low frequencies is

$$Z_o = \sqrt{\frac{l}{c}} \quad \text{ohms} \tag{20.140}$$

The time delay per section for very low frequencies is

$$t_d = \sqrt{lc} \tag{20.141}$$

where $t_d =$ time delay per section, sec

The characteristic impedance decreases with frequency for the network shown in Fig. 20.55a and increases for the network shown in Fig. 20.55b. Equations (20.140) and (20.141) and Fig. 20.56 do not apply for delay lines containing mutual coupling or distributed constants.
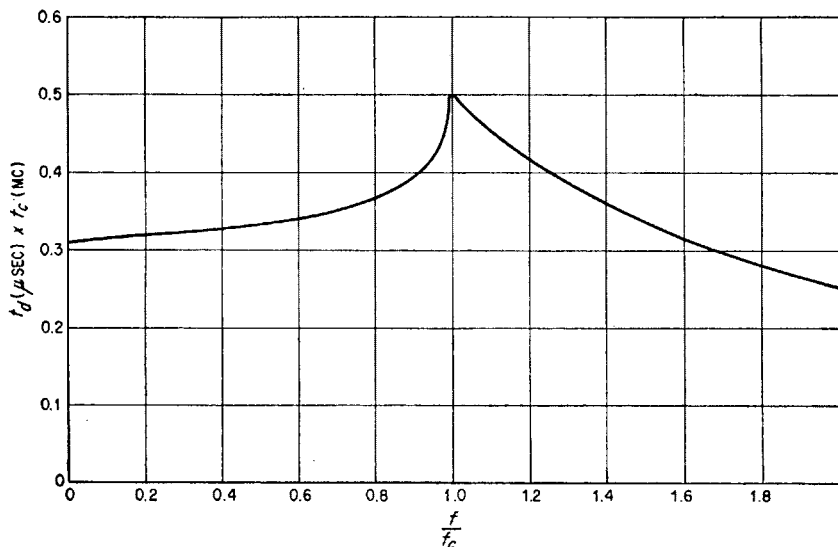
FIG. 20.56. Time delay per section of delay line as a function of frequency ratio.

If mutual coupling exists between the series inductances, the phase distortion of the signal due to unequal time delay of frequency components near f_c will be reduced. The optimum value of coupling coefficient is approximately 0.36 (see Fig. 1.2, Sec. 1).

In general, phase distortion of any type of delay line can be corrected by means of a phase equalizer.[1]

20.8b. Distributed-constant Delay Lines. A distributed-constant delay line is formed by winding a solenoidal layer of wire on a cylindrical metallic form. The metallic form is used as one conductor of the delay line. The solenoidal inductance without the presence of the metallic form is equivalent to the total line inductance, and the capacitance between the winding and the metallic form is equivalent to the total line capacitance. The impedance can be increased by substituting a second solenoidal winding in place of the metallic form. The winding senses of the two coils should be opposed. The phase distortion of a distributed-constant delay line can be equalized by means of an external phase-correction network[1] or bridging capacitance.[2]

[1] See Sec. 17.3.
[2] See Radiation Laboratory Series, vol. 17, "Components Handbook," chap. 6, by J. F. Blackburn, McGraw-Hill Book Company, Inc., New York, 1949. A detailed description of design techniques is given as well as a description of the physical construction of distributed-constant delay lines.

In general, a distributed-constant delay line will be more compact than a lumped-constant line for the same delay and cutoff frequency.

Coaxial lines are sometimes used for signal delay. The total delay T for a coaxial line is given by

$$T = 0.0033 \sqrt{\epsilon_r} \qquad \mu\text{sec/meter} \qquad\qquad (20.142)$$

ϵ_r = relative dielectric constant

The length of a coaxial line for the same time delay will be much greater than the lumped-constant line or distributed line previously discussed.

20.8c. Supersonic Delay Lines. For special applications where very large delay intervals are required, supersonic delay lines are employed. This type of line usually consists of a liquid (usually mercury), quartz, or glass delay medium. Quartz crystals are employed as transducers at the input and output of the liquid line to accomplish the transition of electrical energy to mechanical energy and vice versa. These lines require that the signal be superimposed on a modulated carrier because of attenuation characteristics of the line. A mercury line, however, will give delays of 6.9 μsec/cm.[1]

The following example will illustrate the design of a lumped-constant delay line.

Example 20.8

Design a lumped-constant delay line with an upper cutoff frequency of 3 Mc, a characteristic impedance of 1,000 ohms, and a total delay of 1 μsec.

Solution

1. From Eq. (20.139),

$$\sqrt{lc} = \frac{1}{\pi f_c} = \frac{10^{-6}}{3\pi} = 1.06 \times 10^{-7}$$

2. From Eq. (20.140), assuming $f \ll f_c$,

$$Z_o = \sqrt{\frac{l}{c}} = 1{,}000 \text{ ohms}$$

3. Solve the above expression for l and c.

$$Z_o c = \sqrt{lc}$$

or

$$c = \frac{\sqrt{lc}}{Z_o} = \frac{1.06 \times 10^{-7}}{10^3}$$

$$c = 1.06 \times 10^{-10} = 106 \ \mu\mu\text{f/section}$$
$$l = 106 \ \mu\text{h/section}$$

4. The delay per section, assuming $f \ll f_c$, is

$$T = \frac{10^{-6}}{3\pi} = 0.106 \ \mu\text{sec}$$

5. The required number of sections is therefore

$$n = \frac{1.0}{0.106} = 9.4 \approx 10$$

20.9. Ferrite Elements.[2] Ferrites are materials possessing a high dielectric constant, high resistivity, and a variable permeability. They can produce nonreciprocal

---

[1] For detailed design data see *ibid.*

[2] For additional data see A. G. Fox, S. E. Miller, and M. T. Weiss, Behavior and Applications of Ferrites in the Microwave Region, *BSTS*, Jan., 1955; or C. L. Hogan, The Ferromagnetic Faraday Effect at Microwave Frequencies and Its Applications—The Microwave Gyrator, *BSTJ*, Jan., 1952; or R. H. Fox, Ferrite Devices and Applications at Microwave Frequencies, *MIT Lincoln Laboratory Tech. Report* 69, Sept. 27, 1954.

phase and polarization shift at microwave frequencies. These characteristics provide a very useful means for designing a wide variety of microwave transmission line circuit elements. Ferrites also find application as "dielectric rod" radiators in antenna systems. As such, they possess the advantage of higher directivity than is usually obtainable from an antenna composed of other radiators of equal aperture.

In the usual application of ferrites, a magnetic field H is established parallel to the direction of propagation, as shown in Fig. 20.57a. Ferrites exhibit *gyromagnetic* effects such that in the presence of this external field, energy can be coupled between the ferrite electrons and an applied r-f field. This will cause the permeability of the ferrite to be a function of the external magnetic field. This interchange of energy is a function of the direction of propagation of energy through the material and results in the nonreciprocal transmission properties of the ferrite. The maximum energy will be coupled from the r-f field to the electrons in the ferrite material at a frequency

(a) FERRITE PHASE SHIFT

(b) FERRITE ISOLATOR

(c) WAVEGUIDE CIRCULATORS
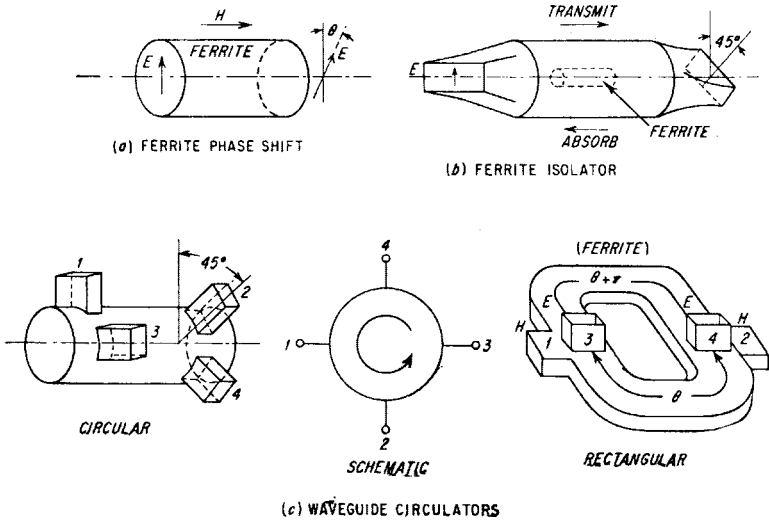
FIG. 20.57. Ferrite applications.

given by $\omega_o = 2.8H$ Mc, where H is the external d-c magnetic field in oersteds. Ferrite devices are usually operated with magnetic field strengths such that ω_o is greater than the signal frequency by a sufficient amount to obtain low loss propagation of the signal through the ferrite material.

For the proper value of magnetic field strength and signal frequency, a signal incident upon the ferrite slab, as shown in Fig. 20.57a, will have its plane of polarization shifted an amount Θ in passing through the ferrite. This same signal, redirected back through the ferrite material, will be rotated an additional angle Θ in the same direction, making the total polarization shift 2Θ for the forward and return paths. The amount of phase shift per unit length is $\Theta/l \approx 65°/\text{cm}$ for a typical ferrite material[1] at 10 kmc. This effect is similar to Faraday rotation in optical devices; however, with ferrite material useful rotation angles can be obtained at frequencies sufficiently far from ω_o to obtain low loss operation. For small values of applied magnetic field, the rotation angle per unit length is not sensitive to variations in H or ω.

A convenient method of utilizing ferrites is to place a thin ferrite rod in the center of a circular waveguide where the r-f magnetic field is transverse to the rod axis and applied magnetic field. Under these conditions, the differential phase shift discussed

[1] See Hogan, *op. cit.*

above will be produced. The proper value of applied magnetic field will vary with the material used; however, it should saturate the ferrite, yet be below the value for resonance. Phase-shift values up to 120° per cm can be obtained with field strengths of the order of 10^3 oersteds.

A precaution to be observed in the use of ferrites is the nonlinear operation obtained at high power levels. The absorption loss can increase rapidly as the power level increases. Therefore, for power levels over 1 kw peak power, all loss measurements should be made at the normally used power level.

An advantage of ferrite materials is the fact that they can be molded into the desired shape and, unlike other metallic magnetic materials, do not require laminated structures.

Ferrites are used in the design of *gyrators*. A gyrator is a device which introduces zero phase shift for one direction of propagation and 180° phase shift for the opposite direction of propagation. Microwave isolators can also be constructed from ferrite sections as shown in Fig. 20.57*b*. The isolator consists of a central section of circular waveguide with an axially mounted ferrite rod. The rectangular-waveguide end sections are mechanically displaced 45° with respect to each other. A signal entering the isolator from the left will propagate in the rectangular waveguide with the dominant TE_{10} mode. This will be converted to the circular waveguide TE_{11} mode in the circular guide where the plane of polarization is shifted by 45° through the ferrite section for transmission through the output waveguide. A signal entering from the right will be shifted 45° in the same sense, resulting in an input signal in the left-hand rectangular guide with an electric field orthogonal to the required direction for propagation. This energy can be absorbed in the rectangular waveguide by the use of resistance cards which will not affect propagation in the desired direction.

A general schematic of a circulator is shown in Fig. 20.57*c*. Energy is transmitted from terminal 1 to 2, terminal 2 to 3, etc. Circulators with four terminals as shown can be operated in tandem, providing circulators with many output and input branches.

Antennas

21.1. Fundamentals of Antennas

21.1a. Nature of Radio Waves. Radio waves consist of time varying electric and magnetic fields related to each other in such a manner that their energy is evenly divided and they are at right angles to each other in any plane perpendicular to the direction of propagation. Their velocity, like the velocity of light, is dependent upon the transmission medium and is approximately equal to 3×10^8 m/sec in free space. The relationship between this velocity c, the wavelength λ, and the frequency f is given by Eq. (21.1).

$$\lambda = \frac{c}{f} \tag{21.1}$$

For the case of radiation in free space,

$$\lambda = \frac{300}{f} \tag{21.2}$$

where λ = wavelength, meters
f = frequency, Mc

Polarization. The polarization of an electromagnetic wave is defined by the direction of the electric field. If, during the cyclic variations of the wave, the tip of the vector representing the instantaneous magnitude and direction of the electric field describes a straight line in a plane perpendicular to the direction of propagation, the polarization is linear. In the general case, the electric field vector rotates 360° in the plane perpendicular to the direction of propagation for each wavelength of travel, causing the tip of the electric vector to describe an ellipse. Circular and linear polarization may both be regarded as special cases of elliptical polarization. An elliptically polarized wave can be resolved into two orthogonal, linearly polarized, component waves of given amplitude and phase. For example, for circular polarization, the amplitudes of the component waves are equal and the phase difference is 90°. For linear polarization either the amplitude of one of the component waves is zero, or the phase difference between them is some multiple of 180°.

Reflection. Radio waves are reflected from any discontinuity encountered in the transmission medium, the magnitude and phase of the reflection depending upon the size and nature of the discontinuity. Substantial reflections occur from metallic objects, from the earth, water, rain, buildings, etc. Reflections from objects whose size is small with respect to the wavelength of the incident energy are usually of negligible amplitude. Radio waves are totally reflected from large conducting surfaces, while only part of the energy is reflected at the surface of a dielectric material.

Refraction. Refraction of radio waves occurs in the same manner as it does for light waves; that is, the wave is bent when crossing at an angle the boundary between two regions in which the velocity of propagation is different. This bending takes place because that part of the wavefront which first encounters the new medium is either advanced or delayed with respect to the part which enters the region last causing the wavefront to change direction.

Ground and Sky Waves. Radio waves can be classified according to the possible paths of propagation. In the *ground wave* the energy is propagated over the surface of the earth. This includes all components of a radio wave except waves reflected

by the ionosphere and troposphere. The ground wave includes direct line of sight transmission and ground-reflected transmission.

The sky wave occurs when a transmitted signal is bent back to a receiving antenna by the ionosphere or troposphere.

Typical sky-wave and ground-wave paths are shown in Fig. 21.1.

Ionospheric Reflections. The ionosphere is composed of several separate layers of charged particles surrounding the earth. These layers have the property of bending electromagnetic waves back toward the surface of the earth. The amount by which a radio wave is refracted in this manner depends upon the frequency of the radiation, atmospheric and geographical factors, and the angle at which the transmitted wave is incident upon the ionized layer. A *critical frequency* f_c exists at which signals vertically incident upon the ionosphere will not be returned to the surface of the earth if they exceed this frequency. Radio signals above the critical frequency will be returned to the earth, however, if the wave intercepts the iono-

FIG. 21.1. Methods of propagation.

sphere at some angle other than 90°. The amount by which the frequency can exceed f_c increases as the angle of incidence decreases, becoming several times the critical frequency at small angles of incidence. The *maximum usable frequency* f_m is related to the critical frequency as follows:

$$f_m = \frac{f_c}{\sin \beta} \tag{21.3}$$

where β = angle of incidence measured from horizontal

In practice, operating frequencies of 70 to 85 per cent of the maximum usable frequency are employed. The value of maximum usable frequency depends upon the time of day, communication distance, and layer height. Data useful in predicting maximum useful frequencies are published periodically by the National Bureau of Standards.

The lowest ionized region is called the E layer. It is approximately 70 miles above the earth and is most effective during the daylight hours. The E layer has the lowest critical frequency, and is lower in the daytime than at night.

The next highest layer is the F layer. This layer is divided into two layers, F_1 below and F_2 above in the daytime, but these merge into the single F layer at night. This layer is between approximately 180 miles in height (F_1) to 200 miles in height (F_2).

The ionosphere varies in its property of reflecting radio waves in accordance with the seasons and sunspot activity.

Propagation Characteristics. The propagation characteristics of a radio wave depend primarily upon the frequency. The following paragraphs summarize the important characteristics of each frequency band.

1. *VLF—very low frequency (10 to 30 kc)* waves are characterized by very low ground attenuation and very good reflection of sky-wave signals. These frequencies are therefore useful for very long-range communication systems. The antenna structures are usually very large and expensive. The atmospheric absorption from these frequencies is very small. Propagation is affected by sunspots and magnetic storms. Daytime ranges of thousands of miles are possible at these frequencies.

2. *LF—low frequency (30 to 300 kc).* In this region the ground-wave attenuation is increased. At the high-frequency end of the band the sky absorption is greatly increased in the daytime. This results in a reduction in the daytime range to several
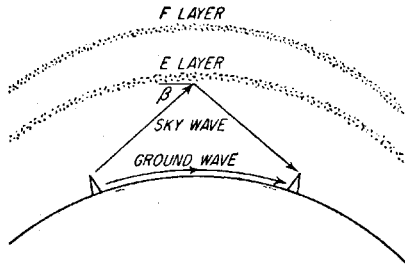
hundred miles under unfavorable circumstances; however, longer-distance sky-wave transmission is possible. Low-frequency antenna structures can be more efficient than vlf antennas, which partially compensates for the increased ground-wave attenuation.

3. *MF—medium frequencies (300 to 3,000 kc)* include the broadcast range of 500 to 1,500 kc. Antenna systems are usually designed to provide good ground-wave coverage in the region of interest. Coverage can extend for a distance of approximately 100 miles from the antenna. Beyond this limit sky-wave coverage is possible at night.

4. *HF—high frequencies (3 to 30 Mc)*. High-frequency propagation is characterized by very short ground-wave coverage. Sky-wave transmission is used exclusively for long-range communication. To achieve the best sky-wave transmission between two points the frequency must be carefully selected. The optimum frequency will vary with time and ionospheric conditions.[1]

5. *VHF—very high frequencies (30 to 300 Mc)*. In general, frequencies above 30 Mc are not reflected from the ionosphere and sky-wave transmission is not possible. Ground-wave coverage is affected by refraction and reflection in the troposphere. Atmospheric absorption is very small.

6. *UHF—ultra high frequencies (300 to 3,000 Mc)*. In this frequency band atmospheric effects become important. In general, except for certain frequencies, the atmospheric absorption increases with frequency. The atmosphere refracts radio waves making "line of sight" transmission beyond the optical horizon possible. The maximum line of sight range depends upon antenna height as follows:

$$D \simeq \sqrt{2H_t} + \sqrt{2H_r} \qquad \text{statute miles} \qquad (21.4)$$
$$\text{for } D \gg H_t, H_r$$

where H_t, H_r = heights of transmitter antenna and receiving antenna, respectively, ft

Equation (21.4) takes into consideration the effect of refraction in the earth's atmosphere increasing the line of sight transmission beyond the optical horizon.

7. *SHF—superhigh frequencies (3,000 to 30,000 Mc)* extend to the practical upper limit of reliable propagation over an appreciable distance. At frequencies in excess of 10,000 Mc, attenuation due to precipitation becomes very pronounced. Very directive antenna systems can be designed at these short wavelengths.

Scatter Propagation.[2] Recent investigations in the field of *beyond-the-horizon propagation* have shown that ionospheric scattering from the lower E layer permits the design of practical communication systems in the frequency range from 25 to approximately 60 Mc over distances from 600 to 1,200 miles. In addition, tropospheric scattering permits the design of practical communication systems over the frequency band of 100 to 10,000 Mc at distances of up to a few hundred miles beyond the horizon.

21.1b. Radiation from Antennas. An antenna is a device used for the purpose of radiating or receiving radio waves. Alternatively, it may be considered as an arrangement for matching a transmission line or r-f generator to a propagation path. In the immediate vicinity of the antenna, the field configuration is called the *induction field*. It is made up of complex local fields which die out very rapidly as the distance from the antenna increases. The only significant field component that exists outside the

---

[1] For a very readable discussion of the detailed considerations of propagation see F. E. Terman, "Radio Engineers' Handbook," sec. 10, McGraw-Hill Book Company, Inc., New York, 1943.

[2] An over-all picture of the theoretical and experimental state of the techniques of scatter propagation is given in the "Scatter Propagation Issue," *Proc. IRE*, vol. 43, no. 10, October, 1955.

immediate vicinity of the antenna is the *radiation field*. The amplitude of this field decreases inversely as the range from the antenna.

21.1c. Antenna Resistance. In order that an antenna provide a match to a source of r-f energy it must present a resistive load of the proper value. This resistive load is called antenna resistance.

The total resistance of an antenna is composed of factors resulting from ohmic loss in the conductors of the antenna, radiation loss, loss from corona, eddy currents, leakage, etc. For most practical purposes the ohmic and radiation resistances only need be considered.

The ohmic loss is equal to $I^2 R_o$, where I is the current at some particular point and R_o is the ohmic resistance of the conductors at the frequency involved. As the current will, in general, vary along a physical antenna, some particular point must be designated if ohmic loss is to have significance. The value of current usually selected is the maximum value present in the antenna.

The radiation loss in an antenna is equal to $I^2 R_R$, where I is the maximum value and R_R is the *radiation resistance*. The radiation resistance is a fictitious quantity representing a resistance which if actually connected to the antenna at the point of maximum current would dissipate an amount of power equal to the power radiated from the antenna. The total power fed into an antenna is therefore

$$P_T = I^2(R_o + R_R) \tag{21.5}$$

As the ohmic loss is usually much smaller than the radiation loss for antennas of length greater than $\lambda/2$, they are efficient radiating structures. The "efficiency" of an antenna can be defined as follows:

$$\eta = \frac{R_R}{R_o + R_R} \tag{21.6}$$

Representative values vary from almost 1.0 for hf antennas to 0.05 for vlf antennas.

If an antenna radiates energy vertically downward, the energy will be reflected and a current will be induced in the antenna because of the reflected energy. The manner in which this induced current affects the total antenna current depends upon its magnitude and phase with respect to the original current, which in turn depends upon the attenuation at reflection and the height of the antenna above the earth. The resultant variation in total antenna current for a constant input power as a function of height of the antenna above the earth can be regarded as a change in radiation resistance. This effect is illustrated in Fig. 21.2 for a simple dipole antenna.
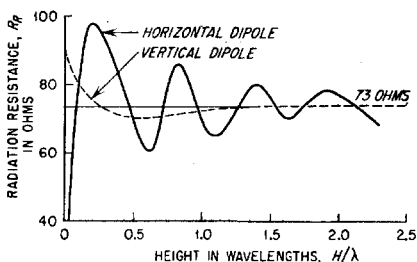
Fig. 21.2. Effect of antenna height above the earth on radiation resistance.

21.1d. Radiation Patterns, Directivity, Gain, and Effective Area. Antenna systems are usually designed to have radiation characteristics which vary with the direction from the antenna. A graphical representation of the radiation of an antenna as a function of direction is called the *radiation pattern*.

Radiation patterns may be specified in various planes through the center of the antenna, and such planes usually are chosen to include the maximum radiation. A pattern in the horizontal plane is called an azimuth pattern, while a vertical-plane pattern is called an elevation pattern. The plane of the pattern is sometimes related to the polarization; for example, an *E*-plane pattern is one measured in a plane con-

taining the electric field. Another classification of patterns depends on the distance from the antenna. If the distance is great enough the pattern is a so-called *Fraunhofer pattern* and its shape is independent of the distance. This distance is given by

$$r \geqq \frac{2D^2}{\lambda} \qquad (21.7)$$

where D = longest dimension of the antenna system in units consistent with r and λ

For distance less than that given by Eq. (21.7), the shape of the pattern varies with the distance from the antenna. Such patterns are called *Fresnel patterns*. Since antennas are ordinarily used only for ranges greater than that given by Eq. (21.7), Fraunhofer patterns are the ones of most interest. Equation (21.7) is useful in determining the minimum distance to use in measuring such patterns.

Two important properties of an antenna system are directivity and gain. Directivity is a function of the radiation pattern and is defined in Eq. (21.8).

$$D = \frac{P_r}{P_a} \qquad (21.8)$$

where P_r = power radiated per unit solid angle in a given direction
P_a = average power radiated per unit solid angle
Gain on the other hand includes the losses in the antenna and is given in Eq. (21.9).

$$G = 4\pi \frac{P_r}{P_t} \qquad (21.9)$$

where P_t = total power delivered to antenna terminals
For an antenna with negligible losses, gain, and directivity are the same.

The *effective area* of an antenna system is defined as

$$A_e = \frac{\lambda^2 G}{4\pi} \qquad (21.10)$$

This quantity is not readily related physically to many types of low-frequency antennas, but in the case of horns, reflectors, and lenses as used in the microwave region, the effective area and physical areas are nearly the same.

21.1e. Bandwidth, Impedance, and Mutual Coupling. The bandwidth of an antenna is defined as a range of frequencies within which its performance in respect to some characteristic conforms to a specified standard. The characteristic in question can be gain, beam width, impedance, or some related quantity. If impedance is chosen, for example, the definition has physical meaning only when a standard value is selected. A representative definition of bandwidth is that range of frequencies for which the reflected power from the terminals of the antenna does not exceed one-third of the incident power. The term broadband is also subject to arbitrary definition. An antenna system possessing fairly uniform characteristics over more than a 15 per cent band of frequencies is usually considered to be broadband.

The input or terminal impedance of an antenna depends on the physical configuration of the antenna and the frequency. The total resistive component of this impedance is the sum of radiation and loss resistances.

Antenna elements separated by not more than several wavelengths will interact and produce mutual coupling. The magnitude and phase angle of the mutual impedance resulting from this coupling will vary with antenna size, spacing between elements, frequency, and orientation of the elements. The real component of the coupling impedance can be either positive or negative, and, in general, the resonant frequency of an antenna will be modified by coupling with another antenna. The

coupling effect of a conducting surface on an antenna is the same as if the antenna were coupled to its image antenna, as reflected in the conducting surface.[1]

21.1f. Reciprocity. The properties of an antenna system are, in general, identical for transmission and reception. In particular, the radiation pattern of an antenna is the same regardless of whether the antenna is used for transmission or reception. The power transferred between two antennas will be the same regardless of which is used for transmission or reception if the generator and load impedance are conjugates of the transmitting and receiving antenna impedances in each case.

21.1g. Images and Ground Effects. Radio waves are reflected from the surface of the earth in the same manner as light is reflected from a mirror; however, since the earth is not a perfect conductor, the process of reflection can attenuate the reflected wave. The attenuation is a function of the frequency and the condition of the ground. In general, however, the attenuation increases as the frequency is increased. For a plane reflecting surface, the angle of reflection is equal to the angle of incidence. Rough or obstructed terrain can cause the reflected waves to depart from this angle, scattering much of the reflected energy in many directions.

Assuming the surface of the earth to be flat and perfectly conducting, Fig. 21.3a shows that the reflected wave can be considered to be radiating from an "image" of the antenna located a distance below the surface of the earth equal to the height of the transmitting antenna.

The reflected wave travels a distance different from the direct wave and at each point in space will combine with the directly transmitted wave with a particular relative phase. In the extreme cases assuming no attenuation, the resultant signal will be equal to twice the directly transmitted signal when the two waves combine in phase and will be zero when the two waves are in phase opposition. As the difference in path length will vary with the elevation angle between the antenna and the point of measurement, the resultant received signal strength will vary from zero to double amplitude as the point of observation changes elevation. This phenomenon has the effect of altering the signal strength pattern from the antenna in the elevation plane.

(a) IMAGE REFLECTION   (b) HORIZONTAL IMAGE

(c) VERTICAL IMAGE   (d) MULTIPLE IMAGES

Fig. 21.3. Images.

The exact manner in which the elevation pattern of the antenna is altered by ground reflection depends upon the polarization of the antenna. If a horizontally polarized antenna is to be used, the image antenna will be as shown in Fig. 21.3b. If a vertical antenna is used, the image antenna will be as shown in Fig. 21.3c. It should be noted that the polarities of corresponding points on the real and image antennas are reversed. This is necessary for the ground plane to be at zero potential at all times. The elevation pattern of such antenna systems can be regarded as equivalent to two antennas 180° out of phase in the horizontal case and in phase in the vertical case, separated a distance equal to twice the height of the real antenna above the ground. The equation for signal intensity as a function of elevation angle is, for horizontal
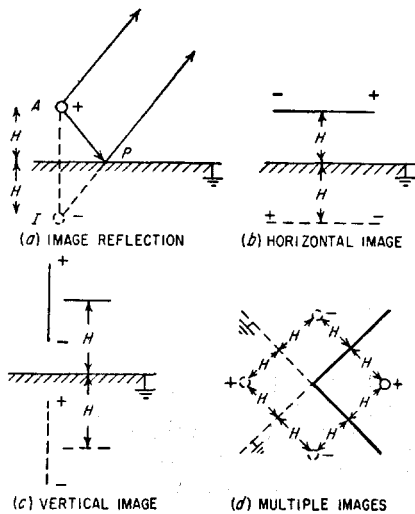
[1] For charts giving mutual impedance values see Terman, *op. cit.*, sec. 11, or J. D. Kraus, "Antennas," chap. 10, McGraw-Hill Book Company, Inc., New York, 1950.
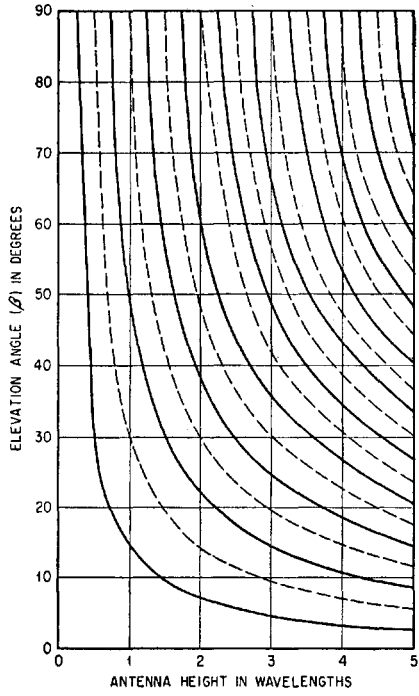
(a)
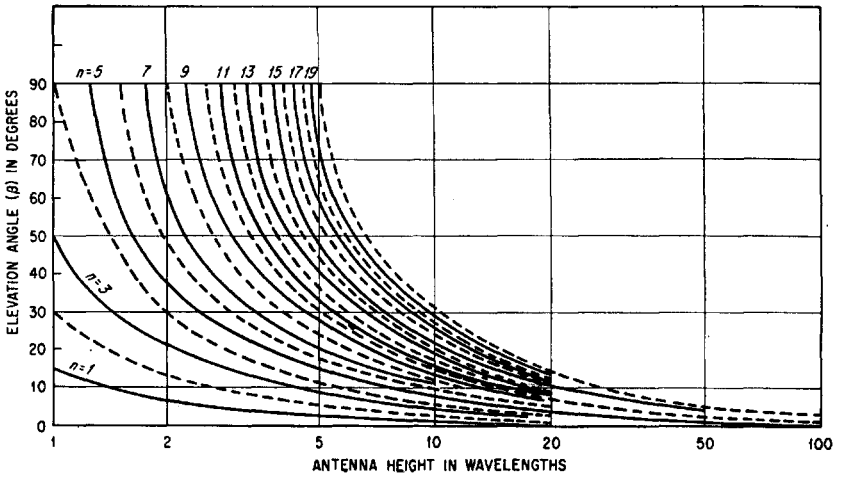
(b)

FIG. 21.4. Maxima and null angles caused by ground reflection.

antennas of any length or for vertical antennas which are an even number of half wavelengths long,

$$G(\beta) = 2F(\beta) \sin \left(\frac{360°H}{\lambda} \sin \beta \right) \qquad (21.11)$$

where $G(\beta)$ = elevation pattern of antenna system including effect of reflection
$\quad F(\beta)$ = normal elevation pattern of antenna without ground reflection
$\quad H$ = antenna height
$\quad \beta$ = angle of elevation measured from horizontal
$\quad \lambda$ = wavelength in units consistent with H

The angular positions of the maxima and nulls can be obtained from Eq. (21.12).

$$\sin \beta = \frac{n\lambda}{4H} \qquad (21.12)$$

This will give a maximum for odd integral values of n and a minimum for even integral values of n.

For a vertical antenna an odd number of half wavelengths long, the positions of the maxima and nulls obtained from Eqs. (21.11) and (21.12) should be reversed. Figure 21.4 gives the angular positions of nulls and maxima due to reflection for various antenna heights. The solid lines correspond to maxima for horizontal antennas and for vertical antennas an even number of half wavelengths long. The dashed lines correspond to minima. These are reversed for vertical antennas an odd number of half wavelengths long.

Figure 21.4 can be used to sketch the elevation pattern of an antenna system, using the following rule to determine the lobe width:

The width between half-power points of any lobe in degrees is approximately equal to half of the difference between adjacent minima in degrees.

A reflecting surface as shown by the heavy lines in Fig. 21.3d is sometimes used in an antenna system. In this case, there are three image antennas as shown. It should be noted that the locations and polarities of the images are such as to maintain the reflecting *planes* at ground potential.

Example 21.1

Determine the required height in wavelengths of a horizontal antenna whose first lobe is to be at an elevation angle of 10°. Sketch the entire elevation pattern assuming the free space pattern of the antenna is omnidirectional.

Solution

1. Determine the height from Fig. 21.4. This is seen to be $H \approx 1.45\lambda$ to produce the first lobe at 10°. This can be checked from Eq. (21.12) as follows:

$$\sin 10° = 0.174 = \frac{n\lambda}{4H}$$

FIG. 21.5. Pattern for Example 21.1.

Since for first maximum $n = 1$, $H = \dfrac{\lambda}{0.696} = 1.44\lambda$.

2. Determine the angles of other lobes and the lobe widths.

From Fig. 21.4 the other lobes are found to be at approximately 32 and 59° with nulls at 20, 44, and 90°. The width at half power points of the first lobe is therefore 10°, the second 12°, and the third 23°. This pattern is sketched in Fig. 21.5.

21.1h. Antenna Feed Systems. Each antenna system requires a transmission line to couple power to it from the transmitter. The type of line used will depend

upon the type of antenna, method of connection, and frequency of operation. Simple, single-wire feeder systems can be used for low-frequency antennas, parallel-wire transmission lines are used at high frequencies, coaxial lines are used at vhf and uhf, and waveguide and coaxial transmission lines are used at microwaves. The primary requirement is for the feed line to couple as much power as possible from the transmitter into the antenna. Characteristics of conventional transmission-line systems are given in Sec. 20.

21.1i. Apertures and Illumination. An antenna may be regarded as consisting of an illuminated aperture in the same sense that such apertures are used in optics. Specifically, at microwaves antenna systems are often composed of shaped metallic
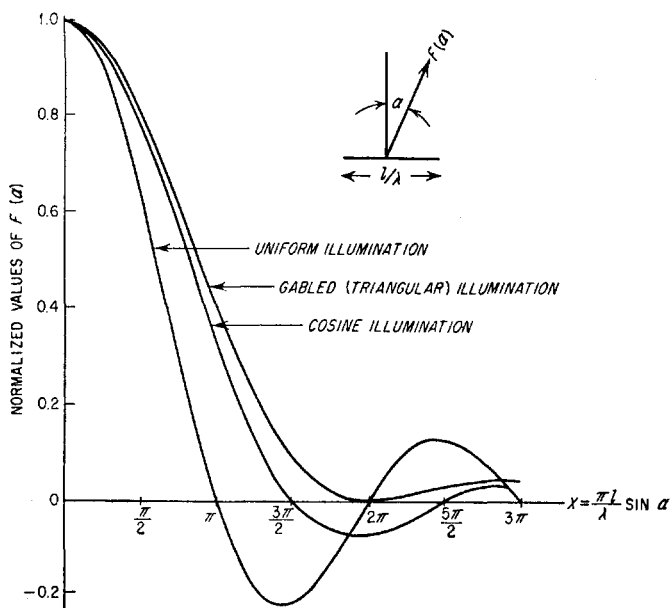
Fig. 21.6. Patterns for rectangular apertures with various types of illumination having zero intensity beyond the aperture edge.

reflecting surfaces and a transmission line terminated in a dipole or horn to "illuminate" the reflector with the radiated signal. The pattern of such antenna systems depends upon the manner in which the amplitude and phase of the radiation are distributed over the aperture subtended by the reflecting surface.[1] In general, however, pattern and illumination functions are Fourier transforms of each other.

The effects of various amplitude distributions are given in Fig. 21.6. In general, tapered illumination broadens the main lobe and suppresses the magnitude of the side lobes. Illuminations which taper at the edges of the aperture to −10 or −15 db below the center amplitude are frequently used and represent a reasonable compromise between beam width and side-lobe level. Symmetrical amplitude distributions are ordinarily used, but even an unsymmetrical amplitude distribution will still yield a

---

[1] For a detailed discussion of the techniques employed in calculating patterns of apertures and synthesizing antenna patterns by phase and amplitude distributions across an aperture, many excellent references are available such as S. Silver, "Microwave Antenna Theory and Design," McGraw-Hill Book Company, Inc., New York, 1949, and *Bell Telephone System Tech J.*, April, 1947, pp. 219.

symmetrical pattern. If the phase distribution is unsymmetrical, however, the pattern will be unsymmetrical.

A linear phase variation causes the beam to be shifted by an angle θ given by

$$\sin \theta = \frac{k\lambda}{2\pi} \tag{21.13}$$

where k is the phase variation per unit length of aperture and θ is measured from the normal to the antenna.

It is possible to use this effect to cause the beam of an antenna to scan, or move in angle, by changing the phase variation along the aperture.

| $l=2\lambda$ | $l=\lambda/2$ | $l=\lambda$ | $l \ll \lambda/4$ | $l=\lambda/4$ | $l=\lambda/2$ | $l=\lambda$ | $l+x=\lambda/4$ |

(a) UNGROUNDED ANTENNAS   (b) GROUNDED ANTENNAS   (c) EFFECT OF TOP LOADING
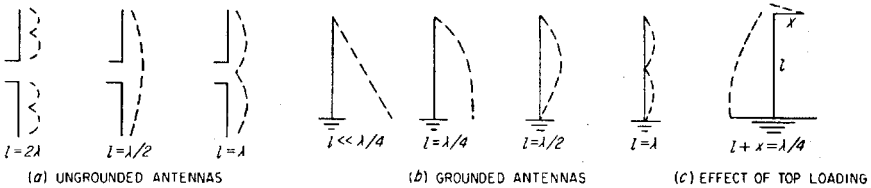
FIG. 21.7. Antenna current distributions.

Nonlinear phase variations can alter the amplitude distribution of the pattern as well as shift the main lobe position.

If an antenna system is composed of a reflector and illuminating field, two patterns are of interest. The *primary pattern* is the pattern of the feed and determines the amplitude and phase distribution of the illumination of the reflector. The *secondary pattern* is the radiation pattern of the illuminated aperture and depends upon the primary pattern and the geometry of the reflector surface.

21.1j. Current Distribution. A useful parameter for low-frequency antennas is the current distribution. The current distribution along a radiating element is analogous to the field distribution across an aperture and, in much the same manner, can be used to calculate its pattern. Current distribution depends upon the ratio of antenna length to wavelength, ground proximity, and antenna loading. Examples of antenna current distributions are shown in Fig. 21.7. Figure 21.8 shows, for reference purposes, an expanded plot of the sinusoidal distribution for a grounded $\lambda/4$ antenna. These current distributions are for resonant antennas or antennas acting as unterminated transmission lines (see Sec. 20) with a standing wave of current distributed as shown. An "end effect" in resonant antennas causes the physical length to be approximately 5 per cent less than the theoretical resonant length. These cur-
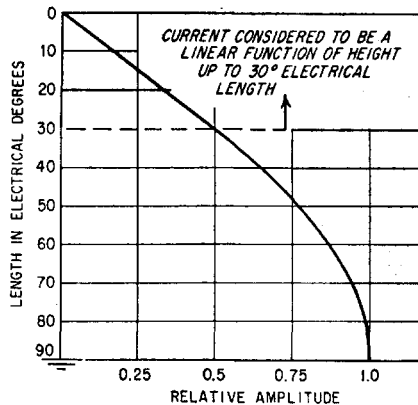
FIG. 21.8. Current distribution on a grounded vertical radiator with an effective height of one-quarter wavelength.

rent distributions can be modified by artificially loading the antenna with lumped inductance or capacitance. In a resonant structure, series inductance decreases the separation between current minima and reduces the effective length of the antenna.

A nonresonant antenna is terminated in such a manner that there are no reflections from the ends of the antenna, standing waves do not exist, and neglecting losses, the current is of uniform amplitude with 2π radians phase shift per wavelength.

Figure 21.7c shows the manner in which top loading by a capacitor element designated by the length x can affect the current distribution. If $x = \lambda/12$ for example, then top loading accounts for 30° of electrical length and the current distribution on the vertical radiator would be as shown in Fig. 21.8 between 30° and 90°.

21.2. Basic Antennas. The performance of a physical antenna system can be analyzed for design purposes by dividing the over-all structure into several basic component parts. A convenient selection of component parts, possessing physical significance is the following:

Over-all antenna = basic source × array factor × ground reflector factor

The basic source is the most elementary antenna that can be physically recognized in the over-all system. For example, in an antenna composed of a number of horizontal radiators, the basic source would be a single horizontal radiator. The array factor in this case would be the effect on the over-all antenna system of the combination of the horizontal radiators into a single system. The ground-reflection factor would be the further effect on the over-all performance of ground-plane reflection. In this manner, a complicated physical structure can be analyzed or designed by several stages, greatly simplifying the process.

The ground-reflection factor is discussed in Sec. 21.1g. The array factor and examples showing the combination of these effects are discussed in Sec. 21.3. It is the purpose of this section to discuss the pattern and characteristics of the basic antenna types which are fundamental to more complicated systems. These basic antennas are often used individually for the required antenna system. In this case, the array factor is unity and the performance of the system depends only on the performance of the basic antenna and ground-reflection effect.

(a) PLANE OF DOUBLET

(b) PLANE NORMAL TO DOUBLET

FIG. 21.9. Field from a vertical doublet antenna as a function of elevation angle.

The fundamental antenna element is the doublet, an infinitesimally short straight antenna in free space with a nonuniform current distribution along its entire length.

The total field from a physical antenna can be found by dividing the antenna into a large number of doublets with current amplitudes changing from doublet to doublet so as to approximate the non-uniform current distribution of the physical antenna. The contribution from each doublet can then be added vectorially to determine the composite pattern.

The electric field of a doublet antenna is given by Eq. (21.14).

$$E(\theta) = \frac{60\pi lI}{r\lambda} \cos \theta \qquad (21.14)$$

where $E(\theta)$ = field intensity, volts/m, measured in a plane containing the axis of the doublet
θ = angle measured from a reference perpendicular to the doublet
l = length of doublet, m
I = current in the doublet, amp
r = distance from the doublet to the point of measurement, m
λ = wavelength, m

A doublet and associated pattern are shown in Fig. 21.9. Equation (21.14) gives the radiation field only, which decreases as r^{-1}. In addition, there is an induction magnetic field which decreases as r^{-2} and an additional static field which decreases as r^{-3}. The radiation field is usually the only one of importance in antenna design.

The field in a plane perpendicular to the doublet axis is a constant, independent of angle.

The radiation resistance of a doublet antenna is given by Eq. (21.15).

$$R_r = 789 \left(\frac{l}{\lambda}\right)^2 \quad \text{ohms} \tag{21.15}$$

where l and λ are in identical units.
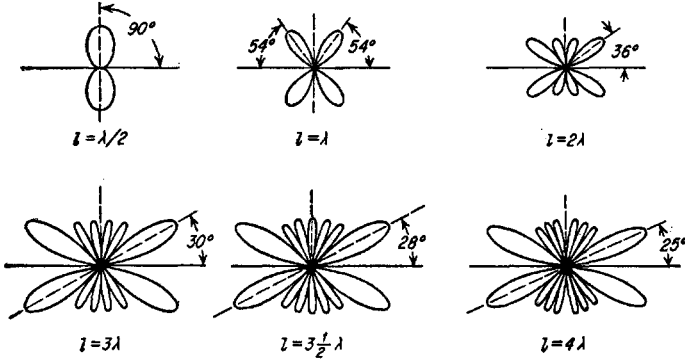
FIG. 21.10. Free-space patterns of long-wire resonant horizontal antennas with standing wave current distribution.

21.2a. Horizontal Antennas. The radiation pattern from a resonant end-fed horizontal wire remote from the earth without ground effects is given by Eq. (21.16).

$$E(\theta) = \frac{60I}{r} \frac{\sin\left[\dfrac{\cos\left(\dfrac{\pi l}{\lambda}\cos\theta\right)}{\sin\theta}\right]}{} \quad \text{volts/m} \tag{21.16}$$

where I = peak current amplitude, amp

r = distance, m

l, λ = antenna length and wavelength in similar units

θ = angle from axis of wire

The sine term is used for antennas an even multiple of $\lambda/2$ long; the cosine term is used for antennas an odd number of $\lambda/2$ long.

Representative patterns are shown in Fig. 21.10. The angular positions of maxima and minima as a function of wire length in wavelengths are plotted from Eq. (21.16) in Fig. 21.11. As the length is increased, the number of lobes increases, the lobe width decreases, and the angle of the first lobe with the wire decreases. For lengths an odd multiple of $\lambda/2$, there is a lobe normal to the wire. For lengths an even multiple of $\lambda/2$, there is a null normal to the wire. Equation (21.16) applies for antenna lengths up to approximately 10 wavelengths. For longer antennas the lobes pointing away from the feed end will be larger than the corresponding lobes in the opposite direction.
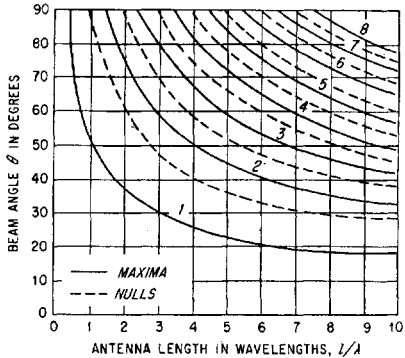
FIG. 21.11. Chart for determining beam patterns of end-fed long-wire resonant antennas.

The radiation pattern of a center-fed horizontal (or vertical) wire of any length in free space is given by Eq. (21.17).

$$E(\theta) = \frac{60I}{r} \frac{\cos\left(\frac{\pi l}{\lambda}\cos\theta\right) - \cos\left(\frac{\pi l}{\lambda}\right)}{\sin\theta} \tag{21.17}$$

where the units are the same as given for Eq. (21.16). For resonant antennas an odd number of $\lambda/2$ long, Eq. (21.17) reduces to the same form as that given by Eq. (21.16). However, for resonant antennas an even number of $\lambda/2$ long, the patterns differ because of the resulting unsymmetrical current distribution in the end-fed case.

The choice between center and end feed for antenna systems depends primarily upon the length of the antenna in wavelengths, the frequency band over which the
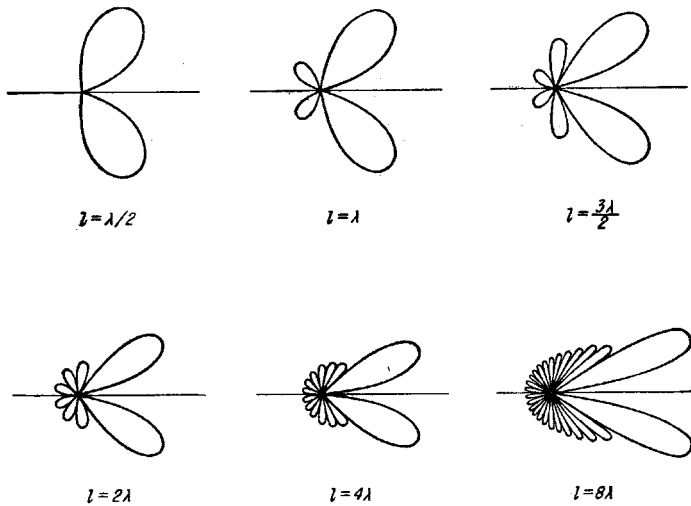
FIG. 21.12. Radiation patterns of nonresonant horizontal wires.

antenna is to operate, and the desired pattern. In center-fed antennas the currents in each half of the antenna are in phase and have a symmetrical amplitude distribution with respect to the feed point. In end-fed antennas the currents in each half of the antenna can be either in phase or out of phase depending upon the electrical length of the antenna. This can result in different directional characteristics for the same length of antenna, depending upon the type of feed. A resonant[1] feed line is usually connected so that a voltage maximum on the line corresponds to the point of connection to a resonant (length a multiple of $\lambda/2$) end-fed antenna. A resonant center-fed antenna can require either a voltage or current maximum on the feed line at the point of connection to the antenna. The pattern of an end-fed antenna may not be symmetrical if antenna losses are not negligible. In this case the pattern will "tilt" away from the feed end, and the lobe amplitudes will be larger away from the feed end. The center-fed antenna has the advantage of providing a balanced termination for the feed line decreasing the radiation from the feeder system as compared to the end-fed case.

Traveling-wave antennas are long-wire antennas terminated so as to eliminate or minimize reflections and the resultant standing wave pattern. The lobe structure is

[1] See Sec. 20.

similar to that given in Figs. 21.10 and 21.11 for the direction toward the termination, as long as the antenna is several wavelengths long. The lobe structure in the direction away from the termination is considerably decreased in amplitude, resulting in an asymmetric, and usually more useful, pattern. Representative patterns of non-resonant, horizontal wires are shown in Fig. 21.12. The variation in amplitude of successive lobes is given by Fig. 21.13. The expression for the directional pattern in a plane containing the wire for the nonresonant case is

$$E(\theta) = \frac{60I}{r} \frac{\sin \theta \sin \left[\frac{\pi l}{\lambda} (1 - \cos \theta) \right]}{1 - \cos \theta} \tag{21.18}$$

where θ = angle from axis of wire
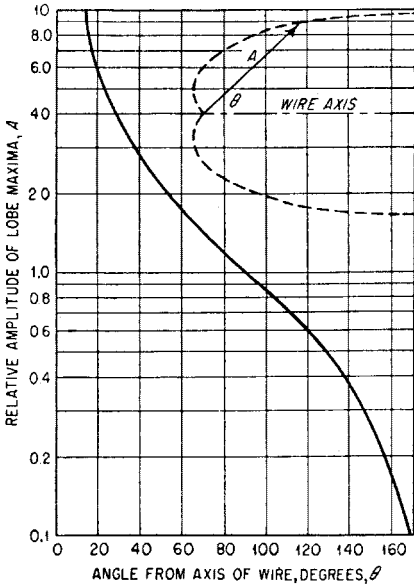
Fig. 21.13. Relative amplitude of lobe maxima as a function of angle from the axis of a nonresonant long-wire antenna.
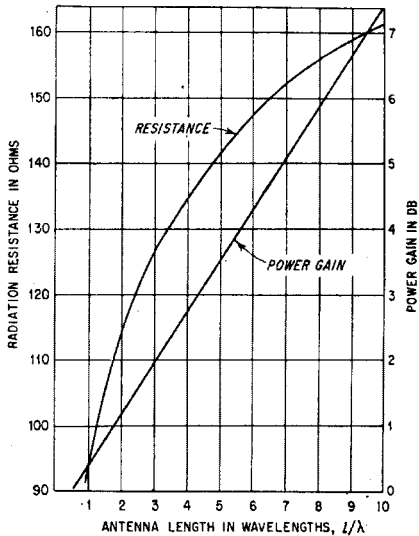
Fig. 21.14. Radiation resistance and gain of a horizontal long-wire antenna neglecting ground reflections.

Equation (21.18) assumes zero attenuation of the traveling wave and a phase velocity equal to the velocity of light. The effect of finite attenuation is to "round out" the pattern by filling in the nulls and decreasing the peak amplitudes of the lobes. The effect of a phase velocity less than the velocity of light is to tilt the main lobe toward the wire, decrease the main-lobe beam width, and introduce additional minor lobes. A decrease in phase velocity of 40 per cent affects the tilt angle and major-lobe beam width by less than 20 per cent.

The effect of ground reflection on horizontal antennas can be accounted for by the use of Eq. (21.11).

The radiation resistance of a resonant horizontal antenna is a function of the antenna height above ground and the ratio of length to wavelength. The radiation resistance of a horizontal wire, neglecting ground effect, is given in Fig. 21.14. The effect of ground reflection on the radiation resistance of a horizontal antenna is a rapidly changing function of height. For increased height, the radiation resistance oscillates

about the nominal value determined from Fig. 21.14 with the variations decreasing in amplitude as the height is increased.[1]

Fig. 21.14 also gives the power gain of a horizontal long-wire antenna compared with a half wavelength wire.

Example 21.2

Determine the required length and height of a horizontal wire antenna, in order to place the main lobe at an elevation angle of approximately 30° and a null normal to the antenna. Sketch the pattern.

--- *GROUND REFLECTION PATTERN*
--- *FREE SPACE PATTERN MODIFIED*
___ *BY GROUND REFLECTION*
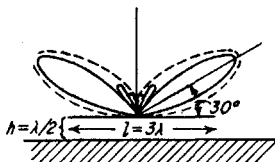
FIG. 21.15. Composite vertical pattern for a horizontal antenna, including ground reflection.

Solution

1. An examination of Fig. 21.11 shows that a length of 3λ will produce the major-lobe maximum at an angle of approximately 30° and will result in nulls at 50, 70, and 90°.
2. The height should now be chosen so that the effect of ground reflection does not seriously modify the pattern. This will be accomplished if the major lo befrom Fig. 21.4 includes the major lobe at 30° in the free-space pattern. A height of λ/2 is acceptable. The composite pattern is sketched in Fig. 21.15.

An example of a nonresonant antenna system is the Beverage, or wave antenna,[2] shown in Fig. 21.16. This antenna has directional properties in the direction of the termination and is used to receive *vertically* polarized signals. This antenna depends upon poor ground conductivity tilting the vertical wavefront in order to provide a horizontal component of electric field. This antenna has greatest application at low frequencies.

21.2b. Vertical Antennas. The free-space orientation of an antenna does not affect the pattern relative to the antenna, assuming ground reflections are not significant. The patterns of Fig. 21.10 apply, therefore, for long vertical radiators in free space, with a sinusoidal

FIG. 21.16. Beverage antenna.

standing wave current distribution. Where ground reflections are significant, Eq. (21.11) gives the resultant modified pattern. Grounded end-fed vertical antennas have a pattern equivalent to a center-fed antenna of twice the length as the vertical radiator with symmetrical current distribution.

In practice, vertical radiators are usually short compared to the wavelength. The expression for the radiation pattern in the elevation plane including the effect of ground reflection is given by Eq. (21.19), and representative patterns are shown in Fig. 21.17.

$$E(\beta) = \frac{60I}{r} \left[\frac{\cos\left(\frac{2\pi l}{\lambda}\sin\beta\right) - \cos\left(\frac{2\pi l}{\lambda}\right)}{\cos\beta} \right] \qquad (21.19)$$

[1] Radiation resistance in any particular case can be determined by considering the ground reflection to be produced by an image antenna spaced twice the height above ground from the real antenna. The real component of the mutual impedance between the real and image antenna accounts for the variation in radiation resistance, usually a reduction for horizontal antennas. Mutual impedance calculations are usually complex but are described in the following: G. H. Brown, *Directional Antennas, Proc. IRE*, vol. 25, p. 78, January, 1937, and P. S. Carter, Circuit Relations in Radiating Systems and Applications to Antenna Problems, *Proc. IRE*, vol. 20, p. 1004, June, 1952.

[2] H. H. Beverage, C. W. Rice, and E. W. Kellog, The Wave Antenna; a New Type of Highly Directive Antenna, *Trans. AIEE*, vol. 42, pp. 215–266, 1923.

where β = elevation angle from horizontal

l/λ = antenna height in wavelengths above ground plane

I = rms value of current at a maximum current point, amp

To refer field strength to the base current, Eq. (21.19) should be divided by sin $2\pi l/\lambda$.
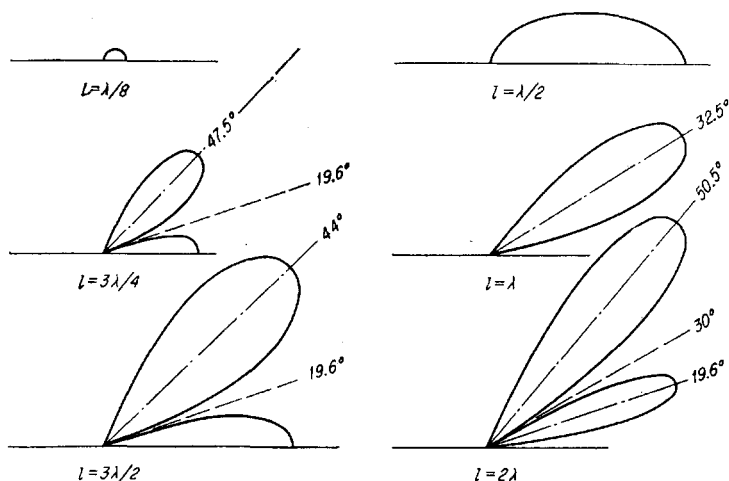
FIG. 21.17. Representative patterns of vertical grounded radiators with sinusoidal current distribution. (Patterns are shown for any quadrant in the vertical plane. The complete pattern is that swept out by rotation of the patterns shown about the axis of the antenna.)

The angular locations of nulls and maxima are given in Fig. 21.18 as a function of l/λ for multiples of one-half wavelength. Equation (21.19) assumes a sinusoidal current distribution.

The radiation resistance referred to the maximum current point on the antenna is shown in Fig. 21.19 as a function of antenna height.[1] To obtain the radiation resistance referred to the base current, the values obtained from Fig. 21.19 should be divided by sin^2 $(2\pi l/\lambda)$.

At low frequencies where vertical antennas are electrically very short, very low values of radiation resistance are obtained (for example, 0.3 ohm for $l = \lambda/36$). This requires the loss resistances to be very low for the antenna to have a reasonable efficiency.[2]

FIG. 21.18. Null and maxima chart for vertical end-fed radiators.

21.2c. V Antennas.[3] The pattern of a V antenna is the resultant obtained by combining the patterns from the two single-wire antennas forming the V. These patterns will vary, depending on the length of the wire in wavelengths, the apex

[1] See G. H. Brown, A Critical Study of the Characteristics of Broadcast Antennas, as Affected by Antenna Current Distribution, *Proc. IRE*, vol. 24, pp. 48–81, January, 1936.

[2] For a general but complete reference on the design of low-frequency antennas, see E. A. Laport, "Radio Antenna Engineering," McGraw-Hill Book Company, Inc., New York, 1952.

[3] See P. S. Carter, W. W. Hansell, and N. E. Lindenblad, Development of Directive Transmitting Antennas by RCA Communication, Inc., *Proc. IRE*, October, 1931, p. 1773, and P. S. Carter, Circuit Relations in Radiating Systems, *Proc. IRE*, June, 1932, p. 1004.

angle of the V, and the termination on each side of the V. Equation (21.16) gives

the radiation pattern of a single end-fed wire with standing waves (no termination). Equation (21.18) gives the radiation pattern of a single end-fed wire terminated with the characteristic impedance of the wire (traveling wave only). Figure 21.20a shows an unterminated V. Figure 21.20b shows a terminated V. The apex angle 2α is chosen to be equal to twice the main-lobe angle of the single wire to get the resultant main lobe aligned in the horizontal plane of the V, assuming free space conditions. To elevate the main lobe above the plane of the V it is necessary to reduce this angle somewhat. The apex angle 2α is shown in Fig. 21.21 as a function of the length of one side of the V in wavelengths. To align the main lobe at a higher elevation angle the apex angle should be reduced slightly. The amount of reduction depends upon the length of the sides and the mutual coupling between the sides. A 10 per cent reduction in apex angle will elevate the beam maximum approximatley 10°.

FIG. 21.19. Radiation resistance of a grounded vertical radiator referred to a current maximum.

(a) UNTERMINATED V

(b) TERMINATED V

(c) HORIZONTAL AND VERTICAL ARRAYS OF V ANTENNAS
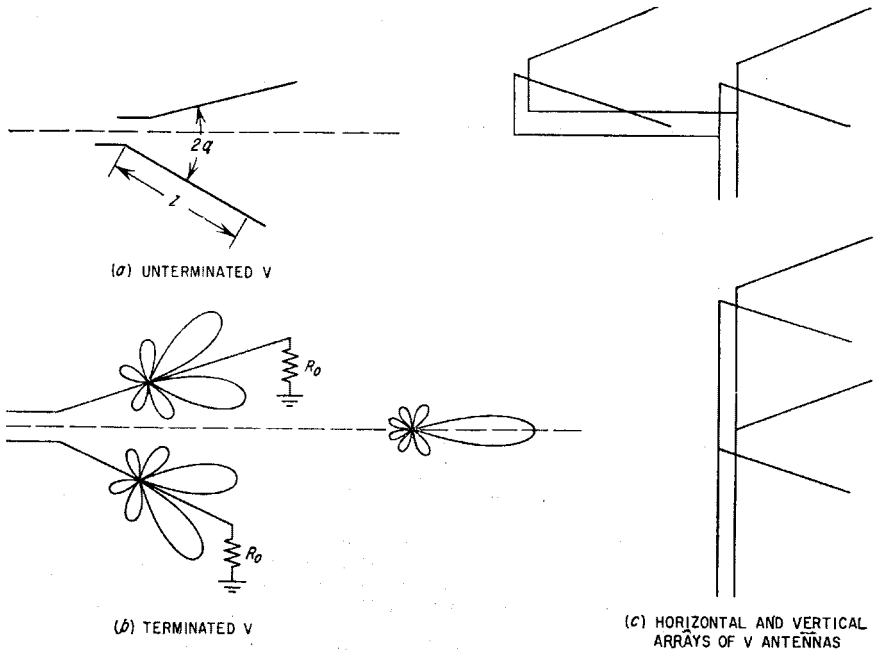
FIG. 21.20. V antennas.

V antennas can be stacked horizontally or vertically to increase the directivity. Rules for computing such array patterns are given in Sec. 21.3.

The characteristic impedance of a V antenna depends upon the apex angle 2α and the ratio between length and diameter of the legs and will vary between several hundred ohms to approximately one thousand ohms, increasing as the length to diameter ratio is increased.

The gain of a V antenna is approximately 3 db greater than the gain of a long wire antenna of length equal to the length of one side of the V. See Fig. 21.14.

21.2d. Rhombic Antennas.[1] A rhombic consists of four nonresonant wires connected as shown in Fig. 21.22. This configuration provides greater gain and directivity than a simple V antenna and is convenient to terminate with a dissipative two-wire transmission line or a resistance. Rhombic antennas are characterized by a relatively broad (for example, 5:1) impedance bandwidth. The radiation pattern is seldom uniform over more than a 2:1 band, however.

FIG. 21.21. Apex angle 2α for main lobe at an elevation angle of zero degrees as a function of the length in wavelengths of a V antenna.

The radiation pattern is characterized by many lobes in the plane of the rhombic and the vertical plane. Rhombic antennas are most useful in applications where the elevation angle of the maximum lobe can be less than 30° from the plane of the rhombic.

The multiple-wire construction shown in Fig. 21.22b lowers the characteristic impedance and improves the impedance bandwidth.

(a) DIRECTIONAL CHARACTERISTICS OF RHOMBIC LEGS   (b) RESULTANT PATTERN

FIG. 21.22. Rhombic antenna.

The polarization of the radiation from a rhombic is horizontal in the main lobe and can be either horizontal, vertical, or a combination of both in the other lobes.

Two bases can be used for rhombic antenna design. The first aligns the main lobe with the desired elevation angle β. The second provides the maximum gain at this angle but does not necessarily place the axis of the main lobe at this angle. Figure 21.23 gives approximate design data for the height H/λ of the rhombic above

[1] For detailed information on the design of rhombic antennas see: D. Foster, Radiation from Rhombic Antennas, *Proc. IRE*, October, 1937, p. 1327. E. Bruce, A. Beck, and L. R. Lowry, Horizontal Rhombic Antennas, *Proc. IRE*, January, 1935, p. 24. Rhombic Transmitting Aerial Efficiency, *Wireless Engr.*, May, 1941, p. 180. A. E. Harper, "Rhombic Antenna Design," D. Van Nostrand Company, Inc., Princeton, N.J., 1941.

the earth in wavelengths, half the included angle of the rhombic ϕ, and the length l/λ of each leg in wavelengths for a compromise design between these two requirements.[1] To align the major lobe at the angle β, values of l/λ should be decreased approximately 10 per cent from those given by the curve. To obtain the maximum gain at the angle β, values of l/λ should be increased approximately 10 per cent from those given by the curve.

The horizontal beam width of the main lobe between nulls decreases from approximately 50° for a leg length of two wavelengths to approximately 25° for a leg length of seven wavelengths.

The power gain of a rhombic antenna varies with the main lobe elevation angle and the height of the antenna above

(a) CIRCULAR LOOP   (b) SQUARE LOOP
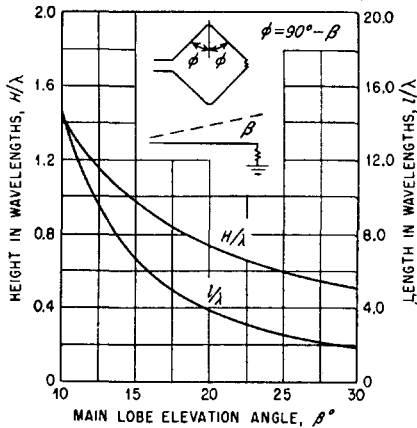
Fig. 21.23. Design chart for a rhombic antenna.

(c) ALFORD LOOP   (d) CLOVERLEAF

COAXIAL FEED LINE

PLANE OF LOOP

$E(\theta)$

(e)

Fig. 21.24. Loop antennas.

the ground. Representative values vary between 6 and 12 db for leg lengths between 1 and 5 wavelengths. These figures are referred to a dipole and neglect loss in the termination.

21.2e. Loop Antennas.[2] Various types of loop antennas are shown in Fig. 21.24. If the loop circumference is small compared to the wavelength, the radiation properties are relatively independent of the specific loop geometry. Uniform, in-phase currents are assumed to exist around the loop conductor. The Alford loop, useful at uhf frequencies, deviates from the constant-current assumption to make the loop elements sections of resonant transmission lines with a current maxima at the center of each leg. This allows the use of longer sides, resulting in greater gain, directivity, and a more efficient antenna.

---

[1] This is one possibility out of the several combinations of height, included angle, and leg length to satisfy a particular design requirement. For additional discussion see: J. D. Kraus, "Antennas," pp. 408–412, McGraw-Hill Book Company, Inc., New York, 1950. E. A. Laport, *op. cit.*, pp. 315–339. E. Bruce, A. C. Beck, and L. R. Lowry, Horizontal Rhombic Antennas, *Proc. IRE*, vol. 23, pp. 24–26, January, 1935.

[2] For additional data see: E. M. Williams, Radiating Characteristics of Short Wave Loop Aerials, *Proc. IRE*, vol. 28, p. 480, October, 1940. Kraus, *op. cit.*, p. 155. D. Foster, "Loop Antennas with Uniform Current," *Proc. IRE*, vol. 32, p. 603, October, 1944. Andrew Alford and A. G. Kandoian, Very High Frequency Loop Antennas, *Trans. AIEE*, vol. 59, p. 843, 1940. P. H. Smith, Cloverleaf Antennas for F. M. Broadcasting, *Proc. IRE*, vol. 35, p. 1556, December, 1947.

The pattern of a small loop is constant in the plane of the loop. The pattern in a plane perpendicular to the loop is given by Eq. (21.20).

$$E(\theta) = \frac{120\pi^2 I A}{r\lambda^2} \sin (\theta) \qquad (21.20)$$

where θ = angle from normal to plane of loop
$\quad A$ = area of loop
$\quad \lambda$ = wavelength in units consistent with A
and the other units are as given for Eq. (21.16).
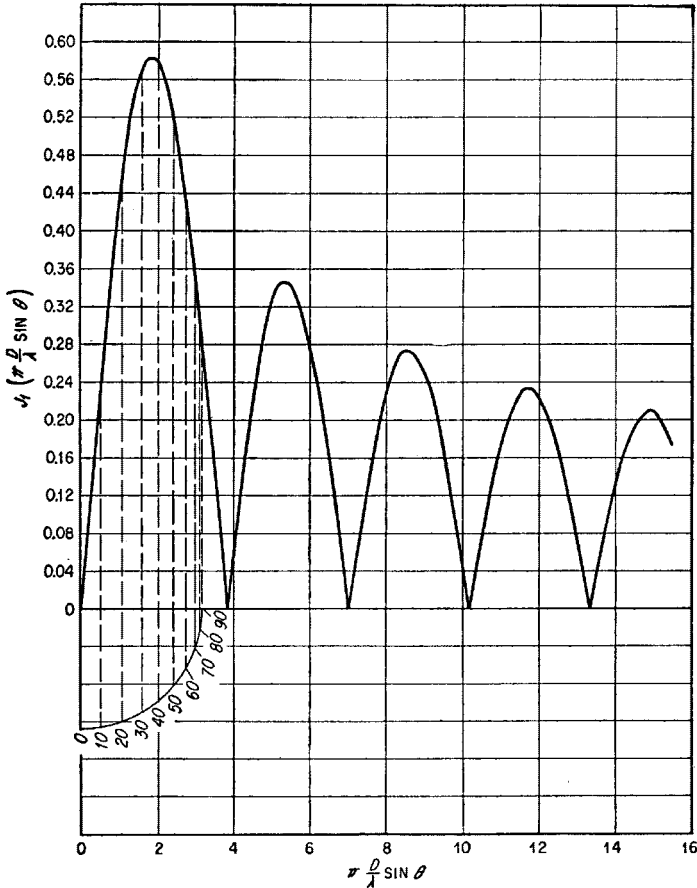
FIG. 21.25. Universal design chart for loop antennas.

This field pattern is illustrated in Fig. 21.24e.

As the loop circumference becomes somewhat greater than one wavelength, the pattern breaks up into successive minor lobes. Figure 21.25 gives the relationship between loop circumference and radiation patterns[1] for circular loops. Regardless of loop size the patterns are still uniform in the plane of the circular loop. For large square loops, however, the pattern in the plane of the loop is no longer uniform.

[1] Adapted from Foster, "Loop Antennas with Uniform Current," *Proc. IRE*, vol. 3ℒ, p. 603, October, 1944.

Figure 21.25 is a universal design curve giving the angular locations of nulls, maxima, and magnitude of the maxima. To use the chart, a circular quadrant of radius $\pi D/\lambda$ is described about the point 0,0 in the figure. D/λ is the diameter of the loop in wavelengths, and the example shown is for a diameter of one wavelength. The quadrant is then divided into angular increments proportional to θ, and the corresponding magnitudes are determined from the graph for each particular angle. The result, for several D/λ ratios, is given in Fig. 21.26.

The radiation resistance depends upon the circumference of the loop in wavelengths and is given in Fig. 21.27 for circular loops.[1] For multiple-turn loops, values obtained from Fig. 21.27 should be multiplied by the square of the number of turns.

Horizontal loops are often stacked vertically to obtain an omnidirectional

Fig. 21.26. Far field patterns of circular loops with uniform inphase current.
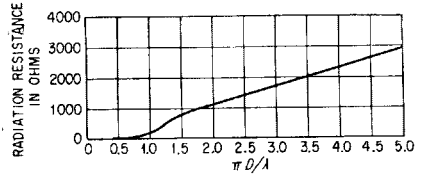
Fig. 21.27. Radiation resistance of circular loop antennas.

horizontal pattern with greater vertical directivity than would be possible with a single loop. The resultant pattern can be determined from the principles discussed in Secs. 21.3a and 21.3b.

Unbalanced current in a loop antenna can cause the pattern to be distorted. A split electrostatic shield, balancing the capacitance to ground of each half of the loop, is often used to eliminate this pattern distortion. In loops with an area large compared to λ^2, the currents will be nonuniform causing additional pattern distortion.

(a) SINGLE DIPOLE   (b) SINGLE FOLDED DIPOLE   (c) MULTI-ELEMENT FOLDED DIPOLE   (d) THICK DIPOLE
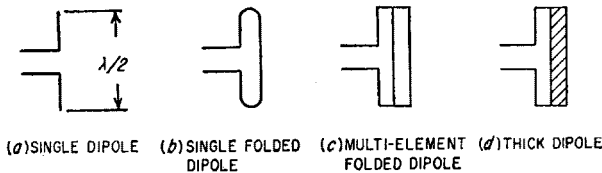
Fig. 21.28. Dipole antenna.

21.2f. Folded Dipole Antennas. Single dipole, two-wire folded dipole and multiple dipole antennas are shown in Fig. 21.28. The parallel dipoles are usually identical and carry currents of equal phase and amplitude. The spacing between dipoles is small compared to the dipole length.

A folded dipole is fed from a balanced two-wire line and provides a convenient means of matching the antenna impedance to the transmission line. The input resistance increases as the square of the number of parallel dipoles. The input

[1] Adapted from Kraus, *op. cit.*, p. 169, or see Foster, "Loop Antennas with Uniform Current," *Proc. IRE*, vol. 32, p. 603, October, 1944.

resistance of a single dipole is 73 ohms; a double dipole is 292 ohms; a triple dipole, 657 ohms; etc.

A single thick dipole as shown in Fig. 21.28d is used to decrease the variation of impedance with frequency.

21.2g. Ground-plane and Sleeve Antennas.[1]   The vertical antennas discussed in Sec. 21.2c can be converted into additional useful forms at higher frequencies where the dimensions of the antennas are physically convenient.   For example, a vertical radiator with an artificial ground plane consisting of a metallic disk is shown in Fig. 21.29a.   For convenience, ground-plane disks are usually less than several wavelengths in diameter, but because they do not duplicate the condition of an infinite ground plane, the antenna input impedance is a function of the ground-plane diameter.
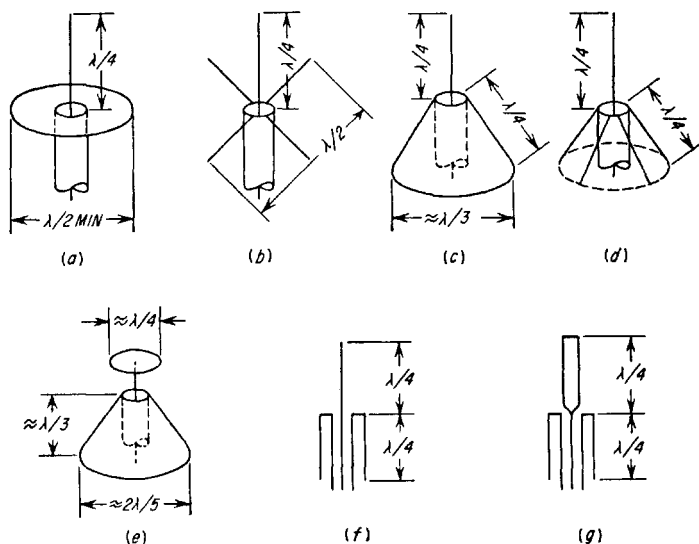
FIG. 21.29. Ground plane and sleeve antennas.

The horizontal pattern of the antenna shown in the figure is omnidirectional, and the polarization is vertical.   These antennas are conveniently fed from a coaxial transmission line.   The vertical patterns for the antennas in Fig. 21.29a and b do not have maxima normal to the antenna but, because of the effect of the finite ground plane, the maxima are inclined upward.   This elevation angle decreases with increasing frequency.   This characteristic is overcome by the construction of Fig. 21.29c in which the ground plane becomes a skirt or cone.   If the apex angle of the skirt is adjusted properly, the maximum radiation will be normal to the vertical radiator. A further modification results in the "discone" antenna of Fig. 21.29e.   This antenna maintains reasonably uniform characteristics over a 3:1 frequency ratio and can be considered to change from a vertical radiator at the lowest frequency of operation to a conical horn at the highest frequency of operation.   The dimensions shown in the figure are for the lowest frequency of operation.

If the ground plane is folded back down along the coaxial line, the sleeve antennas of Fig. 21.29f and g result.   The maximum radiation is in a direction normal to the

[1] A. G. Kandoian, Three New Antenna Types and Their Applications, *Proc. IRE*, February, 1946, pp. 70w–75w.   A. S. Meier and W. P. Summers, Measured Impedance of Vertical Antennas over Finite Ground Planes, *Proc. IRE*, June, 1949, pp. 609–616.   Kraus, *op. cit.*, p. 420.

antenna axis. The thickened upper portion of Fig. 21.29g is one of several possible variations of the sleeve antenna giving broader bandwidth characteristics.

21.2h. Helical Antennas.[1] Helical antennas are shown in Fig. 21.30. Helical antennas radiate in one of two general modes; the axial, or "beam," mode or the normal mode. The normal mode exists when the helix circumference is much smaller than one wavelength. The maximum intensity is normal to the helix axis as shown in Fig. 21.30a, and as the helix diameter becomes smaller and smaller, the pattern approaches that of a straight wire. This mode is elliptically polarized, in general, but can be made circularly polarized.[2] In practice, this mode is limited by low radiation resistance (and, consequently, low radiation efficiency) and narrow bandwidth.

(a) NORMAL MODE

(b) AXIAL MODE-GROUND PLANE

(c) AXIAL MODE-NO GROUND PLANE

(d) AXIAL MODE-LINEAR POLARIZATION
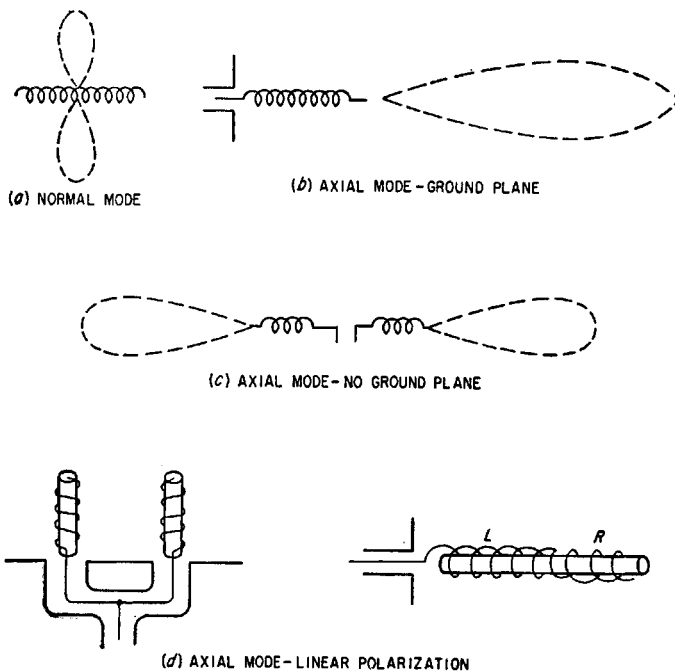
FIG. 21.30. Helix antennas.

The narrow bandwidth results from a requirement for current of uniform amplitude and phase which, in general, requires termination devices which cannot be broadbanded in a simple manner.

The axial mode most generally used in practice is one which occurs when the helix circumference is of the order of one wavelength. The maximum field intensity is along the helix axis, and the helix dimensions are relatively noncritical. The configuration shown in Fig. 21.30b is convenient in practice in that it provides a termination to a coaxial feed line and gives a unidirectional pattern. The ground plane should be greater than $\lambda/2$ in diameter. The configuration of Fig. 21.30c can be used when a bidirectional pattern is wanted. In general, helical antennas operating in the axial mode radiate with circular polarization, with the sense of the polarization

---

[1] This section is a brief adaptation from the excellent and complete discussion in Kraus, *op. cit.*, pp. 173–214. Additional references are cited in this work.

[2] H. A. Wheeler, A Helical Antenna for Circular Polarization, *Proc. IRE*, vol. 35, pp. 1484–1488, December, 1947.

depending on the winding sense of the helix. The configurations of Fig. 21.30d each combine two halves of different winding senses to produce linearly polarized radiation.

The pattern of a helical antenna using the axial radiation mode and designed for maximum directivity[1] can be determined from Eq. (21.21).

$$E(\phi) = \sin (90°/n) \left(\frac{\sin n\psi/2}{\sin \psi/2} \right) (\cos \phi) \qquad (21.21)$$

where n = number of turns of helix
  ϕ = angle measured from axis of helix

The first factor of Eq. (21.21) normalizes the pattern to have a unity maximum amplitude. The second factor is the "array factor" and represents the contribution to the total pattern of the turns of the helix considered as an array of separate sources. The third factor is the pattern of a single-turn helix.

The angle ψ is given by Eq. (21.22).

$$\psi = 360° \left[\frac{S(1 - \cos \phi)}{\lambda} + \frac{1}{2n} \right] \qquad (21.22)$$

where S = spacing between turns of helix
The relationship between S, the circumference of the helix C, and the pitch angle, α is given by Eq. (21.23).

$$\tan \alpha = \frac{S}{C} \qquad (21.23)$$

For helix antennas in the axial mode, the circumference of the helix should be between a lower limit of $3\lambda/4$ and an upper limit of $4\lambda/3$ and α should be between 12 and 15°.

The beam width of a helix antenna in the axial mode decreases inversely with the square root of the length of the helix in wavelengths. Equations (21.21) to (21.23) will be illustrated through the following example.

Example 21.3

Calculate the pattern of a seven-turn helical antenna with a spacing between turns of 0.25λ and a circumference of 0.95λ.

Solution

1. Determine ψ from Eq. (21.22).

   $\psi = 360°[0.25(1 - \cos \phi) + \frac{1}{14}]$

2. Determine $E(\phi)$ from Eq. (21.21).

$$E(\phi) = (\sin 12.8°) \left(\frac{\sin 7\psi/2}{\sin \psi/2} \right) \cos \phi$$

This equation must be evaluated for various values of ϕ including the variation of ψ with ϕ.
This is shown in Fig. 21.31.
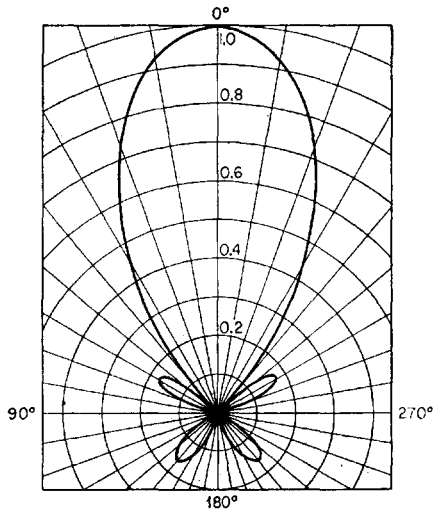
FIG. 21.31. Helical antenna pattern from Example 21.3.

21.3. Systems of Antennas—Antenna Arrays. Individual antennas are often combined in such a way as to produce a composite radiation pattern possessing desirable

[1] See W. W. Hansen and J. R. Woodyard, A New Principle in Directional Antenna Design, *Proc. IRE*, vol. 26, pp. 333–345, March, 1938.

properties not achievable with a single antenna element. Such an assembly of radiating elements is called an array, and the elements are usually all identical and equally spaced in a given direction from one another.

The simplest sort of array is one made up of a number of radiating elements arranged on a straight line, a so-called *linear* array. Two- and three-dimensional arrays are also possible and consist of linear arrays extending in all three coordinate directions as in Fig. 21.32. The properties of such an array can be determined by dividing it into three linear arrays extending in each of the three directions. The total array pattern is then given by the product of the patterns of the three arrays. In the succeeding discussion of arrays it is assumed that the spacings between radiators and the patterns of each radiator are such that mutual coupling effects are negligible.

FIG. 21.32. General array coordinates.

21.3a. Linear Arrays, Uniform Amplitude, and Phase Distribution. The normalized radiation pattern of a linear array of identical elements of equal amplitude and uniform phase distribution is given by Eq. (21.24) in conjunction with Eq. (21.25).

$$E(\phi) = \frac{E_o(\phi)}{n} \frac{\sin (n\psi/2)}{\sin (\psi/2)} \qquad (21.24)$$

and

$$\psi = \frac{2\pi D}{\lambda} \cos \phi + \delta \qquad (21.25)$$

where D = spacing between radiators
λ = wavelength in same units as D
δ = phase shift between successive radiators
$E_o(\phi)$ = pattern of individual radiators
ϕ = angle measured from axis of array
n = number of radiators

The arrangement as given in Eqs. (21.24) and (21.25) is convenient since the first expression depends only on the number of radiators and can be tabulated independently of the spacing or phasing of an array. Normalized array factor charts relating $E(\phi)/E_o(\phi)$ to ψ for various values of n are shown in Fig. 21.33.[1] Then, for a particular array it is only necessary to evaluate ψ for each value of ϕ, and the corresponding values of $E(\phi)$ can be read from the chart.

The array pattern given in Eq. (21.24) has its maximum value of unity whenever $\psi = 0$. It is possible, therefore, to choose the direction of maximum radiation for an array at any angle ψ_o by setting $\psi = 0$ in Eq. (21.25), substituting ϕ_o and choosing D and δ to maintain the equality, i.e.,

$$\delta = -\frac{2\pi D}{\lambda} \cos \phi_o \qquad (21.26)$$

---

[1] A more complete tabulation of Universal Pattern Charts is given in Kraus, *op. cit.*, Appendix.
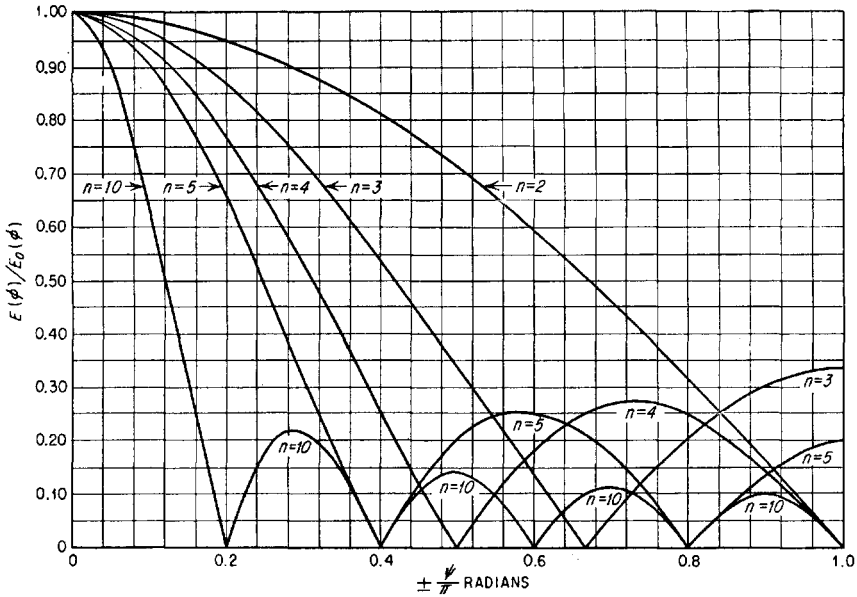
FIG. 21.33. Normalized array-factor chart for linear arrays.

Example 21.4

Determine the pattern of two identical isotropic radiators of equal amplitude and zero phase shift spaced $\lambda/4$ apart.

Solution

In this case

$$\delta = 0$$

$$\psi = \frac{\pi}{2} \cos \phi$$

$$E = \frac{\sin \left(\frac{\pi}{2} \cos \phi \right)}{2 \sin \left(\frac{\pi}{4} \cos \phi \right)} = \cos \left(\frac{\pi}{4} \cos \phi \right)$$

This pattern is plotted in Fig. 21.34a.
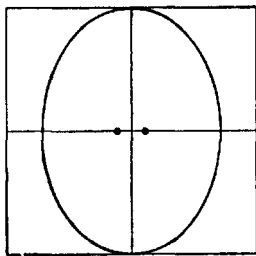
(*a*) PATTERN FOR EXAMPLE 21.4   (*b*) PATTERN FOR EXAMPLE 21.5*a*   (*c*) PATTERN FOR EXAMPLE 21.5*b*
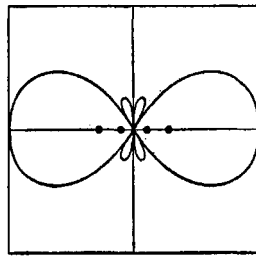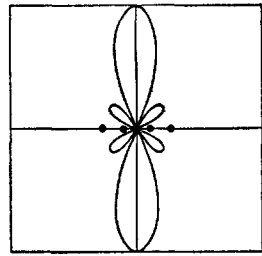
FIG. 21.34. Patterns of linear arrays.

Example 21.5

Determine δ so that maximum field occurs at (a) $\phi = 0°$ and (b) $\phi = 90°$ for four sources spaced $\lambda/2$ apart.

Solution

From Eq. (21.26) for $\phi = 0°$

$$\delta = -\pi \cos 0° = -\pi$$

then

$$\psi = \pi(\cos \phi - 1)$$

$$E = \frac{\sin 2\psi}{4 \sin \dfrac{\psi}{2}} = \frac{\sin [2\pi(\cos \phi - 1)]}{4 \sin \left[\dfrac{\pi}{2} (\cos \phi - 1)\right]}$$

This pattern is shown in Fig. 21.34b.
  For $\phi = 90°$

$$\delta = -\pi \cos 90° = 0$$
$$\psi = \pi \cos \phi$$
$$E = \frac{\sin (2\pi \cos \phi)}{4 \sin \left(\dfrac{\pi}{2} \cos \phi\right)}$$

This pattern is plotted in Fig. 21.34c.

Broadside Arrays. A broadside array is one with zero phase shift between elements; that is, $\delta = 0$. The direction of maximum radiation is normal to the array.
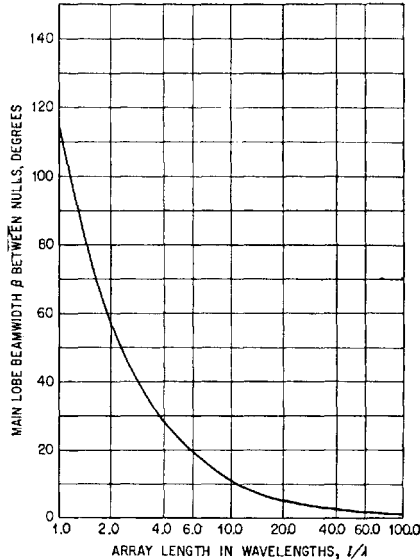
FIG. 21.35. Beamwidth between nulls of broadside arrays as a function of the length of the array in wavelengths.

An example is shown in Fig. 21.34c. The three-dimensional pattern is a figure of revolution about the axis of the array. The gain of a broadside array is approximately proportional to the length of the array in wavelengths. Figure 21.35 gives the main lobe beam width between nulls as a function of the array length in wavelengths. The half-power beam width is approximately one-half the first null beam width. The angular positions of nulls and minor-lobe maxima are given in Fig. 21.36 as a function of the array length in wavelengths.

Ordinary Endfire Arrays. This array is one in which there is a progressive phase shift between elements just equal to the electrical spacing. The radiation maximum is along the line of the array and for spacings less than $\lambda/2$ is unidirectional in the
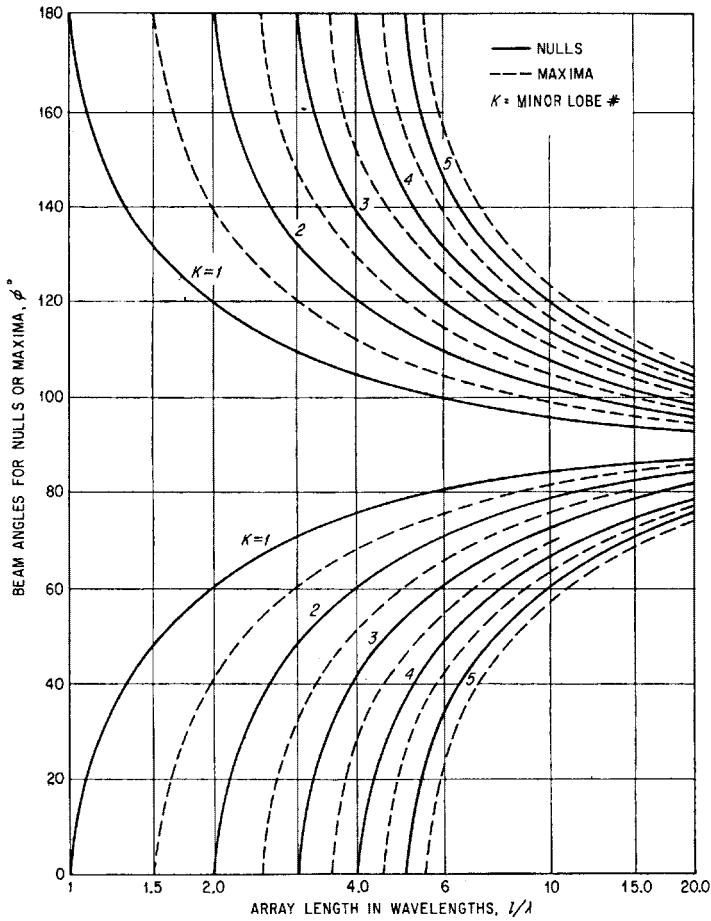
FIG. 21.36. Angular positions of nulls and minor-lobe maxima as a function of array length in wavelengths for broadside arrays.

direction of lagging phase shift. For an endfire array, the expression for ψ becomes

$$\psi = \frac{2\pi D}{\lambda} (\cos \phi - 1) \qquad (21.27)$$

The beam width between nulls of an endfire array is given in Fig. 21.37 as a function of the array length in wavelengths. It can be seen that the beamwidth is greater than for a broadside array of the same length. Because of the omnidirectional properties of the broadside pattern, the power gains of endfire and broadside antennas of the same length are approximately equal. Figure 21.38 gives the angular positions of nulls and minor-lobe maxima as a function of the array length in wavelengths.

Increased-directivity Endfire Arrays.[1] Increased directivity can be obtained in an endfire array by increasing the phase shift between elements to a value equal to the electrical spacing between elements plus π/n where n is the number of elements.

If the phase difference is increased without adjusting the spacing, the increased directivity may not be fully realized because of an increase in the width of the back-lobe pattern. This effect can be eliminated if the spacing in wavelengths is made less than or equal to $(n-1)/4n$. For the increased-directivity endfire array, the value of ψ is given by Eq. (21.28).

$$\psi = \frac{2\pi D}{\lambda}(\cos\phi - 1) - \frac{\pi}{n} \quad (21.28)$$

The gain of an increased-directivity endfire array can be improved by almost 2:1 over an ordinary endfire array. Figure 21.39 gives the beam width between nulls of increased-directivity endfire arrays as a function of array length in wavelengths. Figure 21.40 gives the angular positions of nulls and maxima as a function of the array length in wavelengths.

Minor-lobe Amplitudes. The first three minor lobes of uniformly illuminated broadside and endfire arrays are less than the main-lobe field intensity by approximately 13, 18, and 20 db, respectively. Somewhat greater side-lobe levels will be obtained with increased-directivity endfire arrays.

Fig. 21.37. Beamwidth between nulls as a function of the length of the array in wavelengths for endfire arrays.

21.3b. Linear Arrays—Nonuniform Amplitude Distribution. Linear arrays can be designed with a nonuniform amplitude distribution, a nonuniform phase shift between elements, or both, to obtain improved pattern characteristics over uniform linear arrays. An important special case is a broadside array with nonuniform amplitude distribution. Changes which can be effected by varying the amplitude of each radiating element are an increase or decrease in the gain and main-lobe directivity, an increase or decrease in side-lobe level, and equalization of the level of all side lobes. Any particular design involves a compromise between these factors. In general, a narrow main-lobe beam width will be achieved at the expense of increased side-lobe amplitudes. An amplitude distribution that has a maximum in the center of the array and is reduced (tapered) in amplitude at each end of the array will, in general, produce a slightly wider main lobe with reduced side-lobe amplitudes. Conversely, an amplitude distribution with a minimum at the center and increased amplitudes at the ends of the array will produce a narrow main-lobe beam width and increased side-lobe amplitudes. The amplitude distributions shown in Fig. 21.41 have the following characteristics.

Gabled Distribution. This is shown in Fig. 21.41a. Gabled distribution has the advantages of providing a relatively simple calculation of the individual element amplitudes and a means of controlling beam width and side-lobe level by adjusting

[1] Hansen and Woodyard, *loc. cit.*

the slope of the linear taper. A reasonable compromise between beam width and side-lobe level is obtained by tapering the amplitudes so that the end elements are 10 db below the peak amplitude.

Binomial or Gaussian Distribution. The binomial distribution is shown in Fig. 21.41*b*. The element amplitudes are proportional to the coefficients of a binomial
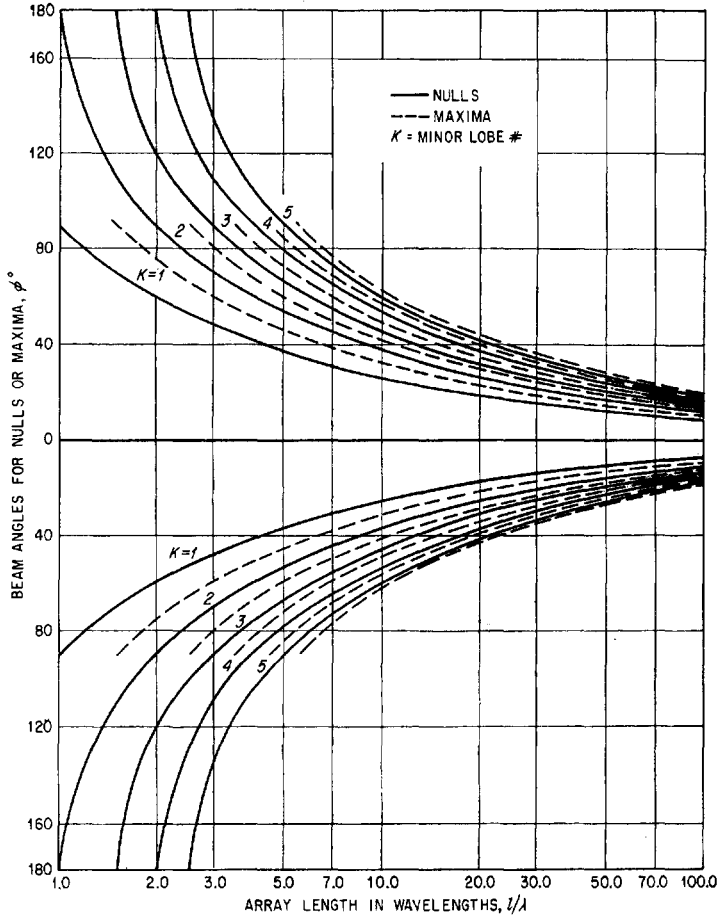
FIG. 21.38. Angular positions of nulls and maxima as a function of array length in wavelengths for endfire arrays.

series in which the number of terms is equal to the number of elements. The binomial coefficients are given below in the form of Pascal's triangle in which each term is the sum of the two adjacent terms in the row above it. Thus for a five-source array the element amplitudes will vary as 1, 4, 6, 4, 1 from one end to the other. This distribution produces no side lobes for element spacings less than $\lambda/2$. The penalty is an increase in the beam width by approximately 30 per cent and a corresponding decrease in gain over the uniform-amplitude case for a long array.

```
                              1
                          1       1
                      1       2       1
                  1       3       3       1
              1       4       6       4       1
          1       5      10      10       5       1
      1       6      15      20      15       6       1
  1       7      21      35      35      21       7       1
1     8      28      56      70      56      28       8       1
1    9     36      84     126     126      84      36      9      1
```
Pascal's Triangle

Cosine Distribution. The cosine distribution of Fig. 21.41*c* is similar in effect and application to the tapered illumination. The amplitudes of the end elements can be established at any two symmetrical points and do not necessarily have to be at the ends of the cosine curve. This distribution will produce slightly wider beam widths and lower side lobes than the linearly tapered distribution for the same peak amplitude.

Dolph-Tschebyscheff Distribution.[1] This distribution provides an optimum compromise between beam width and side-lobe level. It will provide the lowest side lobes for a given beam width or the narrowest beam width for a given side-lobe level. All side lobes will have uniform amplitudes. The binomial distribution is actually a limiting case of a Dolph-Tschebyscheff distribution where the ratio between main-lobe and side-lobe levels becomes infinite. The other limiting case is one in which only the end elements of the array are excited and all lobes have the same amplitude. Uniform amplitude distributions are not a special case in this group.

FIG. 21.39. Beamwidth between nulls as a function of array length in wavelengths for increased-directivity endfire arrays.

A practical difficulty associated with the use of Dolph-Tschebyscheff distributions is the difficulty of adjustment of the excitation of the individual radiators when a large number of radiating elements are required.

21.3c. Parasitic Antennas.[2] Parasitic antennas are shown in Fig. 21.42. A parasitic antenna is coupled to a driven antenna by the mutual impedance between the two antennas. A voltage induced in the parasitic element by the current in the

---

[1] C. L. Dolph, A Current Distribution for Broadside Arrays Which Optimizes the Relationship between Beamwidth and Side Lobe Level, *Proc. IRE*, vol. 34, pp. 335–384, June, 1946. H. J. Riblett, Discussion on Dolph's Paper, *Proc. IRE*, vol. 35, pp. 489–492, May, 1947. Domenick Barbiere, A Method for Calculating the Current Distribution of Tschebyscheff Arrays, *Proc. IRE*, vol. 40, pp. 78–82, January, 1952. Kraus, *op. cit.*, pp. 93–110.

[2] G. H. Brown, Directional Antennas, *Proc. IRE*, vol. 25, pp. 78–145, January, 1937. P. S. Carter, "Circuit Relations in Radiating Systems, *Proc. IRE*, vol. 20, p. 1004, June, 1932.

driven antenna will cause a current to flow in the parasite.   This current will produce a radiation field which will combine with the radiation field of the driven antenna. The resultant pattern of such an array will usually have increased directivity over the pattern of the driven antenna, depending upon the spacing between the driven and parasitic elements and the dimensions of the parasite.   Parasitic arrays are
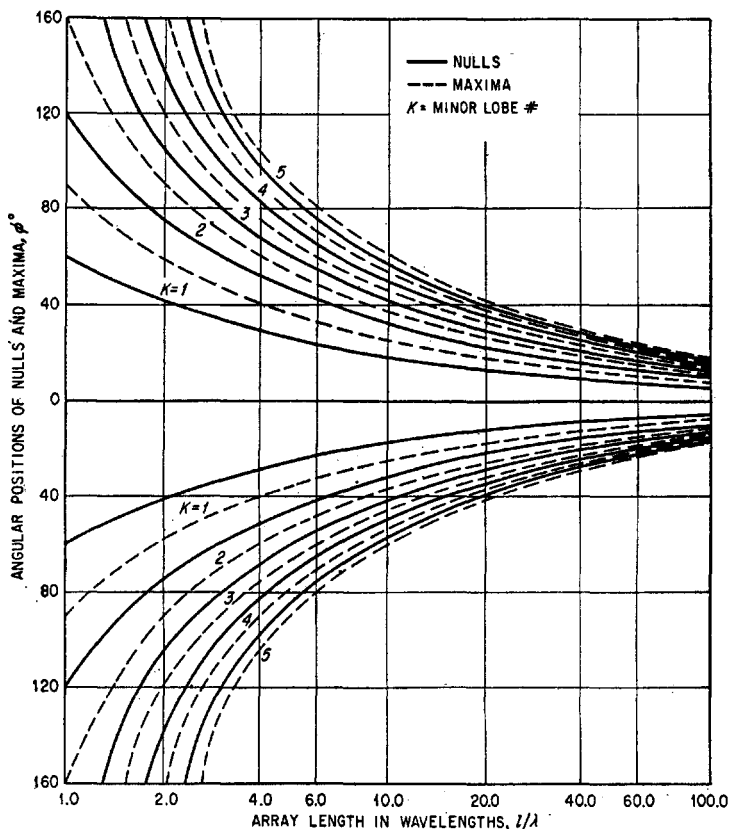
Fig. 21.40. Angular positions of nulls and maxima as a function of array length in wavelengths for increased directivity arrays.
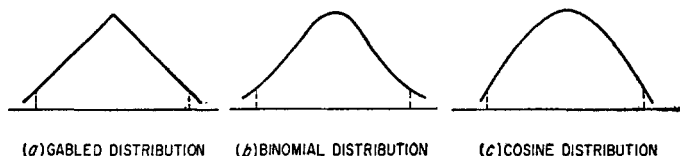
(a) GABLED DISTRIBUTION   (b) BINOMIAL DISTRIBUTION   (c) COSINE DISTRIBUTION
Fig. 21.41. Linear array amplitude distributions.

usually designed with close spacing between elements.   The directive properties of a particular parasite and driven antenna combination are determined by the relative phase and magnitude of the driven and induced currents.   Specifically, for close spacings, if the phase of the current in the parasite lags the phase of the current in the driven antenna, the direction of maximum radiation will be in the direction from the driven element toward the parasite and the parasite is called a director.   If the

phase of the current in the parasite leads the phase of the current in the driven antenna, the direction of maximum radiation will be in the direction from the parasitic element toward the driven element and the parasite is called a reflector. Combinations of reflectors and directors to further increase the directivity are also possible. These arrangements of elements are shown in Fig. 21.42. The Yagi[1] antenna shown in Fig. 21.42b has the disadvantage of narrow effective bandwidth, but it provides a simple method of achieving considerable improvement in directivity. Only one reflecting element is necessary because the residual field behind the reflector is usually quite small.

(a) DRIVEN ANTENNA AND PARASITE

(b) REFLECTOR AND DIRECTOR ANTENNAS
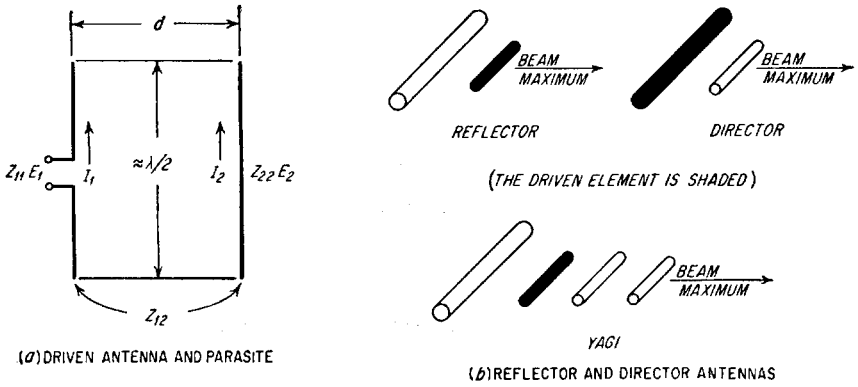
FIG. 12.42. Parasitic antennas.

The equations describing the relationships between one driven and one parasitic element are as follows:

$$V_1 = I_1 Z_{11} + I_2 Z_{12} \qquad (21.29)$$
$$0 = I_1 Z_{12} + I_2 Z_{22}$$

where V_1 = voltage at center of driven antenna

I_1, I_2 = currents in driven and parasitic antennas, respectively

Z_{11}, Z_{22} = self-impedances of driven and parasitic antennas, respectively

and $Z_{12} = Z_{21}$ is the mutual impedance between the two antennas. From Eq. (21.29), the current I_2 is related to I_1 and the antenna impedances by

$$I_2 = -I_1 \frac{Z_{12}}{Z_{22}} \qquad (21.30)$$

or

$$I_2 = I_1 \left| \frac{Z_{12}}{Z_{22}} \right| \underline{/\pi + \theta_{12} - \theta_{22}} \qquad (21.31)$$

where θ_{12} = phase angle of Z_{12}

θ_{22} = phase angle of Z_{22}

The array expression of one driven and one parasitic antenna is given by

$$E(\phi) = kI_1 \left[1 + \left| \frac{Z_{12}}{Z_{22}} \right| \underline{\Big/ \pi + \theta_{12} - \theta_{22} - \frac{2\pi d}{\lambda} \cos\phi} \right] \qquad (21.32)$$

where d = spacing between radiators

ϕ = angle in plane of radiators measured from line connecting centers of radiators

k = an amplitude constant

[1] H. Yagi, Beam Transmission of Ultra Short Waves, *Proc. IRE*, vol. 16, pp. 715–740, June, 1928.

The total pattern will be the pattern as determined from Eq. (21.32) multiplied by the pattern of the $\lambda/2$ driven antenna.

It can be seen from this equation that the design of parasitic antenna systems reduces to the determination of the self and mutual impedances of the radiating elements. Figure 21.43 gives the magnitude and phase angle of the mutual impedance[1] between parallel parasitic and driven antennas. The self-impedance of an antenna approximately $\lambda/2$ in over-all length can be approximated by Eq. (21.33).

$$Z_{22} = 73 + j73 \tan \theta_{22} \quad (21.33)$$

FIG. 21.43. Mutual impedance between driven and parasitic antennas.

where θ_{22} = the phase angle of the impedance Z_{22}

Equation (21.33) is an approximation and is only valid for infinitely thin rods. The magnitude of the reactive portion of Z_{22} varies with length, the rate of variation decreasing as the diameter-to-length ratio of the antenna is increased. In general, for lengths less than $\lambda/2$, the reactance is negative; for lengths greater than $\lambda/2$, the reactance is positive. The self-impedance of antennas has been investigated by many workers.[2] In a practical application the secondary effects of mutual coupling are difficult to evaluate, especially when more than two elements are used in the array. Consequently the final adjustment of antenna lengths is often made experimentally.

Representative values of self-reactance for antennas near one-half wavelength (i.e., $73 \tan \theta_{22}$ ohms) are tabulated below:

TABLE 21.1. SELF-REACTANCE OF CYLINDRICAL ANTENNA, IN OHMS

	$\lambda/2 + 5\%$	$\lambda/2 - 5\%$	$\lambda/2 + 10\%$	$\lambda/2 - 10\%$
$l/r = 400$	+40	+3	+60	−15
$l/r = 2,000$	+50	−3	+75	−30
$l/r = 20,000$	+62	−12	+95	−45

l/r = length-over-radius ratio.

Representative array factors[3] of one driven and one parasitic antenna are shown in Fig. 21.44. An examination of these array factors will reveal that a parasitic element will act as a director, when $(\pi + \theta_{12} - \theta_{22})$ is between 0 and $-180°$. A parasite is a reflector when this difference is between 0 and $+180°$. A director is usually spaced approximately 0.2λ from the driven element.

[1] Figure 21.43 represents a simplification of the problem adequate for preliminary design calculations. For a detailed discussion of the subject, the reader is referred to the source references Brown, op. cit., and Carter, op. cit. In addition, very readable discussions are given in: Kraus, op. cit., chap. 10. E. C. Jordan, "Electromagnetic Waves and Radiation System," chap. 13, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1950. F. E. Terman, "Radio Engineers' Handbook," sec. 11, McGraw-Hill Book Company, Inc., New York, 1943. Additional literature references tabulating the results of considerable experimental data are: Ronald King, Self and Mutual Impedances of Parallel Identical Antennas, Proc. IRE, vol. 40, pp. 981–988, August, 1952. C. T. Tai, Coupled Antennas, Proc. IRE, vol. 36, pp. 487–500, April, 1948.

[2] See the footnote concerning mutual impedance, p. 26–16.

[3] After Brown, op. cit.
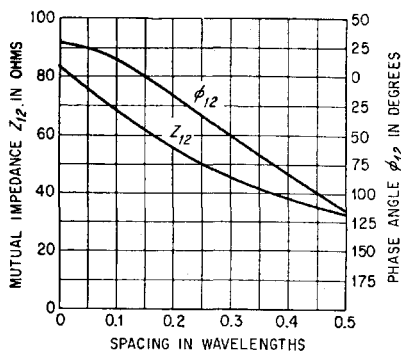
The impedance of the driven antenna is lowered by parasitic elements.   A folded dipole[1] can be used as the driven element to increase the input resistance of the antenna.

The gain of a parasitic array depends upon the number of elements in the array and their spacing.   With one reflector or director spaced so as to obtain maximum gain, values of approximately 5 db over a λ/2 antenna will be obtained.   A three-element array with one director and one reflector properly spaced will yield a maximum gain of over 7 db compared to a λ/2 antenna.   A four-element array consisting
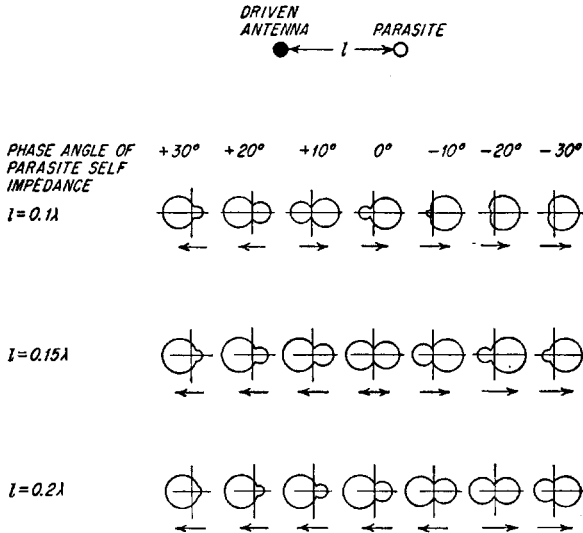
FIG. 21.44. Representative array factors of a driven antenna with one parasite.

of a reflector and two directors properly spaced will yield a maximum gain of over 9 db compared to a λ/2 antenna.

Example 21.6

Determine the array factor of an antenna system composed of a driven λ/2 antenna and a parasitic director spaced 0.1λ from the driven antenna.   The length of the parasitic antenna is shorter than the resonant length so that the phase angle of the self-impedance of the director is −20°.

Solution

1. Determine the ratio of the mutual and self-impedances.
From Fig. 21.43

$$Z_{12} = 69\underline{/+15°}$$

From Eq. (21.33)

$$Z_{22} = 73 + j73 \tan (-20°)$$
$$= 73 - j73 \times 0.364$$
$$= 73 - j26.6 = 77.7\underline{/-20°}$$

Therefore,

$$\frac{Z_{12}}{Z_{22}} = \left| \frac{69}{77.7} \right| \underline{/180 + 15° + 20°}$$
$$= 0.89\underline{/215°}$$

2. Determine the array factor from Eq. (21.32) assuming unity current in the driven antenna.

$$E(\phi) = k[1 + 0.89\underline{/215°} - 36° \cos \phi]$$

---

[1] See Sec. 21.2g.

The array factor is shown in Fig. 21-45. It must be multiplied by the pattern of the $\lambda/2$ driven element to obtain the over-all pattern.

21.3d. Omnidirectional Arrays. For many applications, it is important to provide an antenna pattern that is uniform for all horizontal angles. The over-all antenna gain under these conditions will usually have an undesirably low value unless radiators are stacked vertically to provide directivity in the vertical plane. Many of the basic radiators discussed in Sec. 21.2, such as loops, dipoles, or sleeve antennas possessing axial symmetry, can be arranged in a vertical array for this purpose. Several convenient configurations[1] not previously discussed are shown in Fig. 21.46. The turnstile[2] antenna shown in Fig. 21.46a consists of two half-wavelength dipoles crossed at 90° and excited in phase quadrature. For example, terminals 1-1 would lead or lag terminals 2-2 in phase by 90°. Assuming equal amplitudes of excitation, the
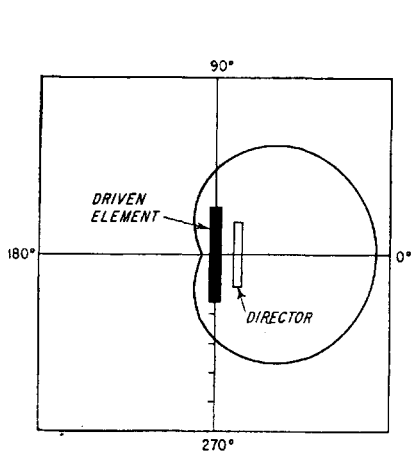
FIG. 21.45. Array factor for Example 21.6.
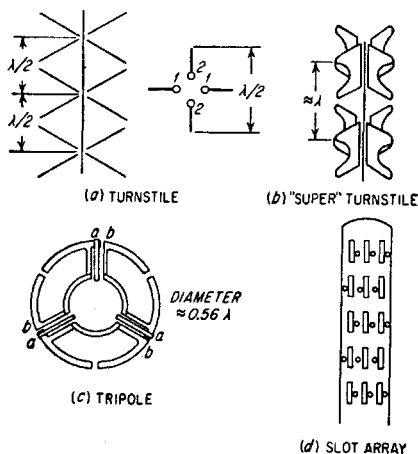
FIG. 21.46. Omnidirectional antennas.

horizontal pattern for this arrangement is approximately circular. Turnstiles are usually stacked with $\lambda/2$ spacing and fed inphase. The superturnstile is used to increase the bandwidth over which it is possible to maintain a low standing wave ratio. The "dipoles" consist of vertical fins or wire structures. The fin length is usually less than $3/4\lambda$, and the vertical separation is approximately one wavelength. The "tripole" radiator shown in Fig. 21.46c is a means of obtaining axial symmetry at microwave frequencies. The phase of the excitation reverses each $\lambda/2$ in the feed line. In order to have the tripoles radiate in phase, the phase of successive tripoles in the vertical array must be reversed by reversing the side connected to the coupling probe, e.g., sides b could be connected to the probes on alternate tripoles. Thus with a vertical spacing of $\lambda/2$ and alternate tripoles fed in opposite phase, it is possible to achieve a vertical broadside array. A slot version of a microwave omnidirectional radiator is shown in Fig. 21.46d. The phase is reversed in adjacent vertical rings

[1] For a discussion of microwave omnidirectional arrays, see S. Silver (ed.), "Microwave Antenna Theory and Design," chap. 9, McGraw-Hill Book Company, Inc., New York, 1949.

[2] For a discussion of omnidirectional systems applicable to lower frequencies, see: Terman, *op. cit.*, pp. 815–818, or E. A. Laport, "Radio Antenna Engineering," McGraw-Hill Book Company, Inc., New York, 1952. George H. Brown, The Turnstile Antenna, *Electronics*, vol. 9, p. 15, April, 1936. R. W. Masters, The Super Turnstile Antenna, *Broadcast News*, January, 1946, p. 42.

by reversing the position of the coupling pins. The polarization of the tripole and slot arrays is horizontal.

21.4. Microwave Radiators. Microwave antennas form an important segment of the antenna field because of their wide range of application and the diversification of techniques which can be employed. The following paragraphs describe the most generally useful microwave radiating elements. Where high gain and narrow or specialized beam shapes are not a requirement, these basic radiators will often provide the necessary characteristics for the complete antenna system. If high gain and narrow or specialized beam shapes are required, arrays of microwave radiators as discussed in Sec. 21.3, or systems using reflectors or lenses as discussed in Sec. 21.5, can be used.

21.4a. Microwave Dipoles. Several forms of microwave dipoles are shown in Fig. 21.47. The balanced-line dipole of Fig. 21.47a is used with a two-wire balanced transmission line and, consequently, is not useful where waveguide or coaxial transmission lines are employed. It has the advantage, however, of equal excitation of each wing of the dipole resulting in a symmetrical pattern. If the dipole length is small with respect to a wavelength, the pattern is given by Eq. (21.14) in Sec. 21.2. If the dipole is not "small," Eq. (21.17) must be used.[1] The pattern in a plane normal to the dipole axis is omnidirectional.

The coaxial sleeve dipole shown in Fig. 21.47b is a coaxial version of the center-fed dipole. The upper half is an extension of the inner conductor and is usually $\lambda/4$ long. This stub is sometimes increased in thickness as shown by the dotted lines to improve the bandwidth. The lower portion is a sleeve formed by folding back the outer conductor of the coaxial line. It is also usually equal to or slightly less than $\lambda/4$. The radiation pattern is affected by the stub length, sleeve length, and the gap between the sleeve and the coaxial line. The symmetry of the pattern is improved by adding a second $\lambda/4$ choke separated from the sleeve by $\lambda/8$ as shown by the dotted lines. The direction of maximum radiation is normal to the axis of the coaxial line.

(a)BALANCED LINE DIPOLE

(b)COAXIAL SLEEVE DIPOLE

(c)OPEN END DIPOLE

(d)STUB SUPPORTED DIPOLE

(e)SLOT FED DIPOLE

(f)STUB SUPPORTED SLOT FED DIPOLE

(g)DOUBLE DIPOLES

(h)DIPOLE DISKS

FIG. 21.47. Microwave dipoles.

A coaxial dipole with the direction of maximum radiation along the axis of the coaxial line is shown in Fig. 21.47c. The currents in each wing of this *open-end dipole* are unbalanced, the greater current being in the wing attached to the inner conductor. This produces a "squint" in the pattern, or displacement of the maximum intensity several degrees toward the side of the wing with the greater current. The over-all pattern distortion caused by leakage currents down the outside of the coaxial line can be reduced by placing a $\lambda/4$ choke around the outer conductor.

The *stub-supported dipole* shown in Fig. 21.47d gives improved mechanical support over the open-end dipole. The stub is approximately $\lambda/4$ long and is short-circuited

---

[1] For example, the half-power beam width of a $\lambda/2$ dipole is 78° compared to 90° for a very short dipole.

at the end. This further decreases the excitation of the wing attached to the outer conductor, however, and consequently increases the squint of the pattern.

The *slot-fed dipole* of Fig. 21.47e achieves a more equal excitation of the dipole wings. Both wings are fastened to the outer conductor; however, the inner and outer conductors are short-circuited on the side of one of the wings. The slot will also decouple the dipole from the outer conductor of the line, eliminating the need for a choke. The slot length is approximately $\lambda/4$ but is adjusted to compensate for mismatches with the particular dipole and line dimensions used.

The *stub-supported slot-fed dipole* shown in Fig. 21.47f can also be used with the same mechanical advantage as before. The total slot length is approximately $\lambda/2$. In general, the addition of a stub support will lower the power-handling capacity of a dipole radiator.

The *double dipole* shown in Fig. 21.47g can be regarded as a primary and parasitic radiation[1] system. It is used to obtain greater gain and directivity over a single dipole. A spacing of $\lambda/8$ between dipoles and a corresponding parasitic length slightly greater than the primary dipole length are customarily used. The combination has an effective center of radiation between the primary and parasitic dipole. In the waveguide case, the pattern will be affected by the waveguide taper which is used to improve the bandwidth and by the dimension and insertion depth of the plate supporting the dipoles.

The *dipole-disk antenna* shown in Fig. 21.47h utilizes a metallic disk spaced $\lambda/4$ behind the dipole to provide an image antenna, increasing the directivity as in the double-dipole systems. The theoretical pattern of such a combination can be determined by the method of images;[2] because of losses, however, the effective center of radiation of such a system is between the dipole and the disk.

When dipole antenna systems are used to feed a reflector system, the dimensions of the feed must be minimized to avoid distortion of the secondary pattern. Accordingly, dipole-disk and other feed systems are often made with less than the normal $\lambda/2$ aperture at some sacrifice in gain and pattern characteristics.

Effect of Dipole Shape. An improvement in bandwidth can be achieved in dipole systems by decreasing the length over diameter ratio l/D for the dipole. In practice, stubby cylindrical conductors rounded at the ends are used for the dipole wings instead of thin cylindrical conductors. In general the variation in input impedance with frequency decreases as the ratio of dipole length to diameter is decreased.

21.4b. Slot Radiators.[3] A slot cut in a waveguide or coaxial line can be either a radiating or a nonradiating element, depending upon its orientation with respect to the fields or currents within the transmission line. A slot cut parallel to lines of current flow will be nonradiating and can be used as a device to make measurements within the transmission line. A slot cut so as to intersect lines of current flow will couple power from the transmission line and will radiate the coupled power. The amount of coupling depends upon the length of the slot, the sine of the angle between the slot axis and the lines of current flow, and the current density intersected by the slot. The polarization of the radiation from a slot is such that the electric field is normal to the long dimension of the slot.

The width of slot radiators is small compared to the length. In cases where the slot length can be made $\lambda/2$, the radiator has characteristics similar to a dipole. Figure 21.48 shows representative slots in rectangular waveguides and coaxial lines. The dotted lines illustrate approximate directions of current flow for the TE_{10} mode

[1] See Sec. 21.3c.
[2] See Sec. 21.1g.
[3] For examples of the design of slot antennas see Radio Research Laboratory Staff, "Very High Frequency Techniques," vol. I, chap. 7, McGraw-Hill Book Company, Inc., New York, 1947.

in rectangular waveguide and the coaxial TEM mode. In this case, slots b, c, and i will be nonradiating; slots a, d, f, e, and j will radiate; and slots g, h, and k will vary in their radiating properties as a function of the tilt angle.

Dumbbell-shaped slots are used to decrease the physical length required for resonance. In cases where the slot must be filled with dielectric material for pressurization, the resonant length will also be reduced.
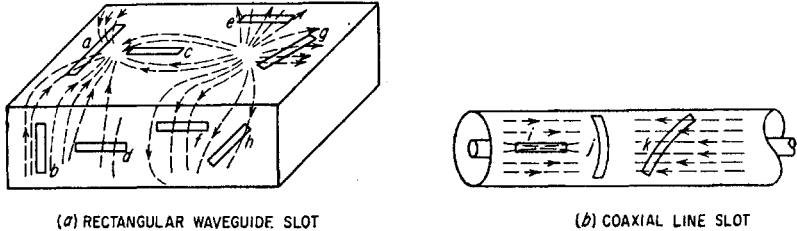
(*a*) RECTANGULAR WAVEGUIDE SLOT   (*b*) COAXIAL LINE SLOT

FIG. 21.48. Transmission-line slots.

A nonradiating slot can be coupled into the transmission line by means of a probe placed adjacent to the slot. The probe distorts the lines of current flow so that the slot is excited. The amount of coupling will be a function of probe location and depth. Bent probes are used to excite transverse slots in the narrow wall of rectangular waveguide. The phase of excitation can be reversed by reversing the side of the slot adjacent to the probe. By this means, linear broadside arrays are possible with $\lambda/2$ slot radiators separated one-half a guide wavelength[1] along the array. This phase reversal can also be achieved for slots in the wide side of a rectangular waveguide by staggering the slots on alternate sides of the guide axis. This reversal of phase for each adjacent slot is necessary because, although the slots must radiate in phase for a broadside array, the current direction is reversed in the waveguide each $\lambda/2$ along the guide. Hence, the alternation of the probe location provides for the

(*a*) WIDE-SIDE PROBE-FED SLOTS   (*b*) NARROW-SIDE PROBE-FED SLOTS

(*c*) REVERSED PARALLEL SLOTS — WIDE-SIDE   (*d*) REVERSED INCLINED SLOTS — NARROW-SIDE

FIG. 21.49. Phase reversal techniques for slot radiators.

proper excitation of all slots. The phase of excitation of inclined slots in the narrow side of a rectangular waveguide can be reversed by reversing the tilt angle. These techniques are shown in Fig. 21.49.

21.4c. Horn Radiators. A horn radiator consists of an open section of waveguide suitably shaped to provide the desired beam width or illumination pattern. Horns are used instead of dipoles because of the convenience of coupling when waveguide transmission lines are involved, because of their greater power-handling capability, and because of the flexibility of control they afford over the radiation patterns in the electric and magnetic planes. Typical horn configurations are shown in Fig. 21.50. Sectoral horns are flared in only one dimension. The E-plane sectoral horn has the flare in the plane of the electric field for the dominant TE_{10} mode. The H-plane sectoral horn has the flare in the plane normal to the E-plane. The radiation pattern from a horn is determined by the amplitude and phase distribution over the aperture of the horn. The amplitude of the electric field is uniform in the direction of the electric field, and consequently, E-plane horns have a relatively constant aperture illumination. The amplitude of the electric field in the H-plane tapers sinusoidally toward zero at the

---

[1] Guide wavelength is longer than free-space wavelength. See Sec. 20.5.

edges of the horn, and consequently, *H*-plane horns have a cosine illumination over their aperture. In either case there is a nonuniform phase distribution over the aperture because the path length along the flared edge is longer than the path length to the aperture along the axis of the horn. Because of these effects, the beam width of the *E*-plane horn is slightly narrower than an *H*-plane horn with the equivalent aperture. The *E*-plane horn has a higher side lobe level and is somewhat more sensitive to frequency changes. The design of a horn radiator involves a compromise between the desirability of increasing the length for a given aperture to minimize the phase error and the physical convenience of a small horn. An *optimum horn* attempts this compromise by allowing a small phase error in order to produce the smallest beam width and side lobe level for a given length of horn. Approximate

(a) E-PLANE SECTORAL   (b) H-PLANE SECTORAL   (c) PYRAMIDAL   (d) CONICAL   (e) TEM BICONICAL COAXIAL-FED
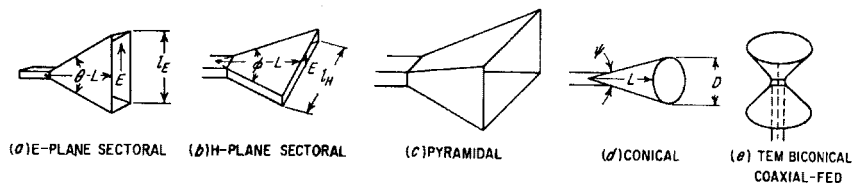
FIG. 21.50. Horn radiators.

expressions for the beam widths of horns with apertures at least several wavelengths long are as follows.[1]

1. Half-power beam width in degrees
   a. Optimum *E*-plane sectoral horn

$$\beta_E = \frac{56\lambda}{l_E} \tag{21.34}$$

   b. Optimum *H*-plane sectoral horn

$$\beta_H = \frac{67\lambda}{l_H} \tag{21.35}$$

2. Beam width between nulls in degrees
   a. Optimum *E*-plane sectoral horn

$$\beta_E = \frac{115\lambda}{l_E} \tag{21.36}$$

   b. Optimum *H*-plane sectoral horn

$$\beta_H = \frac{172\lambda}{l_E} \tag{21.37}$$

where λ = free-space wavelength
   l_E, l_H = lengths of *E*-plane and *H*-plane apertures, respectively, in units consistent with λ

Once the aperture length has been determined for a given beam width, the horn length L and flare angle θ or ϕ can be obtained from the following.[2]

$$L/\lambda = \frac{\left(\dfrac{l}{2\lambda}\right)^2 - \delta^2}{2\delta} \tag{21.38}$$

[1] Radio Research Laboratory Staff, "Very High Frequency Techniques," vol. I, chap. 6. Kraus, *op. cit.*, pp. 371–381.
[2] See Kraus, *loc. cit.* Also: W. L. Barrow and L. J. Chu, Theory of the Electromagnetic Horn, *Proc. IRE*, vol. 27, pp. 51–54, 1939. S. A. Shelkunoff, "Electromagnetic Waves," pp. 359–366, D. Van Nostrand Company Inc., Princeton, N.J., 1943. W. L. Barrow and F. M. Greene, Rectangular Hollow Pipe Radiators, *Proc. IRE*, vol. 26, p. 1498, 1938. For a tabulation of design information based upon a summary of experimental data see G. C. Southworth, "Principles and Applications of Waveguide Transmission," pp. 402–424, D. Van Nostrand Company, Inc., Princeton, N.J., 1950.

where l is l_E or l_H as applicable.

$$\theta \text{ or } \phi = 2 \cos^{-1}\left[\frac{L/\lambda}{L/\lambda + \delta}\right] \tag{21.39}$$

where δ = maximum tolerable path length difference in wavelengths; $\delta = 0.25$ for optimum E-plane sectoral horns and 0.40 for optimum H-plane sectoral horns

θ, ϕ = apex angles for E-plane and H-plane sectoral horns, respectively

These relationships are summarized in Figs. 21.51 and 21.52.
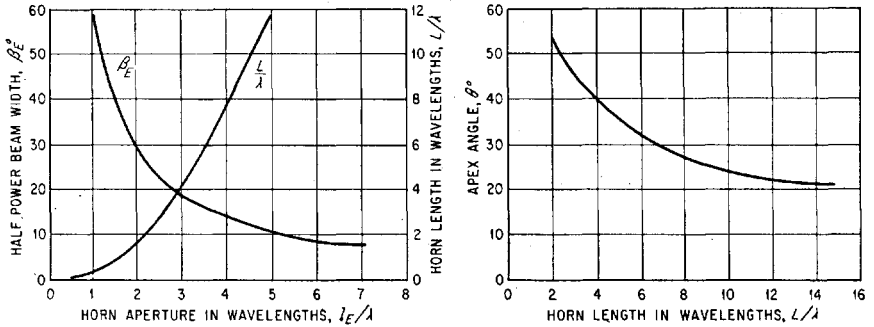
FIG. 21.51. Design curves for optimum E-plane horns.

The pyramidal horn of Fig. 21.50c is the result of combining the designs of E- and H-plane sectoral horns in a single aperture. The E- and H-plane patterns are essentially the same as the patterns of sectoral horns with the same E- or H-plane aperture. Kraus[1] gives the following expression for the gain of a horn:

$$G = \frac{7.5 l_E l_H}{\lambda^2} \tag{21.40}$$

The conical horn of Fig. 21.50d is an adaptation for a circular waveguide. The pattern for the dominant TE_{11} circular-waveguide mode is wider than an E-plane horn with an aperture equal to the diameter of the conical horn, but is narrower than the corresponding H-plane horn. The approximate relationships between apex angle, length, and aperture for an optimum conical horn is given by Eqs. (21.38) and (21.39) where $\delta = 0.32$.

The biconical horn shown in Fig. 21.50e can be considered as a conical horn with an apex angle of 360° and has an omnidirectional pattern normal to the axis of the cones.

Example 21.7

Determine the dimensions of two sectoral horn radiators one of which will have an E-plane beam width of 15° and the other an H-plane beam width of 15° when fed from a rectangular waveguide.

Solution

1. From Fig. 21.51, a horn aperture of 3.7λ is required to provide a half-power beam width of 15° in the E plane. From the same figure, a horn length of 6.8λ and a corresponding apex angle of 30° are required.

2. From Fig. 21.52, a horn aperture of 4.5λ is required to provide a half-power beam width of 15° in the H plane. From the same figure, a horn length of 6λ and a corresponding apex angle of 41° are required.

[1] *Op. cit.*, p. 380.

Note that it is not possible to combine these two optimum sectoral horns into a pyramidal horn. Optimum pyramidal horns are possible, however, and are discussed in Southworth.[1]

21.4d. Dielectric Rod Antenna.[2] A dielectric rod will act as a waveguide, containing most of the signal energy within the dielectric material if the rod diameter or cross section is sufficiently large. As the rod is tapered in thickness, appreciable energy is radiated from the sides of the rod. The transition between the radiating and nonradiating condition depends upon the dielectric material, but for materials such as polystyrene ($\epsilon_r = 2.5$) the power outside a circular dielectric is greater than the power inside the dielectric for diameters less than $\lambda/2$. Dielectric rods act as end-fire antennas and are useful in that they occupy less area than a horn of equivalent
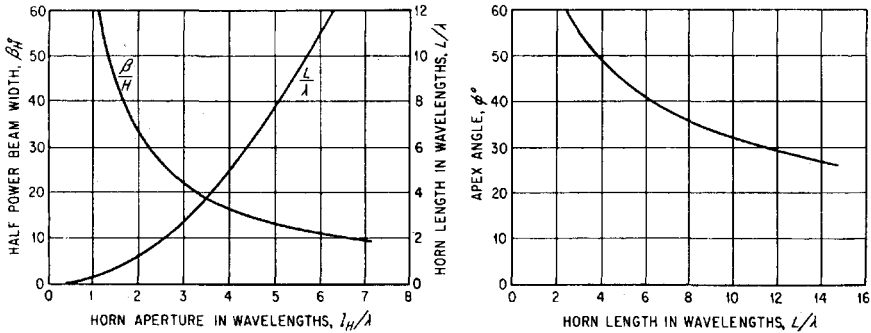
FIG. 21.52. Design curves for optimum *H*-plane horns.

beam width although there is the possibility of cross coupling between fields of adjacent rods if the spacing is too close. The directivity of a dielectric rod antenna is roughly proportional to its length.

21.5. Reflector Systems and Lenses. In general, any highly directive microwave antenna system will consist of a radiating element "illuminating" a reflector or lens. "Pencil" beams and other special-purpose beam shapes can be produced in this manner. The problem of designing such an antenna system is divided into the following parts:

1. Specification, or determination of the over-all (secondary) antenna pattern.

2. Determination of the aperture illumination required to produce this secondary pattern.

3. Determination of the primary pattern which, together with the reflector or lens, will produce the necessary aperture illumination.

The particular radiation element used to illuminate the reflector can be a dipole system, horn system, series of slot radiators, or an array[3] composed of many of these individual elements. The reflector can be cylindrical, paraboloidal, or of complex shape. There is often no unique solution to a given problem, it being a matter of the designer's choice and experience to select a practical combination.[4]

21.5a. Parabolic and Cylindrical Reflectors.[5] The parabola shown in Fig. 21.53a has the useful property that energy radiating from a point source at the focus F

[1] *op. cit.*, p. 412.
[2] See G. E. Mueller and W. A. Tyrrell, Polyrod Antennas, *Bell Telephone System Tech. J.*, vol. 26, pp. 837–851, October, 1947.
[3] See Sec. 21.3.
[4] See Sec. 21.1i and references thereto.
[5] See H. T. Friis and W. D. Lewis, Radar Antennas, *Bell Telephone System Tech. J.*, vol. 26, pp. 219–317, April, 1947, and S. Silver, *op. cit.*

of the parabola will be reflected as a plane wavefront, in phase over the aperture, forming a pattern with high directivity along the axis of the parabola. Conversely, incident energy intercepted by the aperture of the parabola will be reflected so as to concentrate at the focal point. This property makes parabolic reflectors useful in the formation of narrow-beam antenna patterns. As the geometrical properties of the reflecting surface do not depend upon the frequency, such reflectors are very

(a)   (b)   (c)   (d)
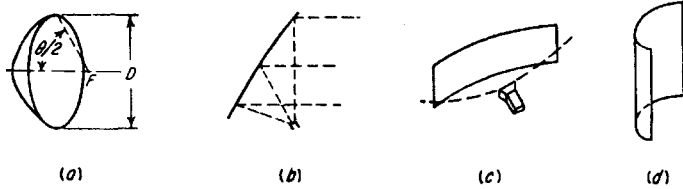
FIG. 21.53. Reflector shapes.

broadband devices assuming the aperture is large compared to the wavelength. The beam width will change with frequency, however, in accordance with Eq. (21.41) for a circular parabola.

$$\beta = k \frac{58\lambda}{D} \text{ deg} \tag{21.41}$$

where β = beam width between half power points
   λ = wavelength
   D = diameter of aperture in units consistent with λ
   k = a factor that depends upon uniformity of illumination of the parabola by the feed; k = unity for perfectly uniform illumination and about 1.25 for illumination tapered down about 10 db at edges of reflector

The gain over a $\lambda/2$ dipole of a large uniformly illuminated circular parabolic reflector is

$$G \approx \frac{6D^2}{\lambda^2} \tag{21.42}$$

The angle θ subtended by the reflector surface from the focal point F is given by

$$\theta = 2 \tan^{-1} \frac{D}{4F} \text{ radians} \tag{21.43}$$

where D = diameter of parabola as shown in Fig. 21.53a in consistent units with F

One of the practical difficulties encountered in the use of parabolic reflectors is the necessity of close tolerances on the shape of the reflector surface to preserve uniform phase across the aperture. Other problems arise in connection with the location of the feed. These include the shadowing effect of the feed, reflections from the center of the parabola back into the feed, and the difficulty of proper illumination of the edges of the parabola by a feed located at the focal point. Horns or coaxial-fed dipoles with reflecting plates presenting direct radiation along the parabolic axis are usually used.[1]

The offset feed shown in Fig. 21.53b illuminating a portion of the parabolic surface is helpful in eliminating feed problems. The cut paraboloid shown in Fig. 21.53c is useful where directivity is required in only one plane.

If high directivity in one plane is required, the cylindrical parabolic reflector shown in Fig. 21.53d is used. This reflector is parabolic in a plane normal to the axis of the

[1] C. C. Cutler, Parabolic Antenna Designs for Microwaves, *Proc. IRE*, vol. 37, pp. 1284–1294, November, 1947.

cylinder, providing whatever directivity is desired in this plane, and is illuminated by an array[1] of radiators along the axis of the cylinder, providing the normal directivity associated with long arrays.

21.5b. Corner Reflectors.[2] The reflectors shown in Fig. 21.54a are a simple means of increasing the directivity of a radiating system. The main lobe of the directional pattern can be determined by considering the equivalent image[3] system of the reflecting sheets as shown in Fig. 21.54a. Reflector angles less than 90° can be used; however, the increased number of images is offset by the decreased aperture for a given

IMAGES FOR A 90° REFLECTOR   WIRE CORNER   SOLID DIHEDRAL   TRIANGULAR TRIHEDRAL   SQUARE TRIHEDRAL

(a)

(b)
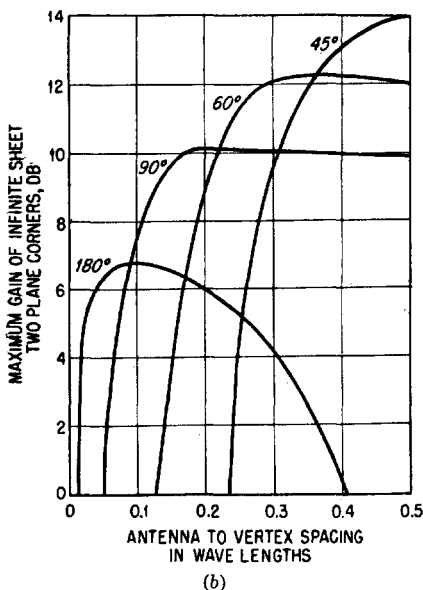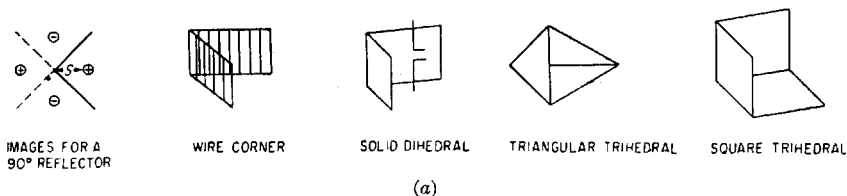
Fig. 21.54. Corner reflectors. (a) Types of reflectors; (b) corner reflector gain as a function of antenna spacing for various corner angles (one ohm antenna loss resistance assumed).

size of reflector. The two-sided, dihedral reflectors are used where directivity in only one plane is required. Power gains of over 10 db compared to a $\lambda/2$ antenna are achievable as shown in Fig. 21.54b. The spacing between wires should be less than 0.1λ, and the wire length should be greater than 0.6λ for the corner of Fig. 21.54a. The feed location for two-sided corners is noncritical. Spacings from the vertex of 0.25λ to 0.65λ are commonly used. The bandwidth decreases, and the gain increases as the spacing is decreased. The length of the 90° corner should be a minimum of 2S (see Fig. 21.54a).

[1] See Sec. 21.3.
[2] See S. D. Robertson, Targets for Microwave Radar Navigation, *Bell Telephone System Tech. J.*, vol. 26, pp. 852–869, October, 1947.
[3] See Sec. 21.1g.

The trihedral reflectors shown in Fig. 21.54a have the property of reflecting an incident signal back toward the transmitting source. They are useful as artificial radar targets and in other applications where it is necessary to reflect a signal back toward the transmitter over a larger angle than is practical with a plane sheet reflector. The sides of trihedral corners should be a minimum of several wavelengths long. The effective aperture of the square corner is approximately equal to the projected area in the direction of the incident signal. The effective aperture of the triangular reflector is approximately 65 per cent of the square reflector. The allowable tolerance on the 90° angles between the reflecting surfaces is of the order of a few tenths of a degree for proper operation.

21.5c. Lenses.[1]   A microwave lens provides directivity by forming an equiphase wavefront over the aperture of the lens. This is accomplished by equalizing the

(a) DIELECTRIC LENS   (b) E-PLANE METAL LENS-UNZONED

(c) E-PLANE METAL LENS-ZONED   (d) ARTIFICIAL DIELECTRIC LENSES
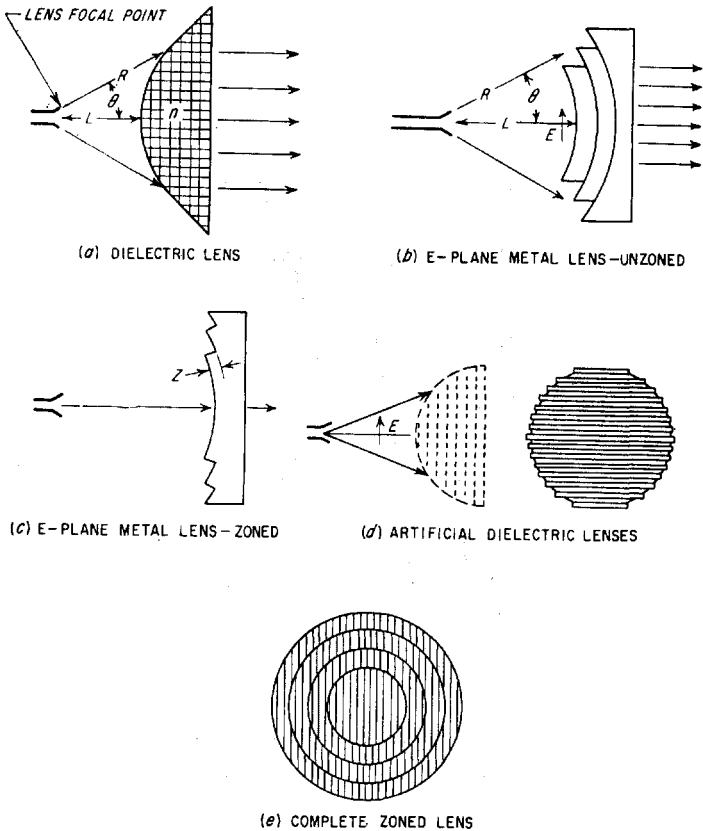
(e) COMPLETE ZONED LENS
Fig. 21.55. Lens configurations.

travel times of "rays" emanating from the illuminating feed as they pass through the lens. The dielectric lens of Fig. 21.55a is directly analogous to optical lenses in that the transit time of the signal through the dielectric is multiplied by the index of

[1] W. E. Koch, Metal-lens Antennas, *Proc. IRE*, vol. 34, pp. 828–836, November, 1946; Metallic Delay Lenses, *Bell Telephone System Tech. J.*, vol. 27, pp. 58–82, January, 1948. J. R. Risser, "Microwave Antenna Theory and Design" edited by S. Silver, chap. 11, McGraw-Hill Book Company, Inc., New York, 1949.

refraction of the material. A hyperbolic surface described by Eq. (21.44) will equalize the delay times in the desired manner.

$$R = \frac{(n - 1)L}{n \cos \theta - 1} \tag{21.44}$$

where L = focal length of lens
R = distance from focal point to lens surface
θ = angle from lens axis
n = index of refraction of lens material (equal to ratio of velocity in air to velocity in dielectric)

The E-plane metal-plate lens shown in Fig. 21.55b takes advantage of the fact that the phase velocity of electromagnetic energy through metallic plates parallel to the electric field is faster than in the unbounded medium. Consequently, the thickness of the lens increases as the path length from the feed increases in order to equalize the transit times for all rays passing through the lens. The effective index of refraction of an E-plane lens is given by Eq. (21.45).

$$n = \frac{v_o}{v} = \sqrt{1 - \left(\frac{\lambda}{2l}\right)^2} \tag{21.45}$$

where λ = wavelength in unbounded medium
l = separation between plates in units consistent with λ

The index of refraction as defined by Eq. (21.45) will always be less than 1. The surface of the lens is ellipsoidal in accordance with Eq. (21.46).

$$R = \frac{(1 - n)L}{1 - n \cos \theta} \tag{21.46}$$

Lenses of this type are frequency-sensitive because of the variation of the apparent index of refraction with frequency. This can be decreased approximately 2:1 by "zoning," as shown in Fig. 21.55c. Zoning is affected by decreasing the path length through the lens by one wavelength whenever the lens thickness exceeds this amount. The minimum thickness of each zone step is given by

$$z = \frac{\lambda}{1 - n} \tag{21.47}$$

where λ = wavelength in free space

The length L of Eq. (21.46) is replaced by L_k given in Eq. (21.48) for zoned lenses.

$$L_k = L + (k - 1)z \tag{21.48}$$

where $k = 1, 2, 3$, etc., is the number of the zone, beginning with the first zone ($k = 1$, $L_k = L$) on the axis of the lens.

The directivity and gain of lenses are comparable to parabolic reflectors of the same aperture. They do not have as close contour tolerances as reflectors. The thickness tolerance is comparable to the contour tolerance of reflectors, but it is easier to control. Lenses are conveniently fed by horn radiators.

A final form of lens construction uses an artificial "dielectric" formed by metallic spheres, rods, or strips as shown in Fig. 21.55d.[1] The effective "dielectric constant" of these particles has been determined by Koch. In general, the metallic particles are smaller than $\lambda/4$ in a direction parallel to the electric field, and the separation is less than 1λ.

[1] For further details on these metallic delay lenses see Koch, *loc. cit.*

Waveform Analysis

Many of the problems encountered by the electronic designer are concerned with the transmission of signals or "waveforms" through electrical networks. Important considerations in these problems are the characteristics of the signals as a function of both time and frequency. This section discusses techniques useful in determining these characteristics. This information and similar data on the properties of the electrical network used for signal transmission are required to determine the distortion which will occur in the signal as it passes through the network.

The analysis of waveforms which are recurrent, or which repeat at a fixed repetition frequency, is very similar in many respects to the analysis of nonrecurrent signals. In either case, it is often possible through the use of a few relatively uncomplicated analysis techniques to determine satisfactorily the characteristics of complex signals. The process can often be simplified by dividing the signal into simple component parts and properly combining the known characteristics of these parts.

22.1. Peak, Root-mean-square, and Average Values of Recurrent Waveforms. The analytical expressions relating the *peak, root-mean-square,* and *average* values of recurrent waveforms are given by

$$A_{\text{rms}} = \left[\frac{1}{T} \int_0^T f(t)^2 \, dt \right]^{1/2} \tag{22.1}$$

$$A_{\text{av}} = \frac{1}{T} \int_0^T f(t) \, dt \tag{22.2}$$

$$C = \frac{A_{\text{max}}}{A_{\text{rms}}} \tag{22.3}$$

$$F = \frac{A_{\text{rms}}}{A_{\text{av}}} \tag{22.4}$$

where $f(t)$ = recurrent function of time under analysis

T = period of one cycle of the function

t = variable, time

A_{rms} = root-mean-square (rms) value of $f(t)$

A_{max} = peak or maximum value of $f(t)$

A_{av} = average value of $f(t)$

C = crest factor (also called peak or amplitude factor)

F = form factor

These quantities are illustrated for a sinusoid in Fig. 22.1.

22.2. Effect of a Superimposed D-C Component. If a d-c component A_{dc} is added to a waveform $f(t)$, the effects upon the peak, rms and average values are as follows.

The instantaneous value of the recurrent function $f(t)$ is changed by direct algebraic addition of the instantaneous and d-c components, i.e.,

$$f'(t) = f(t) + A_{\text{dc}} \tag{22.5}$$

Therefore, the resultant peak value is given by

$$A'_{\text{max}} = A_{\text{max}} + A_{\text{dc}} \tag{22.6}$$

22–2

The variation in $f(t)$ (peak-to-peak difference) is unchanged by the superimposed term.

The rms value A_{rms} is modified as follows:

$$A'_{rms} = \left[\frac{1}{T}\int_0^T [f(t) + A_{dc}]^2\, dt\right]^{1/2} \quad (22.7)$$

The average value is increased or decreased by the superimposed term by direct algebraic addition.

$$A'_{av} = A_{av} + A_{dc} \quad (22.8)$$

22.3. Effect of Superposition of Waveforms. Many commonly encountered waveforms can be regarded as being formed by the superposition of two or more of the waveforms shown in Table

FIG. 22.1. Sinusoidal waveform.

22.1. In this case, the average value of the composite waveform is the sum of the average values of the component waveforms where the amplitude of each is measured from its base. The rms value of the composite waveform is obtained by Eq. (22.1).
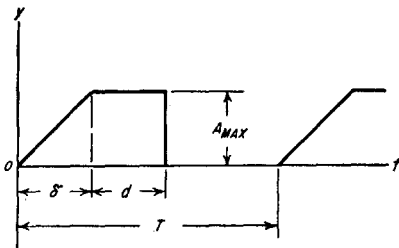
FIG. 22.2 Asymmetrical trapezoidal wave for Example 22.1.

Example 22.1

Calculate the rms value of the asymmetrical trapezoidal wave shown in Fig. 22.2. This waveshape is defined by

$$y_1 = A_{max}\frac{t}{\delta} \quad \text{for } 0 < t < \delta$$

$$y_2 = A_{max} \quad \text{for } \delta < t < d + \delta$$
$$y = 0 \quad \text{for } d + \delta < t < T$$

Solution

1. Evaluate Eq. (22.1) for this waveform:

$$A_{rms} = \left[\frac{1}{T}\int_0^T f(t)^2\, dt\right]^{1/2}$$

or

$$A_{rms}^2 = \frac{1}{T}\left[\int_0^\delta A_{max}^2\frac{t^2}{\delta^2}\, dt + \int_\delta^{\delta+d} A_{max}^2\, dt\right]$$

$$= \frac{1}{T}\left[\frac{A_{max}^2\,\delta}{3} + A_{max}^2\, d\right]$$

$$A_{rms} = A_{max}\sqrt{\frac{\delta + 3d}{3T}}$$

Example 22.2
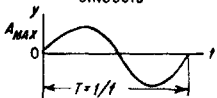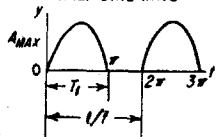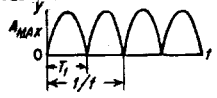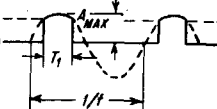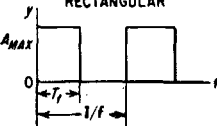
From the triangular waveform in Table 22.1, obtain the rms and average values of the waveform shown in Fig. 22.3.

Solution

1. Evaluate the constants in Table 22.1 for the given waveform as follows:

$$R = 0,\ S = T = 1/f$$

TABLE 22.1. CHARACTERISTICS OF RECURRENT WAVEFORMS

Function	$f(t)$	A_{rms}	A_{av}	C	F
SINUSOID	$y = A_{max} \sin 2\pi ft$	$0.707 A_{max}$	$\dfrac{2A_{max}}{\pi}$ (½ cycle 0 to π or π to 2π) 0 (1 cycle)	1.414	1.11 (½ cycle)
HALF SINE WAVE	$y = A_{max} \sin 2\pi ft$ for $\dfrac{n}{f} < t < \dfrac{n}{f} + T_1$ $y = 0$ for $\dfrac{n}{f} + T_1 < t < \dfrac{n+1}{f}$ $n = 0, 1, 2, \ldots$	$0.5 A_{max}$	$\dfrac{A_{max}}{\pi}$	2	1.57
FULL WAVE RECTIFIED SINE WAVE	$y = |A_{max} \sin 2\pi ft|$	$0.707 A_{max}$	$\dfrac{2A_{max}}{\pi}$	1.414	1.11
PORTION OF A SINUSOID	$y = A_{max} \sin 2\pi ft$ for $\dfrac{1}{4f} - \dfrac{T_1}{2} < t < \dfrac{1}{4f} + \dfrac{T_1}{2}$ $y = 0$ everywhere else	$A_{max} \sqrt{\dfrac{(2\theta + \sin 2\theta)}{4\pi}}$ $2\theta = 2\pi f T_1$	$\dfrac{A_{max} \sin \theta}{\pi}$	$\dfrac{1}{\sqrt{\dfrac{(2\theta + \sin 2\theta)}{4\pi}}}$	$\pi \dfrac{\sqrt{\left(\dfrac{2\theta + \sin 2\theta}{4\pi}\right)}}{\sin \theta}$
RECTANGULAR	$y = A_{max}$ for $\dfrac{n}{f} < t < \dfrac{n}{f} + T_1$ $y = 0$ for $\dfrac{n}{f} + T_1 < f < \dfrac{n+1}{f}$ $n = 0, 1, 2, 3, \ldots$	$A_{max} \sqrt{T_1 f}$	$A_{max} T_1 f$	$\dfrac{1}{\sqrt{T_1 f}}$	$\dfrac{1}{\sqrt{T_1 f}}$

TRAPEZOIDAL 	$y = \dfrac{A_{\max}}{R} t$ for $\dfrac{n}{f} < t < \dfrac{n}{f} + R$ $y = A_{\max}$ for $\dfrac{n}{f} + R < t < \dfrac{n}{f} + R + D$ $y = A_{\max}\left(1 - \dfrac{t-(R+D)}{S}\right)$ for $\dfrac{n}{f} + R + D < t < \dfrac{n}{f} + T_1$ where $S + R + D = T_1$ $y = 0$ for $\dfrac{n}{f} + T_1 < t < \dfrac{n+1}{f}$ $n = 0, 1, 2, 3, \ldots$	$A_{\max}\sqrt{\dfrac{f}{3}(R + S + 3D)}$	$A_{\max}f\left(\dfrac{R}{2} + \dfrac{S}{2} + D\right)$	$\dfrac{1}{\sqrt{\dfrac{f}{3}(R + S + 3D)}}$	$\dfrac{\sqrt{\dfrac{f}{3}(R + S + 3D)}}{f\left(\dfrac{R}{2} + \dfrac{S}{2} + D\right)}$
TRIANGULAR 	$y = A_{\max}\dfrac{t}{R}$ for $\dfrac{n}{f} < t < \dfrac{n}{t} + R$ $y = A_{\max}\left(1 - \dfrac{t - R}{S}\right)$ for $\dfrac{n}{f} + R < t < \dfrac{n}{f} + T_1$ where $R + S = T_1$ $y = 0$ for $\dfrac{n}{f} + T_1 < t < \dfrac{n+1}{f}$ $n = 0, 1, 2, 3, \ldots$	$A_{\max}\sqrt{\dfrac{f}{3}(R + S)}$	$A_{\max}f\left(\dfrac{R}{2} + \dfrac{S}{2}\right)$	$\dfrac{1}{\sqrt{\dfrac{f}{3}(R + S)}}$	$\dfrac{\sqrt{\dfrac{f}{3}(R + S)}}{f\left(\dfrac{R}{2} + \dfrac{S}{2}\right)}$

2. Calculate A_{rms} and A_{av} from the expressions given in Table 22.1.

$$A_{rms} = A_{max} \sqrt{\frac{f}{3}(R + S)}$$

$$= A_{max} \sqrt{\frac{S}{3S}}$$

$$= \frac{A_{max}}{\sqrt{3}}$$

$$A_{av} = A_{max} f \left(\frac{R + S}{2} \right)$$

$$= \frac{A_{max}}{2}$$
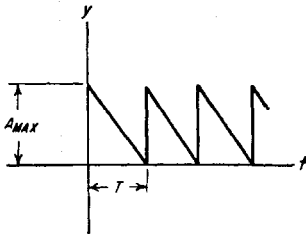
Fig. 22.3. Saw-tooth waveform for Example 22.2.
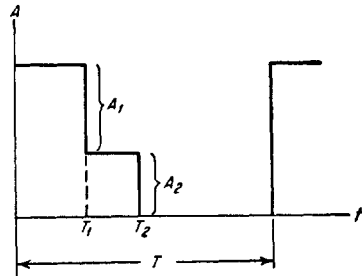
Fig. 22.4. Composite waveform for Example 22.3.

Example 22.3

Calculate the average and rms values of the composite waveform in Fig. 22.4.

Solution

1. Determine the average value of the composite waveform. From the rule given in Sec. 22.3, the average value is equal to the sum of the average values of the two component rectangular waveforms.

From Table 22.1, the average value is

$$A_{av} = \frac{1}{T}(A_1 T_1 + A_2 T_2)$$

2. Determine the rms value of the composite waveform.
From Eq. (22.1), the rms value is

$$A_{rms}^2 = \frac{1}{T} \int_0^T f(t)^2 \, dt = \frac{1}{T} \int_0^{T_1} (A_1 + A_2)^2 \, dt + \frac{1}{T} \int_{T_1}^{T_2} A_2^2 \, dt$$

$$A_{rms} = \left\{ \frac{1}{T} [(A_1 + A_2)^2 T_1 + A_2^2 (T_2 - T_1)] \right\}^{1/2}$$

22.4. Fourier Series. Any periodic waveform can be analyzed as a series of harmonics of the repetition frequency. These signals will have varying amplitudes and phases, the exact nature of the variations depending upon the configuration of the waveform. The peak amplitudes of these frequency components plotted as a function of frequency is known as the frequency spectrum of the signal. The separation of a periodic waveform into its frequency components is accomplished by applica-

tion of the Fourier series. Thus, any periodic waveform can be expressed as follows:

$$e(t) = B_o + \sum_{n=1}^{\infty} (A_n \sin n\omega t + B_n \cos n\omega t) \qquad (22.9)$$

where $e(t)$ = periodic time function

$$B_o = \frac{1}{T} \int_{-T/2}^{+T/2} e(t)\, dt \qquad (22.10)$$

= average or d-c value of waveform

$$B_n = \frac{2}{T} \int_{-T/2}^{T/2} e(t) \cos n\omega t\, dt \qquad (22.11)$$

$$A_n = \frac{2}{T} \int_{-T/2}^{T/2} e(t) \sin n\omega t\, dt \qquad (22.12)$$

where $\omega = 2\pi/T$

If, at a given frequency, the sine and cosine harmonic components are combined, the result is

$$A \sin \omega t + B \cos \omega t = M \sin (\omega t + \psi) \qquad (22.13)$$
$$M = (A^2 + B^2)^{\frac{1}{2}}$$
$$\psi = \tan^{-1} \left(\frac{B}{A}\right)$$

The signal $e(t)$ can then be expressed as

$$e(t) = B_o + M_1 \sin (\omega t + \psi_1) + \cdots + Mn \sin (n\omega t + \psi_n) + \cdots \qquad (22.14)$$

The $B_o, M_1, M_2, \ldots, M_n$ coefficients are the amplitudes of successive harmonics comprising the Fourier amplitude spectrum of $e(t)$. The phase angle ψ_n is the starting phase required of the nth harmonic at $t = 0$. The presence of both sine and cosine terms in the expression for $e(t)$ indicates that the harmonic components have different relative phase angles as well as different amplitudes.

Any periodic function can be classified as an even function, an odd function, or the combination of an even and an odd function. Even functions are symmetrical with respect to the origin, for example, $y = x^2, x^4, x^{-2}, \cos x$, etc. Odd functions are diametrical with respect to the origin, for example, $y = x, x^3, \sin x$, etc. Even and odd functions are illustrated in Fig. 22.5. An even function satisfies Eq. (22.15) whereas an odd function satisfies Eq. (22.16).

(a) EVEN FUNCTION   (b) ODD FUNCTION
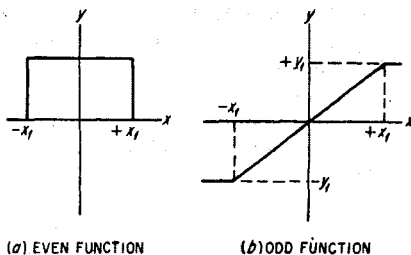
FIG. 22.5. Even and odd functions.

$$f(t) = f(-t) \qquad \text{for an even function} \qquad (22.15)$$
$$f(-t) = -f(t) \qquad \text{for an odd function} \qquad (22.16)$$

Periodic functions which do not satisfy either Eq. (22.15) or (22.16) can be expressed as the sum of an even function $c(t)$ and an odd function $s(t)$.

$$e(t) = c(t) + s(t) \qquad (22.17)$$

For even functions, which contain only cosine terms, $A_n = 0$, $\psi_n = \pi/2$, $M_n = B_n$, and Eq. (22.9) reduces to

$$c(t) = B_o + B_1 \cos \omega t + B_2 \cos 2\omega t + \cdots + B_n \cos n\omega t \qquad (22.18)$$

where $\omega = 2\pi f$

f = fundamental (recurrence) frequency

For odd functions, which contain only sine terms, $B_n = 0$, $\psi_n = 0$, $M_n = A_n$, and Eq. (22.9) reduces to

$$s(t) = A_1 \sin \omega t + A_2 \sin 2\omega t + \cdots + A_n \sin n\omega t \qquad (22.19)$$

The starting phase of each harmonic component of an even or an odd function is the same. To shift the "phase" of a complex wave by θ degrees, the phase of the fundamental is shifted θ degrees, and each harmonic component is shifted $n\theta$ degrees, where n is the order of the harmonic.

The effects of harmonic content upon waveform symmetry are:

1. An even function defined by Eq. (22.15) does not contain sine terms.
2. An odd function defined by Eq. (22.16) does not contain cosine terms.
3. If $f(x)$ is such that $f\left(x + \dfrac{T}{2}\right) = -f(x)$, where T is the period, then $f(x)$ contains only odd harmonics.
4. If $f(x)$ is such that $f\left(x + \dfrac{T}{2}\right) = f(x)$, where T is the period, then $f(x)$ contains only even harmonics.
5. A square wave is an even or odd function containing only odd harmonics.
6. A saw-tooth wave is an odd function containing only even harmonics.

In general, a choice of the origin causing a function to be either even or odd will simplify the harmonic analysis. The use of these rules will often save the attempted computation of harmonic amplitudes which are not present.

22.5. Graphical Analysis. Engineering data are often available in graphical form. A harmonic analysis can be made from these data by the following procedure[1] for periodic functions.

1. Select one complete interval T of the graphical plot of the function.
2. Divide the interval into k equal subintervals of width W.
3. Determine the coefficients A_n of the sine terms in the equivalent Fourier series as follows:

$$A_n = 2 \sum_{m=1}^{m=k} \left[\frac{a_m \sin n\theta_m}{k} \right] \qquad (22.20)$$

where A_n = peak amplitude of nth harmonic

m = number of subinterval = 1, 2, 3, . . . k

a_m = amplitude of mth subinterval in graphical plot of waveform

θ_m = phase angle of mth subinterval at fundamental frequency

$= \dfrac{2\pi m}{k}$ radians

4. Determine the coefficients B_n of the cosine terms from

$$B_n = 2 \sum_{m=1}^{m=k} \left[\frac{a_m \cos n\theta_m}{k} \right] \qquad (22.21)$$

[1] See, for example, pp. 82–87, "Mathematics of Modern Engineering" vol. 1, by R. E. Dogherty and E. G. Keller, John Wiley & Sons, Inc., New York, 1936.

The accuracy of the analysis will increase as the number of intervals is increased. In particular, if the number of intervals is k, the highest harmonic n determined by this method should be less than $k/2$.

This procedure is illustrated by the following example.

Example 22.4

Determine the amplitudes of the fundamental and the first significant harmonic contained in the waveform of Fig. 22.6.

It can be seen that this is an even function about the y axis, and consequently only cosine terms need be evaluated. In addition, $f[x + (T/2)] = -f(x)$, and consequently only odd harmonics are present. The first significant harmonic is the third, and it is only necessary to divide the function into more than six intervals to obtain the value of the 3rd harmonic. Select 10 intervals of 36° each.
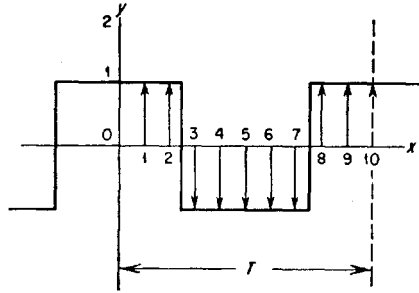
FIG. 22.6. Waveform for Example 22.4.

Solution

1. Determine the amplitude of the fundamental.

Ordinate m	Amplitude a_m	Angle θ_m	Cos θ_m	$a_m \cos \theta_m$	Sin θ_m	$a_n \sin \theta_m$
1	+1	36°	+0.809	+0.809	+0.588	+0.588
2	+1	72°	+0.309	+0.309	+0.951	+0.951
3	-1	108°	-0.309	+0.309	+0.951	-0.951
4	-1	144°	-0.809	+0.809	+0.588	-0.588
5	-1	180°	-1.000	+1.000	0	0
6	-1	216°	-0.809	+0.809	-0.588	+0.588
7	-1	252°	-0.309	+0.309	-0.951	+0.951
8	+1	288°	+0.309	+0.309	-0.951	-0.951
9	+1	324°	+0.809	+0.809	-0.588	-0.588
10	+1	360°	+1.000	+1.000	0	0
				$\Sigma = 6.472$		$\Sigma = 0$

Therefore, $A_1 = \dfrac{6.472}{10} \times 2 = 1.294$.

2. To illustrate the absence of any 2nd harmonic, determine the 2nd-harmonic amplitude.

Ordinate m	Amplitude a_m	Angle θ_m	$2\theta_m$	Cos $2\theta_m$	$a_m \cos 2\theta_m$	Sin $2\theta_m$	Sin $2\theta_m$
1	+1	36°	72°	+0.309	+0.309	+0.951	+0.951
2	+1	72°	144°	-0.809	-0.809	+0.588	+0.588
3	-1	108°	216°	-0.809	+0.809	-0.588	+0.588
4	-1	144°	288°	+0.309	-0.309	-0.951	+0.951
5	-1	180°	360°	+1.000	-1.000	0	0
6	-1	216°	432°	+0.309	-0.309	+0.951	-0.951
7	-1	252°	504°	-0.809	+0.809	+0.588	-0.588
8	+1	288°	576°	-0.809	-0.809	-0.588	-0.588
9	+1	324°	648°	+0.309	+0.309	-0.951	-0.951
10	+1	360°	720°	+1.000	+1.000	0	0
					$\Sigma = 0$		$\Sigma = 0$

Therefore, $A_2 = 0$.

3. Determine the amplitude of the 3rd harmonic.

Ordinate m	Amplitude a_m	Angle θ_m	$3\theta_m$	$\cos 3\theta_m$	$a_m \cos 3\theta_m$
1	+1	36°	108°	−0.309	−0.309
2	+1	72°	216°	−0.809	−0.809
3	−1	108°	324°	+0.809	−0.809
4	−1	144°	432°	+0.309	−0.309
5	−1	180°	540°	−1.000	+1.000
6	−1	216°	648°	+0.309	−0.309
7	−1	252°	756°	+0.809	−0.809
8	+1	288°	864°	−0.809	−0.809
9	+1	324°	972°	−0.309	−0.309
10	+1	360°	1080°	+1.000	+1.000
					$\Sigma = -2.472$

Therefore, $A_3 = \dfrac{2 \times 2.472}{10} = 0.494$.

The minus sign indicates a 180° phase relationship between the 3rd harmonic and the fundamental. The exact amplitude of the fundamental is $4/\pi = 1.272$, and the exact amplitude of the 3rd harmonic is $4/3\pi = 0.425$. The accuracy of the graphical method of analysis increases as the number of intervals per cycle increases.

22.6. Spectra of Periodic and Nonperiodic Signals. The amplitude and phase spectra of periodic and nonperiodic signals can be obtained in a variety of ways. The Fourier series relationship given in Sec. 22.4 can be used directly to obtain the amplitudes of the successive harmonic components in a periodic signal. The phase spectrum can also be determined from Fourier series relationships. It is often more convenient, however, to obtain an expression for the envelope of the amplitude and phase spectra. The amplitude or phase shift at any particular frequency can then be determined by finding the amplitude or phase spectrum envelope at that frequency.

Spectra envelopes are obtainable by use of the Fourier integral or by Laplace transform analysis.

22.6a. Fourier Integral. If the repetition frequency of a periodic signal is continuously decreased, the number of harmonic components in its spectrum will increase correspondingly until, when the limiting condition of a single nonrecurrent signal is reached, the spectrum will contain an infinite number of components spaced an infinitesimal distance apart and the spectrum can be considered to be continuous. By this limiting process, the Fourier series expression becomes the Fourier integral expression:

$$f(t) = \int_{-\infty}^{\infty} A(\omega) \cos \omega t \, d\omega + \int_{-\infty}^{\infty} B(\omega) \sin \omega t \, d\omega \qquad (22.22)$$

This can be expressed more generally as

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)e^{j\omega t} \, d\omega \qquad (22.23)$$

or

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t} \, dt \qquad (22.24)$$

In these expressions $f(t)$ is the expression for the time-amplitude characteristic of the signal and $F(\omega)$ is the complex frequency spectrum. $F(\omega)$ contains an amplitude function describing the relative amplitudes of the harmonic components and a phase function describing the starting phases of all the harmonic components.

If the shape of one interval of a periodic waveform is identical to the shape of a nonperiodic waveform, the complex frequency spectra $F(\omega)$ obtained by Eq. (22.24) for the nonperiodic waveform will be the envelope of the line spectra of the periodic waveform and will differ only in absolute amplitude.

In using Eq. (22.24) to obtain spectra of various waveforms, the analytical expression for the waveform is substituted for $f(t)$, dividing the integral into several segments if necessary. The expression is then integrated and, in general, will result in an algebraic expression containing real and imaginary terms. The amplitude spectrum is the absolute magnitude of the complex expression. The phase spectrum is given by the phase angle of the complex expression. The real part of $F(\omega)$ gives the even part of $f(t)$, while the imaginary part of $F(\omega)$ gives the odd part of $f(t)$.

Equations (22.23) and (22.24) form a Fourier transform pair and complete tables[1] of $F(\omega)$ and $f(t)$ are available for many useful functions. Several functions most useful in engineering application are shown in Fig. 22.7. The use of this table can be extended by noting that $f(t)$ and $F(\omega)$ are interchangeable. For

FIG. 22.7. Fourier integral relationships.

example, the $\dfrac{\sin \omega T/2}{\omega T/2}$ value of $F(\omega)$ corresponding to a pulse $f(t)$ means that a $\sin x/x$ function for $f(t)$ will result in a square spectra $F(\omega)$.

Spectra for more complicated $f(t)$ can be obtained by using the functions in Fig. 22.7 as building blocks. If a waveform can be constructed by the addition of several $f(t)$, for example, $f_1(t) + f_2(t) + f_3(t)$, the spectrum of the resultant waveform is $F_1(\omega) + F_2(\omega) + F_3(\omega)$. A precaution should be observed in the use of this rule, however, in that the transform pairs listed in Fig. 22.7 are for the particular time references shown. If $f_1(t)$, $f_2(t)$, and $f_3(t)$ do not have the indicated time references with respect to $t = 0$, but are shifted in time amounts T_1, T_2, and T_3, respectively, then their respective transforms as given in the table must be multiplied by $e^{-j\omega T}$ where T is the delay interval, that is, $F_1(\omega)$ must be multiplied by $e^{-j\omega T_1}$, $F_2(\omega)$ must be multiplied by $e^{-j\omega T_2}$, and $F_3(\omega)$ must be multiplied by $e^{-j\omega T_3}$ before addition. This procedure is illustrated in Fig. 22.8.

22.6b. Spectrum Analysis by Laplace Transform Analysis. An equivalent but sometimes easier method of determining the spectrum of a function is the use of Laplace transform analysis.[2] The method is identical to that described for the Fourier integral method except the $F(\omega)$ values are obtained from a table of Laplace transform pairs. There is the additional restriction, however, that the function must be zero for all values of time less than zero. After obtaining the Laplace transform $F(s)$ of the original function $f(t)$, the substitution $s = j\omega$ is made, and then the absolute magnitude and phase angle of the resulting $F(\omega)$ are determined. The absolute

---

[1] See, for example, Campbell and Foster, "Fourier Integrals for Practical Application," D. Van Nostrand Company, Inc., Princeton, N.J., 1948.

[2] See Sec. 23.

magnitude squared of the complex expression containing $j\omega$ is obtained by multiplying the expression by its conjugate. The Laplace transform of a function $f(t - T_1)$ starting at time $t = T$, instead of $t = 0$, is $e^{-sT_1}F(s)$ where $F(s)$ is the Laplace transform of $f(t)$.
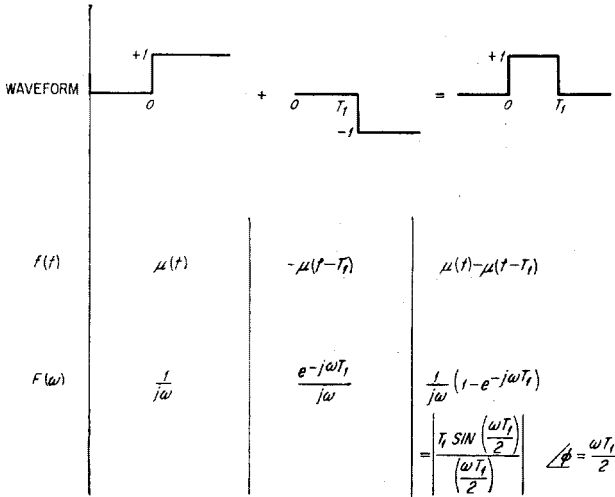
Fig. 22.8. Spectrum addition.
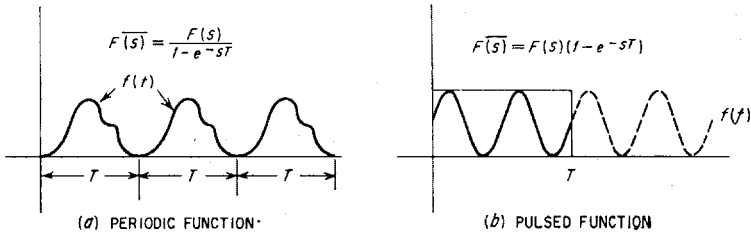
Fig. 22.9. Periodic and pulsed functions.

Several additional rules useful in determining the Laplace transform expression for a function are as follows:

1. *Periodic functions:* If $f(t)$ is the time function for one interval T of a periodic waveform and the Laplace transform of $f(t)$ is $F(s)$, the Laplace transform $\overline{F(s)}$ of the periodic function is

$$\overline{F(s)} = \frac{F(s)}{1 - e^{-sT}} \qquad (22.25)$$

This is illustrated in Fig. 22.9a.

2. *Pulsed or sampled functions:* If a function $f(t)$ has a Laplace transform $F(s)$ and is "pulsed" or "sampled" for an interval T so that it exists only over an interval from $t = 0$ to $t = T$, the Laplace transform $\underline{F(s)}$ of the sampled function is

$$\underline{F(s)} = F(s)(1 - e^{-sT}) \qquad (22.26)$$

This is illustrated in Fig. 22.9b.

Example 22.5

Determine by Fourier integral and Laplace transform analysis the sin x/x spectra for the rectangular pulse shown in Fig. 22.10.

(*a*) PULSE SHAPE $\qquad$ (*b*) AMPLITUDE SPECTRUM
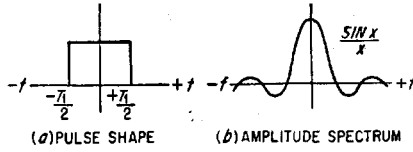
Fɪɢ. 22.10. Spectrum of a rectangular pulse.

Solution

1. Fourier integral analysis
From Eq. (22.24)

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega T} \, dt$$

$$f(t) = A \quad \text{for} \ -\frac{T_1}{2} < t < \frac{T_1}{2}$$

$$= 0 \quad \text{elsewhere}$$

$$\therefore F(\omega) = \int_{-\frac{T_1}{2}}^{\frac{T_1}{2}} A e^{-j\omega T} \, dt$$

$$= -\frac{A}{j\omega} \left(e^{-\frac{j\omega T_1}{2}} - e^{+\frac{j\omega T_1}{2}} \right)$$

$$= A T_1 \frac{\sin (\omega T_1/2)}{\frac{\omega T_1}{2}}$$

2. Laplace transform analysis
a. Express the square wave as a function of time.

$$f(t) = A\mu(t) - A\mu(t - T_1)$$

where $\mu(t)$ is a unit step and the pulse is considered for convenience to start at $t = \boldsymbol{\rho}$
b. Express $f(t)$ in complex form.[1]

$$F(s) = \frac{A}{s} - \frac{A}{s} e^{-sT_1} = \frac{A}{s} (1 - e^{-sT_1})$$

c. Substitute $s = j\omega$.

$$F(j\omega) = \frac{A}{j\omega} (1 - e^{-j\omega T_1})$$

$$= \frac{jA}{\omega} (e^{-j\omega T_1} - 1)$$

d. Obtain $|F(j\omega)|^2$ by multiplying $F(j\omega)$ by its conjugate.

$$|F(j\omega)|^2 = \left[\frac{A}{j\omega} (1 - e^{-j\omega T_1}) \right] \left[\frac{A}{-j\omega} (1 - e^{j\omega T_1}) \right]$$

$$= \frac{A^2}{\omega^2} [2 - (e^{j\omega T_1} + e^{-j\omega T_1})]$$

$$= \frac{A^2}{\omega^2} (2 - 2 \cos \omega T_1)$$

$$= \frac{2A^2}{\omega^2} \left(2 \sin^2 \frac{\omega T_1}{2} \right) = A^2 T_1^2 \frac{\sin^2 (\omega T_1/2)}{(\omega T_1/2)^2}$$

---
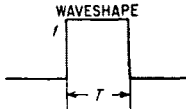
[1] See Sec. 23.

e. Solve for $|F(j\omega)|$.

$$|F(j\omega)| = A T_1 \frac{\sin (\omega T_1/2)}{\omega T_1/2}$$

22.6c. General Characteristics of Amplitude Spectra. Expressions for the envelopes of the amplitude spectra of three representative pulse-type waveforms are given in Table 22.2. These spectra are shown for comparison in Fig. 22.11. The function

TABLE 22.2. REPRESENTATIVE SPECTRUM ENVELOPES

WAVESHAPE	*Amplitude Spectrum*
	$K \dfrac{\sin \omega T/2}{\omega T/2}$
	$K \left(\dfrac{\sin \omega T/4}{\omega T/4}\right)^2$
	$K \left(\dfrac{\sin \omega T/6}{\omega T/6}\right) \left(\dfrac{\sin \omega T/3}{\omega T/3}\right)$

$\sin x/x$, which finds repeated application in spectrum analysis, is plotted in detail in Fig. 22.12 for convenience.

The following general characteristics apply to signal spectra.

1. *Effect of signal repetition frequency on periodic waveforms.* The harmonic components of a spectrum are spaced by an interval equal to the repetition frequency of the waveform. The shape of the envelope of the amplitude spectrum does not vary as the repetition frequency of the waveform is changed, since it is only dependent upon the shape of the signal. The number of harmonic components included in a given portion of the frequency spectrum will increase, however, as the repetition frequency is decreased. The d-c or average value will vary in proportion to the repetition frequency. As the number of frequency components is increased with decreasing repetition frequency the energy carried by each component must decrease to keep the waveform energy constant; consequently, the maximum amplitude of the frequency spectrum decreases with decreasing repetition frequency and increases with increasing repetition frequency.

2. *Effect of signal amplitude.* As the amplitude of the waveform is varied with all other characteristics remaining constant, the amplitude of the frequency spectrum and the average value will vary in proportion to the signal amplitude.

3. *Effect of signal duration.* As the duration of the signal is decreased or increased, the intercepts of the spectrum on the frequency axis will move away from or toward the origin, respectively. The intercept spacing is inversely proportional to the signal width. This is sometimes referred to as "reciprocal spreading," since for short signal durations the spectrum is widened and for long signal durations the spectrum is shortened. It can be stated, in general, that the steeper the slope or the greater the maximum rate of change of any portion of the waveform, the higher will be the number of frequency components of significant amplitude required to reproduce the waveform.

4. *Effect of an additional variable on spectra.* It is sometimes necessary to determine the spectrum of a periodic signal which is subjected to amplitude, position, duration, or some other form of modulation. A general approach is to determine first the

spectrum of the unmodulated signal and then determine the effect of the modulation on the spectrum. A general treatment of this problem is beyond the intended scope of this text; however, the following rules can often be applied.[1]

1. *Amplitude modulation.* The spectrum of one waveform amplitude modulated by another is the spectrum of the product of the basic signal and the modulating
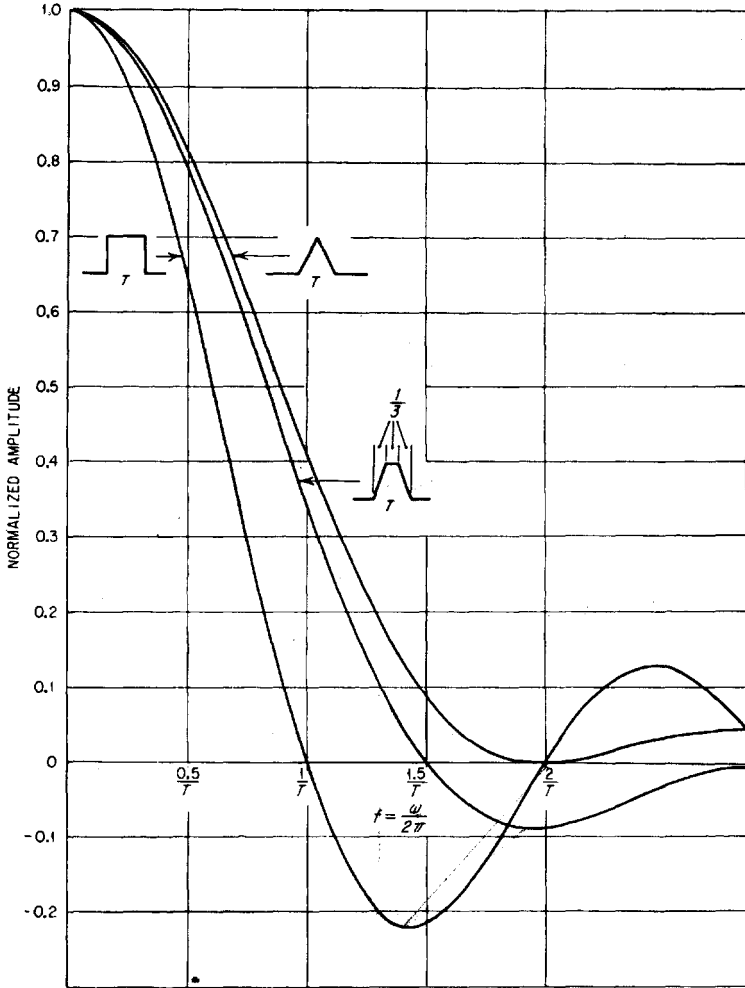
FIG. 22.11. Comparative spectra of pulse waveforms.

signal. This results in the production of sidebands on each side of each spectrum component of the basic signal separated by the modulation frequency from each basic signal component for each frequency component present in the modulation signal. The envelope of the spectrum of the basic signal will not be affected by the modulation. The amplitude of the modulation sideband components of the spectrum will be proportional to the modulation amplitude.

[1] For further information see S. C. Losier, Spectrum Analysis of Pulse Modulated Waves, *Bell Telephone System Tech. J.*, April, 1947.

2. Position modulation. Pulse position modulation will, in general, not alter the envelope of the amplitude spectrum. The harmonic spacing within the envelope will be nonuniform, however, and the phase spectrum will be altered at each frequency by an amount equal to the angular frequency times the time displacement.
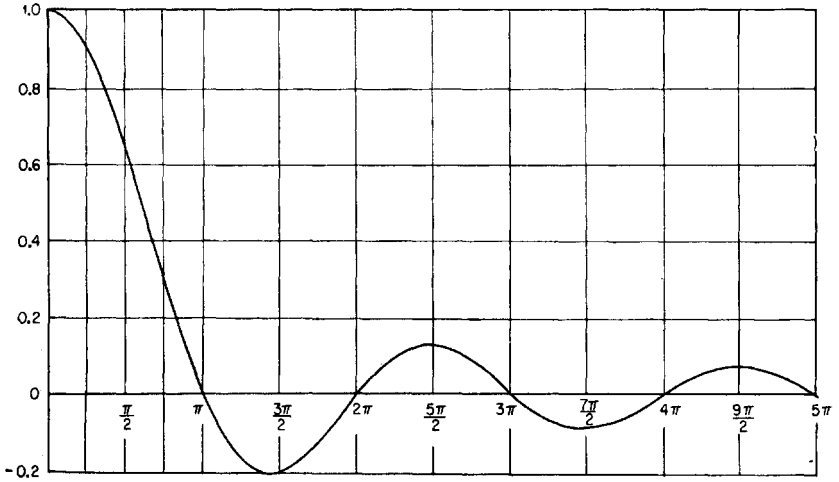
FIG. 22.12. Plot of the function $\dfrac{\sin x}{x}$.
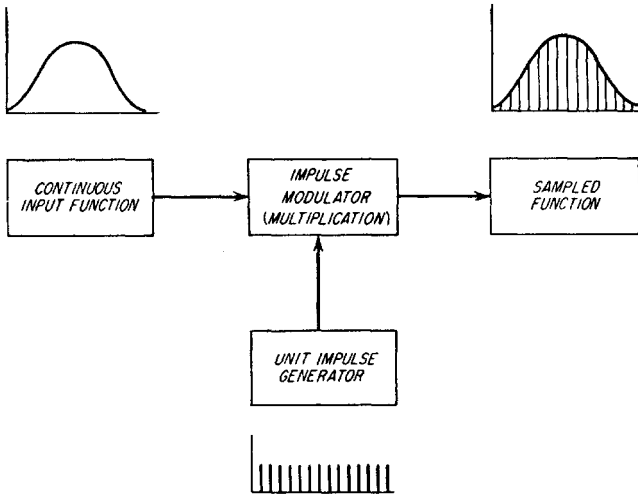
FIG. 22.13. Sampling process.

3. Duration modulation. Pulse duration modulation will affect the amplitude spectrum by altering the spectrum width inversely with the pulse width. If the duration change is symmetrical about the center of the pulse, there will be no effect on the phase spectrum; otherwise there will be an additional effect corresponding to the position modulation of the pulse by an amount equal to the difference in center position of the signal before and after modulation.

22.7. Spectra of Sampled Signals.[1] It is sometimes convenient or necessary to sample a continuous function by measuring the instantaneous value of the function at regular intervals of time.   A new function is formed by this process consisting of a series of regularly spaced impulses whose amplitudes equal the amplitude of the continuous function at the time of sampling.   This is illustrated in Fig. 22.13.

The Laplace transform $U^*(s)$ of the train of impulses is

$$U^*(s) = \frac{1}{1 - e^{-sT}} \qquad (22.27)$$

where T = period between impulses (sampling interval)

If $F(s)$ is the Laplace transform of a continuous input function $f(t)$, the transform of the function sampled at intervals of T is[2]

$$F^*(s) = F(s) * \frac{1}{1 - e^{-sT}} = \sum_{K=-\infty}^{\infty} \frac{1}{T} F(s + jK\Omega) \qquad (22.28)$$

where $K = 1, 2, 3, \ldots$

Ω = angular sampling frequency = $2\pi/T$

Sampling transforms the spectrum of the input function to an infinite series of spectra spaced at harmonics of the sampling frequency and with upper and lower sidebands around these frequencies.   This is illustrated in Fig. 22.14.
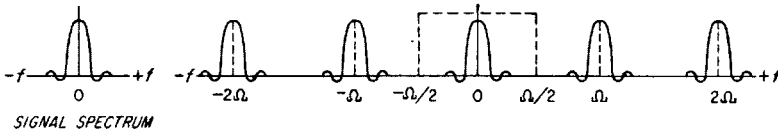
SIGNAL SPECTRUM

FIG. 22.14. Spectrum of a sampled signal.

Desampling is accomplished by passing the sampled signal through a filter which passes only the frequencies contained in one of the spectra of Fig. 22.14, such as the region indicated by the dotted line from $-\Omega/2$ to $+\Omega/2$.

The maximum frequency that can be identified in the sampled function is $\Omega/2$. Therefore, the sampling frequency must be at least twice the highest significant frequency in the input function.   If the polarity of alternate samples is reversed, the resulting spectrum of the sampled function consists of upper and lower sidebands around only the odd harmonics of the sampling frequency, that is, $\pm\Omega$, $\pm3\Omega$, $\pm5\Omega$, etc.

[1] For additional information, see S. Goldman, "Information Theory," Prentice-Hall, Inc., Englewood Cliffs, N.J., 1953.

[2] Where * indicates convolution in the s domain.   See J. G. Truxal, "Automatic Feedback Control System Synthesis," pp. 500–507, McGraw-Hill Book Company, Inc., New York, 1955.

Network Analysis

23.1. Network Geometry. Any network can be considered as a geometric array, each part of which has a physical equivalent. To describe adequately the conditions within a network, each part must be completely defined. The following material establishes such definitions and gives a criterion for choosing between loop and nodal bases of network analysis.

23.1a. Nodes and Elements. If the numbered points in Fig. 23.1a are designated as *nodes*, they may be considered as the terminals of the lines connecting these points and the lines may be designated as *elements* of the network. Each element necessarily has associated with it two terminal nodes; however, a series connection of two elements

(a) AN ARRAY OF NODES AND ELEMENTS

—— TREE BRANCH   ———— ASSOCIATED LINK

(b) TREES AND LINKS
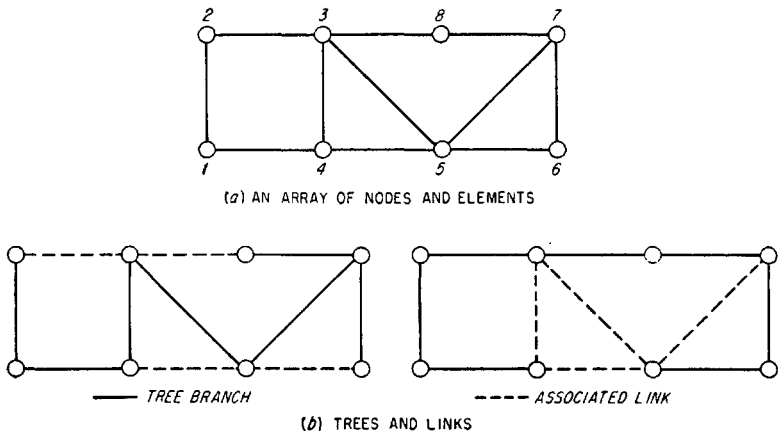
FIG. 23.1. Geometrical network.

results in two separate nodes being replaced by a single node at the point of connection. A parallel connection of two elements also results in two separate nodes being replaced by a single node at each point of connection. If one node of a network is designated as a "reference node," e.g., a point at ground potential, then the "pairs of nodes" formed by taking this reference node and each of the other nodes of the network are called the *independent node pairs* of the network.

23.1b. Trees, Links, and Branches.[1] A network can be divided into two portions called *trees* and *links*. A tree of a given network is any combination of the elements of the network which connects all of the nodes but which contains no closed loops. The links of a network are those elements associated with a given tree of the network which are required to complete the network, i.e., to close all the loops in the tree. Figure 23.1b shows two possible sets of trees and links associated with the network of Fig. 23.1a. The elements of a network form the *branches* of the network. Branches associated with a given tree are *tree branches*. The remaining branches are the links of the network, associated with the given tree.

23.1c. Choice of Loops and Nodes. The tree concept of network analysis provides

[1] See E. A. Guilleman, "Introductory Circuit Theory," John Wiley & Sons, Inc., New York, 1953.

a means of determining the number of unknown currents in a network and also specifies an independent set of these unknowns. In particular, the number of unknown loop currents is equal to the number of links associated with a given network, and the independent loops are those loops formed by successively inserting each link into the tree.

In a similar manner, the number of unknown branch voltages is equal to the number of tree branches in the network tree, and the tree-branch voltages form an independent set of such voltages.

23.1d. Separate Parts. A single branch, loop, or group of loops and branches having no conductive or capacitive coupling with any other part of the network is called a *separate part.* Two separate parts may be inductively coupled.

23.1e. Criterion for Selection of Loop or Nodal Analysis. Networks can be analyzed on either the loop or the nodal basis. The following equations describe the relation between the portions of the network defined above and are useful in establishing the form of the network equations most easily solved, i.e., the form of solution containing the least number of unknowns. The number of independent loop currents in a network is equal to the number of independent loops, and the number of independent node-pair voltages in a network is equal to the number of independent node pairs or tree branches. In establishing the number of independent node pairs in a network, a reference node must be designated for each separate part of the network. These quantities are computed from Eqs. (23.1) and (23.2).

$$l = e - n + s \qquad (23.1)$$
$$n_s = n - s \qquad (23.2)$$

where l = number of independent loops

n_s = number of independent node pairs or tree branches

e = number of elements

n = number of nodes

s = number of separate parts

Example 23.1

Determine the number of independent loops and node pairs in Fig. 23.1.

Solution

Number of separate parts.................... $s = 1$
Number of nodes............................. $n = 8$
Number of elements.......................... $e = 11$
    Number of independent node pairs........... $n_s = 8 - 1 = 7$
    Number of independent loops............... $l = 11 - 8 + 1 = 4$

Example 23.2

Determine the least number of independent loops or independent node pairs in the following networks.

Solution

In Fig. 23.2a

$$s = 1, n = 6, e = 9$$
$$n_s = 6 - 1 = 5$$
$$l = 9 - 6 + 1 = 4$$

Therefore, there are fewer loop currents.

In Fig. 23.2b

$$s = 1, n = 5, e = 8$$
$$n_s = 5 - 1 = 4$$
$$l = 8 - 5 + 1 = 4$$

Therefore, there are an equal number of loop currents and node-pair voltages.

In Fig. 23.2c

$$s = 2, \ n = 4, \ e = 6$$
$$n_s = 4 - 2 = 2$$
$$l = 6 - 4 + 2 = 4$$

Therefore, there are fewer node-pair voltages.

(a) (b) (c)
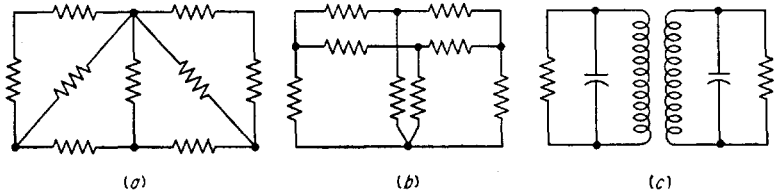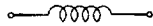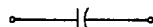
FIG. 23.2. Networks for Example 23.2.

23.2. Passive Elements Including Mutual Inductance

23.2a. Resistance, Capacitance, and Inductance. Energy dissipation and storage of energy in electric fields and in magnetic fields denote passive elements within the network. These passive elements are the resistances, capacitances, and inductances, respectively, of the network. Table 23.1 gives the relationships for the voltage between terminals of each passive element, the current through the element, and the magnitude of the element.

When the loop basis is used for network equations, the passive elements are most simply expressed in terms of resistance R, inductance L, and elastance S. When the nodal basis is used for network equations, the passive elements are most simply

TABLE 23.1. PASSIVE ELEMENTS

Element	Symbol	Unit	Voltage-current relationships	Eq .
Resistance or conductance	R $\;\;$ —⋀⋀⋀— $\;\;$ G	ohm $\;\;\;$ mho	$e = iR$ $R = 1/G$ $i = Ge$	(23.3) (23.4) (23.5)
Inductance or inverse inductance	L $\;\;$ —⟨0000⟩— $\;\;$ Γ_L	henry $\;\;\;$ 1/henry	$e = L \dfrac{di}{dt}$ $L^* = 1/\Gamma_L$ $i = \Gamma_L \displaystyle\int_0^t e \, dt + i_o$ $i_o = $ current through L at $t = 0$	(23.6) (23.7) (23.8)
Elastance or capacitance	S $\;\;$ —⊣⊢— $\;\;$ C	1/farad $\;\;\;$ farad	$e = S \displaystyle\int_0^t i \, dt + e_o$ $C = 1/S$ $i = C \dfrac{de}{dt}$ $e_o = $ voltage across C at $t = 0$	(23.9) (23.10) (23.11)

* Assuming no mutual inductance associated with L (see Sec. 23.2d).

expressed in terms of conductance G, inverse inductance Γ_L, and capacitance C. The relationships between R and G, L and Γ_L, and S and C are given in Table 23.1. The relationship between L and Γ_L given in the table only applies for inductances with no mutual coupling, as explained in Sec. 23.2d.

23.2b. Linearity of Passive Elements. It is ordinarily assumed that passive elements are linear, i.e., their magnitude does not depend upon the voltage applied to the element or the current through the element. When the passive elements are not linear, then either a succession of solutions must be obtained for all values of the nonlinear element which are of interest, or the applicable nonlinear differential equations must be solved.

23.2c. Mutual Inductance. Mutual inductance is present in any magnetic circuit where flux from a given inductance links with another inductance.

Inductance can be defined as the number of flux linkages occurring per unit value of current. Applying this definition to the circuit of Fig. 23.3 the following expressions for mutual inductance M are obtained.

$$M_{12} = \frac{N_1\,\Delta\phi_{12}}{\Delta I_2} \tag{23.12}$$

$$M_{21} = \frac{N_2\,\Delta\phi_{21}}{\Delta I_1} \tag{23.13}$$

where $\Delta\phi_{21}$ = change in magnetic flux that links turns of L_2 due to a current change ΔI_1 in L_1

$\Delta\phi_{12}$ = change in magnetic flux that links turns of L_1 due to a current change ΔI_2 in L_2

N_1 = number of turns of L_1

N_2 = number of turns of L_2

If the units of ϕ and I are webers and amperes, respectively, M will be in henrys. The first subscript of M refers to the inductance across which a voltage has been induced because of the current through another inductance, designated by the second subscript. For linear inductances, the mutual inductances as seen from either inductance are identical; that is, $M_{12} = M_{21} = M$.

Equations corresponding to Eqs. (23.12) and (23.13) for the case of self-inductance are

$$L_1 = \frac{N_1\,\Delta\phi_{11}}{\Delta I_1} \tag{23.14}$$

$$L_2 = \frac{N_2\,\Delta\phi_{22}}{\Delta I_2} \tag{23.15}$$

where $\Delta\phi_{11}$ = change in magnetic flux that links turns of L_1 due to a current change ΔI_1 in L_1
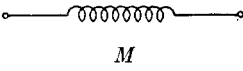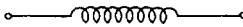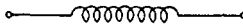
$\Delta\phi_{22}$ = change in magnetic flux that links turns of L_2 due to a current change ΔI_2 in L_2

Table 23.2 gives the relationships for mutual inductance corresponding to those given in Table 23.1 for self-inductance.

In Table 23.2, the plus sign is used with positive M and the minus sign is used with negative M.
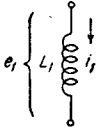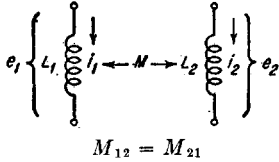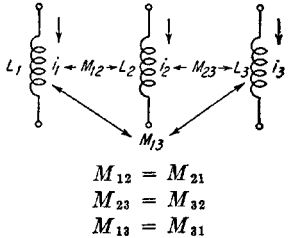
23.2d. Inverse Mutual and Inverse Self-inductance. Once the magnitudes of the self and mutual inductances have been obtained, they can be used directly in the equations of Tables 23.1 and 23.2. The inverse of self-inductance Γ_L is not the inverse of the self-inductance of the coil when mutual inductance is present. This is because the current through the coil is proportional to the integral of the voltage across the coil. The constant of proportionality accounts for the effect on the current of mutual

TABLE 23.2. MUTUAL INDUCTANCE

Element	Symbol	Unit	Voltage-current relationships
Mutual inductance	M	henry	$e_1 = (\pm)M\dfrac{di_2}{dt}$ $e_2 = (\pm)M\dfrac{di_1}{dt}$
Inverse mutual inductance	Γ_M	1/henry	$i_2 = -(\pm)\Gamma_M \displaystyle\int e_1\,dt$ $i_1 = -(\pm)\Gamma_M \displaystyle\int e_2\,dt$ $\Gamma_M = \dfrac{M}{L_1 L_2 - M^2}$

coupling with other coils and includes the self and mutual inductances of the coupled circuits. For a similar reason, the inverse mutual inductance Γ_M is not the direct inverse of the mutual inductance of two coupled coils, but includes the effect of the self-inductances of the coils. Exact expressions for the inverse inductance coefficients depend upon the number of coupled elements. There are, however, a few arrangements of elements which will satisfy a great majority of the situations likely to be encountered in practice. Table 23.3 gives the correct expressions for these cases.

TABLE 23.3. INVERSE MUTUAL INDUCTANCE

Circuit configuration	Inverse inductance
	$M = 0$ $\Gamma_L = L_1^{-1}$
  $M_{12} = M_{21}$	$\Gamma_M = \dfrac{M}{L_1 L_2 - M^2}, \ \Gamma_{L1} = \dfrac{L_2}{L_1 L_2 - M^2}$ $\Gamma_{L2} = \dfrac{L_1}{L_1 L_2 - M^2}$
  $M_{12} = M_{21}$ $M_{23} = M_{32}$ $M_{13} = M_{31}$	$\Gamma_{L1} = \dfrac{L_2 L_3 - M_{23}{}^2}{\Delta}, \ \Gamma_{M12} = \dfrac{[M_{12}L_3] - [M_{23}M_{13}]}{\Delta}$ $\Gamma_{L2} = \dfrac{L_1 L_3 - M_{13}{}^2}{\Delta}, \ \Gamma_{M23} = \dfrac{[M_{23}L_1] - [M_{12}M_{13}]}{\Delta}$ $\Gamma_{L3} = \dfrac{L_1 L_2 - M_{12}{}^2}{\Delta}, \ \Gamma_{M13} = \dfrac{[M_{13}L_2] - [M_{12}M_{23}]}{\Delta}$ where $\Delta = L_1 L_2 L_3 - L_1 M_{23}{}^2 - L_2 M_{13}{}^2 - L_3 M_{12}{}^2$ $+ 2M_{12}M_{13}M_{23}$

The following procedure, applied to the case of two inductances, can be used to determine the inverse inductances in any arbitrary network.

L_1 HAS N_1 TURNS
L_2 HAS N_2 TURNS
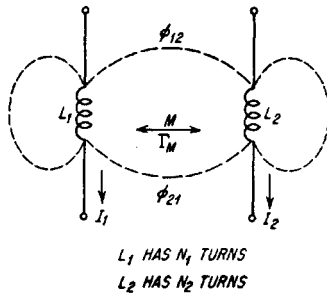
FIG. 23.3. Inductances mutually coupled.

General Step	*Example*
1. Write the set of equations for the voltage across each inductance. See Fig. 23.3.	1. $L_1 \dfrac{di_1}{dt} + M \dfrac{di_2}{dt} = e_1$ $M \dfrac{di_1}{dt} + L_2 \dfrac{di_2}{dt} = e_2$
2. Solve the equations for the current through each inductance.	2. $i_1 = \dfrac{L_2}{L_1 L_2 - M^2} \displaystyle\int e_1\, dt - \dfrac{M}{L_1 L_2 - M^2} \displaystyle\int e_2\, dt$ $i_2 = \dfrac{-M}{L_1 L_2 - M^2} \displaystyle\int e_1\, dt + \dfrac{L_1}{L_1 L_2 - M^2} \displaystyle\int e_2\, dt$
3. From Tables 23.1 and 23.2 write the general equations for total current through an inductance.	3. $i_1 = \Gamma_{L_1} \int e_1\, dt$ from Table 23.1 $i_1 = -(\pm)\Gamma_M \int e_2\, dt$ from Table 23.2 Total $i_1 = \Gamma_{L_1} \int e_1\, dt - (\pm)\Gamma_M \int e_2\, dt$ Total $i_2 = -(\pm)\Gamma_M \int e_1\, dt + \Gamma_{L_1} \int e_2\, dt$
4. Equate the coefficients of the integrals obtained in step 2 with those obtained in step 3.	4. $\Gamma_{L_1} = \dfrac{L_2}{L_1 L_2 - M^2}$ $\Gamma_{L_2} = \dfrac{L_1}{L_1 L_2 - M^2}$ $\Gamma_M = \dfrac{M}{L_1 L_2 - M^2}$

23.2e. Algebraic Signs of Mutual-inductance Terms. In circuit analysis, the algebraic sign associated with the mutual-inductance term is either positive or negative. To determine the proper sign it is convenient to place dots adjacent to the end of each coupled coil having the same instantaneous polarity. These dots can be located by either of the following two procedures:

1. *Winding sense and orientation.*
   a. Assume a potential across one of the circuit inductances and, with a dot on the circuit schematic, indicate the terminal of the inductance having the positive potential.
   b. For the assumed voltage in step a, establish the direction of the resulting flux through the inductance (see Fig. 23.4).
   c. For each of the other inductances, establish the terminal of the inductance to which an applied positive potential would generate flux in the same direction as the flux developed in step b.
   d. Indicate these positive terminals established in step c with dots on the circuit schematic.

2. *External measurements.* If the coils are not available for inspection as to winding sense, the following procedure can be used:

   a. Momentarily connect a battery across one coil.

   b. Establish the induced polarity of the voltage across each coil.

   c. Place a dot at the end of each coil having a positive polarity.

A positive sign for M should be used in writing circuit equations if the assumed currents flow in the same direction, as determined by the dots, through mutually coupled coils. A negative sign should be used if the directions of current flow are opposite. In writing nodal equations, the signs of the inverse mutual-inductance coefficients in Table 23.3 can be determined by substituting the proper positive or negative sign to each of the mutual inductance terms, as determined above.

23.2f. Effect of Series and Parallel Connections. If two coils are connected in series or in parallel, the effective inductance of the combination is given by the following:

$$L_{\text{series}} = L_1 + L_2 + (\pm)2M \qquad (23.16)$$

$$L_{\text{parallel}} = \frac{L_1 L_2 - M^2}{L_1 + L_2 - (\pm)2M} \qquad (23.17)$$

M can be either positive or negative depending upon the coil connections.

FIG. 23.4. Right-hand rule for determining the relation between flux and current directions. The direction of ϕ is the direction the fingers of the right-hand curl around the conductor with the thumb pointing in the direction of conventional current flow (opposite to electron flow).

These expressions are often useful in measurements of coupled coils. For example, if the effective inductance of two coils in series is measured and then the terminals of one of the coils are interchanged and the measurement repeated, half the sum of the measured values will equal $(L_1 + L_2)$ and one-quarter of the difference between the measurements will equal M. This can be seen by addition and subtraction of the two possible values of Eq. (23.16).
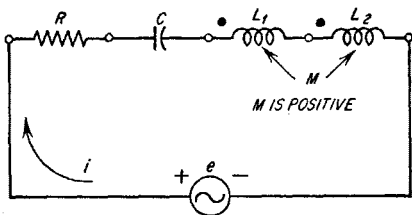
FIG. 23.5. Series circuit for Example 23.3.
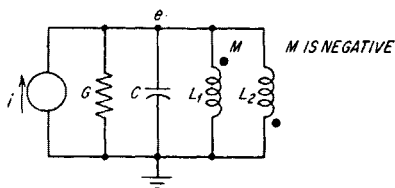
FIG. 23.6. Parallel circuit for Example 23.3.

Example 23.3

1. Write the expression for the total voltage across the series circuit shown in Fig. 23.5.

$$e = iR + S \int i\, dt + L_1 \frac{di}{dt} + L_2 \frac{di}{dt} + (+)2M \frac{di}{dt}$$

2. Write the expression for the total current through the parallel circuit shown in Fig. 23.6.

$$i = Ge + C\frac{de}{dt} + \Gamma_{L_1}\int e\,dt + \Gamma_{L_2}\int e\,dt - (-)2\Gamma_M\int e\,dt$$

where $\quad \Gamma_{L_1} = \dfrac{L_2}{L_1L_2 - M^2}, \Gamma_{L_2} = \dfrac{L_1}{L_1L_2 - M^2}, \Gamma_M = \dfrac{M}{L_1L_2 - M^2}$

3. Determine the sign of the mutual inductance in the coupled circuit shown in Fig. 23.7.
By application of the right-hand rule, the flux due to i in L_1 will be downward. The flux due to i in L_2 will also be downward. Therefore, the sign of M is positive.

23.3 Active Elements. An active element is a source of energy within the network, such as a battery, vacuum tube, etc. Active elements can be regarded as either voltage generators or current generators. An ideal voltage generator has zero internal impedance, and consequently, the terminal voltage equals the generated voltage regardless of the load connected across the generator terminals. An ideal current generator has an infinite internal impedance, and the current is independent of the load applied to the terminals. In practice, however, both voltage and current generators have finite internal impedances. The equivalent circuit of a voltage generator e_s having an internal impedance Z_s is shown in Fig. 23.8a. The equivalent circuit of a current generator i_s having an internal admittance Y_s is shown in Fig. 23.8b.

FIG. 23.7. Determination of the sign of mutual inductance.

Voltage sources can be replaced by current sources, and current sources can be replaced by voltage sources. Figure 23.8c shows the equivalence between voltage and current generators.

Where tubes are included in the network, reference should be made to Sec. 3.2 for the method of reduction to equivalent generators.

(a) EQUIVALENT CIRCUIT OF A VOLTAGE GENERATOR

(b) EQUIVALENT CIRCUIT OF A CURRENT GENERATOR

(c) RELATIONSHIPS FOR THE EQUIVALENCE OF A VOLTAGE AND A CURRENT SOURCE

FIG. 23.8. Equivalent generators.

$$i = \frac{e_1}{Z_1} + \frac{e_2}{Z_2} + \frac{e_3}{Z_3} \qquad Y = \frac{1}{Z_1} + \frac{1}{Z_2} + \frac{1}{Z_3}$$
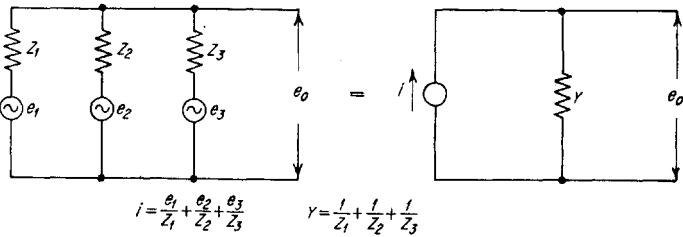
(a) COMBINATION OF VOLTAGE SOURCES

$$e = \frac{I_1}{Y_1} + \frac{I_2}{Y_2} + \frac{I_3}{Y_3}$$
$$Z = \frac{1}{Y_1} + \frac{1}{Y_2} + \frac{1}{Y_3}$$
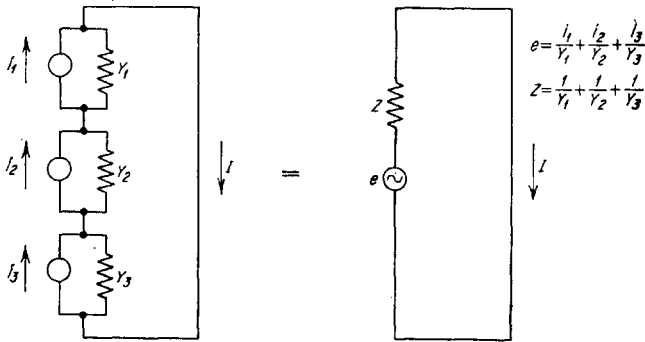
(b) COMBINATION OF CURRENT SOURCES

FIG. 23.9. Combination of sources.

Multiple current or voltage sources can be combined into a single equivalent voltage or current source as shown in Fig. 23.9.

23.4. Loop Equations. The loop equations of a network are defined as the system of equations obtained by equating the voltage drops across loop elements to the voltage sources within the closed loops of the network. These equations can be obtained by the following procedure:

1. *Given a network, determine the number of independent loop currents.* The number of independent loop currents is given by

$$\mu_i = l - i_s \tag{23.18}$$

where μ_i = number of independent loop currents
l = number of independent loops
i_s = number of independent current generators included in closed loops of network

In any network composed of two or more meshes, there may be several closed-loop paths which can be traced around the network and for which loop equations can be written. However, only l of these equations will be independent. l, as used in Eq. (23.18), is given by

$$l = e - n + s \tag{23.19}$$

where e = number of active and passive elements included in the closed loops of the network
n = number of nodes
s = number of separate parts

2. Select a particular set of the independent loops determined in step 1.

3. Assume a direction of current flow around each loop. These directions are to be considered the directions of positive current flow in each loop. The probable current directions can often be determined by inspection. If the actual current flow is opposite in direction to an assumed current it will be indicated by a negative sign in the final expression for the loop current.

4. Write the equation for each loop, equating the sum of the voltage drops in the direction of loop current to the sum of the voltage sources in the same direction. Loop equations are expressions of voltage equilibrium around each loop. It should be noted that where an element of the network is common to two or more loops, the total current through the element is the sum of all the loop currents passing through the element. These loops are said to be mutually coupled through the element. In this case, the voltage changes across such mutual elements caused by currents in loops other than the one being considered must be included in the loop equation.

Each loop equation will, in general, be of the following form illustrated for loop 1 of a network containing n loops.

$$z_{11}i_1 + z_{12}i_2 + z_{13}i_3 + \cdots + z_{1n}i_n = e_1 \qquad (23.20)$$

where z_{11} = total *self-impedance* of loop 1

$\qquad z_{12}$ = total *mutual impedance* between loop 1 and loop 2

$z_{13} \ldots z_{1n}$ = similar mutual impedances between loop 1 and other loops

$\qquad i_1$ = loop current for loop 1

$i_2 \ldots i_n$ = other loop currents

$\qquad e_1$ = total generated (source) voltage in loop 1

In the most general case the impedances are composed of R, L, and S elements and are expressed as

$$z_{jk} = R_{jk} + sL_{jk} + \frac{S_{jk}}{s} \qquad (23.21)$$

where R_{jk} = total resistance through which i_j and i_k flow

$\quad L_{jk}$ = total inductance through which i_j and i_k flow

$\quad S_{jk} = 1/C_{jk}$ = total elastance through which i_j and i_k flow

$\quad s = d/dt$; $1/s = \int dt$ for general case

$\quad s = j\omega$ for steady-state case

5. Tabulate the loop equations obtained in step 4. The loop currents and voltages are related by the following set of simultaneous equations:

$$
\begin{aligned}
z_{11}i_1 + z_{12}i_2 + z_{13}i_3 + \cdots + z_{1n}i_n &= e_1 \\
z_{21}i_1 + z_{22}i_2 + z_{23}i_3 + \cdots + z_{2n}i_n &= e_2 \\
z_{31}i_1 + z_{32}i_2 + z_{33}i_3 + \cdots + z_{3n}i_n &= e_3 \\
& \qquad\qquad (23.22) \\
z_{n1}i_1 + z_{n2}i_2 + z_{n3}i_3 + \cdots + z_{nn}i_n &= e_n
\end{aligned}
$$

which may be expressed in condensed form as

$$\sum_{k=1}^{n} z_{jk}i_k = e_j \qquad (j = 1, 2, \cdots n) \qquad (23.23)$$

where for each value of j, the summation is taken for all n values of k. The subscript j refers to the loop whose equation is being written, and the subscript k refers to the current being considered in the equation for loop j.

Example 23.4

1. Write the loop equations for the network shown in Fig. 23.10.
a. From Eq. (23.19), the number of independent loops is

$$l = e - n + s = 5 - 4 + 1 = 2$$

and the number of unknown loop currents from Eq. (23.18) is

$$\mu_i = l - i_s = 2 - 0 = 2$$

b. Select the loops as shown in the figure and assign directions of positive current in each loop.
c. Write the loop equations by equating voltage drops and rises in each loop.   For loop 1,

$$i_1 R_1 + i_1 R_2 - i_2 R_2 = E$$

and for loop 2,

$$-i_1 R_2 + i_2 R_2 + i_2 R_3 + i_2 R_4 = 0$$

These may be rearranged as follows:

$$(R_1 + R_2)i_1 - R_2 i_2 = E_1$$
$$-R_2 i_1 + (R_2 + R_3 + R_4)i_2 = 0$$

which is the same form as Eq. (23.22) where

$$z_{11} = R_1 + R_2, \; z_{12} = z_{21} = -R_2,$$
$$z_{22} = R_2 + R_3 + R_4$$

FIG. 23.10. Network for Example 23.4 with two independent loops.

It should be noted that i_1 and i_2 pass through R_2 in opposite directions, causing voltage drops of opposite polarity, resulting in the negative signs for the terms in the above equations.

Example 23.5

Determine the loop equations in the network shown in Fig. 23.11.
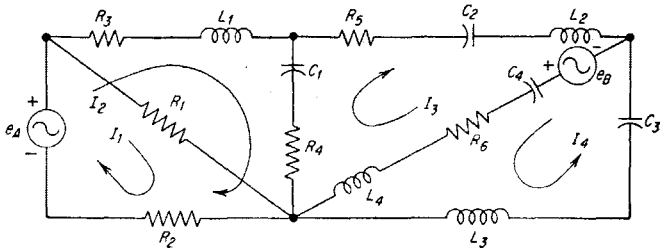
FIG. 23.11. Network for Example 23.5.

Solution

a. From Eq. (23.19)

$$l = e - n + s = 16 - 13 + 1 = 4$$

b. From Eq. (23.18)

$$\mu_i = l - i_s = 4 - 0 = 4$$

c. These four unknown loop currents may be selected as shown in Fig. 23.11.   There will be four loop equations of the following form:

$$z_{11}i_1 + z_{12}i_2 + z_{13}i_3 + z_{14}i_4 = e_1$$
$$z_{21}i_1 + z_{22}i_2 + z_{23}i_3 + z_{24}i_4 = e_2$$
$$z_{31}i_1 + z_{32}i_2 + z_{33}i_3 + z_{34}i_4 = e_3$$
$$z_{41}i_1 + z_{42}i_2 + z_{43}i_3 + z_{44}i_4 = e_4$$

d. Determine the z_{jk} and e_j terms in these equations by inspection of the network as follows:

$$z_{11} = R_1 + R_2$$

$$z_{22} = (R_2 + R_3 + R_4) + sL_1 + \frac{1}{sC_1}$$

$$z_{33} = (R_4 + R_5 + R_6) + s(L_2 + L_4) + \frac{1}{s}\left(\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_4}\right)$$

$$z_{44} = R_6 + s(L_3 + L_4) + \frac{1}{s}\left(\frac{1}{C_3} + \frac{1}{C_4}\right)$$

$$z_{12} = z_{21} = +R_2 \qquad z_{23} = z_{32} = -\left(R_4 + \frac{1}{sC_1}\right)$$

$$z_{13} = z_{31} = 0 \qquad z_{24} = z_{42} = 0$$

$$z_{14} = z_{41} = 0 \qquad z_{34} = z_{43} = \left(R_6 + sL_4 + \frac{1}{sC_4}\right)$$

$$e_1 = e_A,\, e_2 = e_A,\, e_3 = e_B,\, e_4 = e_B$$

These expressions when substituted in the four equations will give the complete loop equations for the network.

23.5. Nodal Equations. The nodal equations of a network are defined as the system of equations obtained by equating the incoming currents to the outgoing currents at each node. These equations can be obtained by the following procedure:

1. *Given a network, determine the number of independent node-pair voltages.* The number of independent node-pair voltages is given by the following:

$$u_v = n_s - e_s \tag{23.24}$$

where u_v = number of independent node-pair voltages

n_s = number of independent node pairs

e_s = number of independent voltage generators

The expressions for current flowing into or away from a given node involve the admittances of the elements connected to that node and the potential difference between the given node and each of the other nodes to which it is connected.

To compute the number of independent node-pair voltages u_v existing in the network, the number of independent node pairs n_s must be determined. For every separate part of the network, one of the nodes must be established as a reference node in order that potential differences throughout the network can be measured with respect to a common reference. This reference is usually taken as the ground point of the network. The number of independent node pairs n_s as used in Eq. (23.24), is, therefore, given by

$$n_s = n - s \tag{23.25}$$

where n = number of nodes

s = number of separate parts

2. *Arbitrarily select the reference node(s) and identify all other nodes in the network.*

3. *Assume relative polarities of each node with respect to each adjacent node.* This consists of drawing an arrow adjacent to each element showing an assumed direction of current flow through the element and thereby designating currents flowing into and away from each node. Current flow should be shown toward the lower potential end of each element.

4. *Write the equation for each node by equating the expressions for current flowing into each node to the expressions for current leaving each node.* Nodal equations are expressions of current equilibrium at each node. Note that each element of the network must be common to two nodes and provides mutual coupling between these nodes.

5. *Tabulate the nodal equations obtained in step 4.* Each nodal equation will, in general, be of the following form illustrated for node 1 of n nodes:

$$y_{11}e_1 - y_{12}e_2 \cdots - y_{1n}e_n = i_1 \tag{23.26}$$

where y_{11} = total "self-admittance" of node 1 and is admittance between node 1 and all others when other nodes are all shorted together $(y_{11} = y_1 + y_{12} + y_{13} + \cdots + y_{1n})$.

$y_{12}, y_{13}, \quad \ldots , y_{1n}$ = admittance of element connected directly between nodes 1-2, 1-3, etc. Mutual admittances must be connected directly between two nodes in question without passing through any other node. Nodes having no such direct connection have zero mutual admittance coupling.

e_1 = potential between node 1 and reference node.

i_1 = net current from all generators connected directly to node 1.

The internal impedances of all generators must be taken into account in determining the self or mutual admittances of all nodes. In the general case the network admittances are composed of resistance, inductance, and capacitance and are expressed as

$$y_{jk} = G_{jk} + \frac{\Gamma_{L_{jk}}}{s} + sC_{jk} \qquad (23.27)$$

where G_{jk} = total conductance between nodes j and k

Γ_{jk} = total inverse inductance between nodes j and k

C_{jk} = total capacitance between nodes j and k

The nodal equations can be expressed by the following set:

$$
\begin{aligned}
y_{11}e_1 - y_{12}e_2 - y_{13}e_3 - \cdots - y_{1n}e_n &= i_1 \\
-y_{21}e_1 + y_{22}e_2 - y_{23}e_3 - \cdots - y_{2n}e_n &= i_2 \\
-y_{31}e_1 - y_{32}e_2 + y_{33}e_3 - \cdots - y_{3n}e_n &= i_3 \\
\vdots \qquad \vdots \qquad \vdots \qquad\qquad \vdots \qquad \vdots \\
-y_{n1}e_1 - y_{n2}e_2 - y_{n3}e_3 - \cdots + y_{nn}e_n &= i_n
\end{aligned}
\qquad (23.28)
$$

which may be further condensed into

$$\sum_{k=1}^{n} (-1)^{\delta} y_{jk}e_k = i_j \qquad (j = 1, 2, \cdots n) \qquad (23.29)$$

where $\delta = 0$ for $j \neq k$; 1 for $j = k$

For each value of j the summation is taken for all n values of k; j refers to the node whose equation is being written.

Example 23.6

Write the nodal equations for the network shown in Fig. 23.12.

FIG. 23.12. Network for Example 23.6.
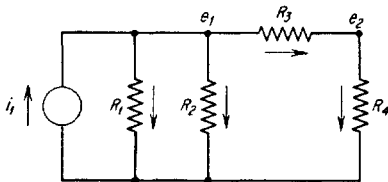
Solution

1. From Eq. (23.25) the number of independent node pairs is

$$n_s = n - s = 3 - 1 = 2$$

and the number of unknown node-pair voltages is, from Eq. (23.24),

$$u_v = n_s - e_s = 2 - 0 = 2$$

2. Designate the nodes, assign polarities with respect to the reference, and designate current directions through each element as shown in the figure.

3. Write the nodal equations by equating expressions for the currents flowing into and away from the points of unknown node-pair voltage e_1 and e_2. For node 1,

$$\frac{e_1 - o}{R_1} + \frac{e_1 - o}{R_2} + \frac{e_1 - e_2}{R_3} - i_1 = 0$$

and for node 2,

$$\frac{e_2 - o}{R_4} + \frac{e_2 - e_1}{R_3} = 0$$

These may be rearranged as follows:

$$\left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right) e_1 - \frac{1}{R_3} e_2 = i_1$$

$$-\frac{1}{R_3} e_1 + \left(\frac{1}{R_4} + \frac{1}{R_3}\right) e_2 = 0$$

which is the same form as Eq. (23.28) where

$$y_{11} = \left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right) = G_1 + G_2 + G_3$$

$$y_{22} = \left(\frac{1}{R_4} + \frac{1}{R_3}\right) = G_3 + G_4$$

$$y_{12} = y_{21} = \frac{1}{R_3} = G_3$$

Example 23.7

Write the nodal equations for the network shown in Fig. 23.13.
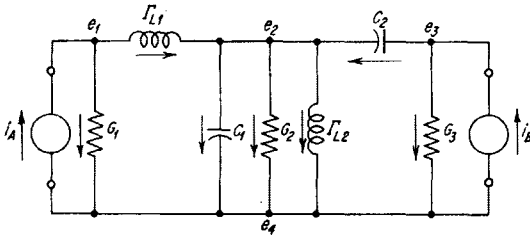
Fig. 23.13. Network for Example 23.7 with three unknown node-pair voltages.

Solution

1. From Eq. (23.25)
$$n_s = n - s = 4 - 1 = 3$$
2. From Eq. (23.24)
$$u_v = n_s - e_s = 3 - 0 = 3$$

3. The three unknown node-pair voltages may be selected as shown in the figure with node $4(e_4)$ taken as reference. There will be three nodal equations of the following form:

$$y_{11}e_1 - y_{12}e_2 - y_{13}e_3 = i_1$$
$$-y_{21}e_1 + y_{22}e_2 - y_{23}e_3 = i_2$$
$$-y_{31}e_1 - y_{32}e_2 + y_{33}e_3 = i_3$$

Determine the y_{jk} and i_j terms in these equations by inspection of the network as follows:

$$y_{11} = \left(G_1 + \frac{\Gamma_{L_1}}{s}\right) \qquad y_{12} = y_{21} = \left(\frac{\Gamma_{L_1}}{s}\right)$$

$$y_{22} = G_2 + \frac{\Gamma_{L_1} + \Gamma_{L_2}}{s} + s(C_1 + C_2) \qquad y_{13} = y_{31} = 0$$

$$y_{33} = G_3 + sC_2 \qquad y_{23} = y_{32} = sC_2$$

$$i_1 = i_A, \ i_2 = 0, \ i_3 = i_B$$

These expressions, when substituted in the above three equations, will give the complete nodal equations for the network.

23.6. Solution of Network Equations. Any given network can be analyzed on either a loop or nodal basis. In general, the method involving the least number of

unknowns is preferable (see Sec. 23.1e). Loop analysis deals with the impedances of a network and involves expressions of voltage equilibrium. Loop currents are the dependent variables, and the source voltages of the network are usually known. Nodal analysis deals with the admittances of a network and involves expressions of current equilibrium. Node-pair voltages are the dependent variables, and the source currents of the network are usually known. Nodal analysis is most applicable to networks with a majority of shunt elements, and loop analysis is most applicable to networks with a majority of series elements.

23.6a. Algebraic Solution. When there are no reactive (i.e., inductance or capacitance) elements within the network, a complete solution to the loop or nodal equations can be obtained by algebraic means. This requires writing the set of loop or nodal equations describing the behavior of the network. A set of simultaneous linear algebraic equations will result which can be solved for the required unknown(s) by conventional methods. This process is most easily explained by a specific example.
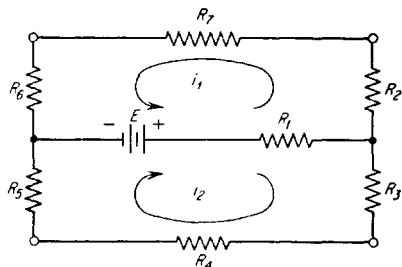
Fig. 23.14. Resistance network of Example 23.8.

Example 23.8

Given the network shown in Fig. 23.14 determine all the loop currents.

Solution

1. The loop equations are as follows:

$$(R_1 + R_2 + R_6 + R_7)i_1 + R_1 i_2 = E$$
$$R_1 i_1 + (R_1 + R_3 + R_4 + R_5)i_2 = E$$

The currents i_1 and i_2 are the dependent variables. The voltage E is the independent variable.

2. To solve for the currents, the system determinant is formed from the coefficients of the current (dependent variable) terms

$$\Delta = \begin{vmatrix} (R_1 + R_2 + R_6 + R_7) & (R_1) \\ (R_1) & (R_1 + R_3 + R_4 + R_5) \end{vmatrix}$$

This determinant is evaluated as shown below (see Sec. 23.15c).

$$\Delta = (R_1 + R_2 + R_6 + R_7)(R_1 + R_3 + R_4 + R_5) - R_1{}^2$$

3. The minor determinants are formed as follows:

$$\Delta_1 = \begin{vmatrix} E & R_1 \\ E & (R_1 + R_3 + R_4 + R_5) \end{vmatrix} \qquad \Delta_2 = \begin{vmatrix} (R_1 + R_2 + R_6 + R_7) & E \\ R_1 & E \end{vmatrix}$$

These determinants are evaluated as shown below.

$$\Delta_1 = E(R_1 + R_3 + R_4 + R_5) - ER_1 = E(R_3 + R_4 + R_5)$$
$$\Delta_2 = E(R_1 + R_2 + R_6 + R_7) - ER_1 = E(R_2 + R_6 + R_7)$$

4. Obtain the required solution by forming the quotient Δ_1/Δ and Δ_2/Δ.

$$i_1 = \frac{\Delta_1}{\Delta} = E \frac{R_3 + R_4 + R_5}{(R_1 + R_3 + R_4 + R_5)(R_1 + R_2 + R_6 + R_7) - R_1{}^2}$$

and

$$i_2 = \frac{\Delta_2}{\Delta} = E \frac{R_2 + R_6 + R_7}{(R_1 + R_3 + R_4 + R_5)(R_1 + R_2 + R_6 + R_7) - R_1{}^2}$$

23.6b. Limitation of Algebraic Solution. The expressions given above are the complete solutions to a network containing only resistive elements. It can be seen, however, that if reactive elements were present the terms for voltage across and current through these elements would involve integrals and derivatives. The solution to the network would then require the solution of a set of linear integrodifferential equations with constant coefficients. The procedure outlined in the following paragraphs can be used to solve these equations.

23.6c. Solution of General Network Equations by Laplace Transforms. This treatment is not intended as a complete exposition of the use of transform methods in the solution of network problems. It is expected, however, that the reader who has previously mastered the theoretical development will find here a concise summary of the most useful relationships. It is further expected that the reader who has not previously used this technique will find the simplified material useful in solving many problems which would be quite difficult by other techniques.[1] In particular, the material contained in Sec. 23.6d covering graphical solutions of networks is a very practical means of reducing the amount of computational labor required in the solution of network problems.

The use of Laplace transforms in the solution of network problems may be compared with the use of logarithms to simplify the solution of mathematical computations. In the use of logarithms, a group of numbers is transformed into a different group of numbers called the logarithms of the original numbers. This is usually accomplished by means of a previously prepared table. The logarithms are then manipulated to achieve the desired mathematical operation, and the results interpreted in terms of the original numbers by means of another previously prepared table or the same table. The advantage gained by this procedure is the reduction of the difficulty and complexity of the mathematical operation. For example, the process of multiplication of the original numbers is reduced to the process of addition of their logarithms; the process of division is reduced to the process of subtraction, etc. In a similar manner, Laplace transforms can be used to reduce the problem of solving integrodifferential equations to a purely algebraic process, and moreover, complex excitations such as square waves, etc., can be analyzed by these same algebraic procedures. This requires a table of "transform pairs" which performs a function similar to the table of logarithms.

Laplace Transforms. The Laplace transform of a given function of time $f(t)$ is defined as

$$\pounds f(t) = F(s) = \int_0^\infty f(t)e^{-st}\, dt \qquad (23.30)$$

The quantity s is a complex variable composed of a real part α and an imaginary part ω as given in Eq. (23.31), and t is the real time variable

$$s = \alpha + j\omega \qquad (23.31)$$

The following examples illustrate the use of Eq. (23.30) to construct a table of transform pairs.

Example 23.9

1. Obtain the Laplace transform of the unit step function $u(t)$ defined as follows:

$$u(t) = 0 \qquad t < 0$$
$$u(t) = 1 \qquad t > 0$$

From Eq. (23.30)

$$F(s) = \int_0^\infty u(t)e^{-st}\, dt = \int_0^\infty e^{-st}\, dt = \frac{1}{s}$$

Therefore, the transform of $u(t)$ is $1/s$.

2. Obtain the transform of $e^{-\alpha t}$. From Eq. (23.30)

$$F(s) = \int_0^\infty e^{-\alpha t}e^{-st}\, dt$$

$$= \int_0^\infty e^{-(s+\alpha)t}\, dt = \frac{1}{s+\alpha}$$

---

[1] See M. F. Gardner and J. L. Barnes, "Transients in Linear Systems," vol. I, John Wiley & Sons, Inc., New York, 1942.

Table 23.4 contains a few transform pairs. Extensive tables covering a wide variety of functions will be found in standard references.[1]

The first portion of Table 23.4 gives the transforms of certain functions of time. The second portion gives a few transforms of mathematical operations and some identities useful in network calculations. These require some additional explanation to demonstrate their utility.

The *translation theorem* (1) is used to determine the transform of a function $f(t - \alpha)$ where $F(s)$ is the transform of $f(t)$. This theorem is useful in evaluating the effect of time delays in a physical system.

The *differentiation theorem* (2) is most useful in transforming differential equations into algebraic equations. The transform includes the "initial condition" $f(0+)$ in the transformed expression. The initial condition is the value of $f(t)$ for $t = 0$. In the event of a discontinuity at $t = 0$, the limiting value for $t > 0$ is selected.

The *integration theorem* (3) is most useful in transforming integral equations into algebraic equations. As before, the additional term $f(0+)$ is used to account for initial conditions at $t = 0$.

It can be seen from the differentiation and integration theorems that derivative and integral expressions can be reduced to a simplified form by substitution of s for the derivative, and $1/s$ for the integral.

The *linearity theorem* (4) is useful in the transformation of sums of $f(t)$. The quantity K must be independent of t and s.

The *final-value theorem* (5) is used to evaluate the limiting (d-c steady-state) value of a function of time in terms of the limit of its transform as s approaches zero. Restrictions on this theorem are that both $f(t)$ and its first derivative must be Laplace-transformable and there must not be singularities[2] of $sF(s)$ on the imaginary axis or in the right half plane. This excludes the use of sinusoidal functions for $f(t)$.

The *initial-value theorem* (6) is used to evaluate the value of $f(t)$ just after $t = 0$ in terms of the limiting value of its transform as s approaches infinity. There are no restrictions on singularities for this theorem.

The *steady-state theorem* (7) is not, strictly speaking, a theorem of operational calculus. It is a substitution that is made for convenience, in reducing complex variable expressions in s to a form suitable for analysis by ordinary a-c circuit theory for steady-state conditions. This substitution amounts to setting $\alpha = 0$ in Eq. (23.31), which is valid for sinusoidal excitations.

The *scale-change theorem* (8) is useful in altering the value of the variable t for ease of handling.

The *complex differentiation theorem* (9) shows that a derivative expression in complex form reduces to a multiplication when expressed as a function of the real variable.

The *complex integration theorem* (10) shows that an integral expression in complex form reduces to a division when expressed as a function of the real variable.

The Inverse Laplace Transform. Before undertaking the general solution of networks by transform methods, it is necessary to consider the methods of performing the inverse transform process; i.e., methods of obtaining a function of time $f(t)$ from a complex expression $F(s)$. Where a complex expression can be identified

---

[1] More complete tables can be found in the following references: M. F. Garner and J. L. Barnes, *op. cit.*, Appendix A. C. A. Campbell and R. M. Foster, "Fourier Integrals," D. Van Nostrand, Company, Inc., Princeton, N.J., 1948. Admiralty Computing Service, Department of Scientific Research and Experiment, London, "Dictionary of Laplace Transforms," 1948. A. Erdelyi (ed.), "Tables of Integral Transforms," McGraw-Hill Book Company, Inc., New York, 1954.

[2] A knowledge of functions of a complex variable is necessary for proper application of this restriction. See M. F. Gardner and J. L. Barnes, *op. cit.*, or N. W. McLaughlan, "Complex Variable and Operational Calculus with Technical Applications," The Macmillan Company, New York, 1944.

TABLE 23.4. LAPLACE TRANSFORM PAIRS

Real function, $f(t)$	Complex function, $F(s)$
Mathematical functions	
1. Unit impulse	1. 1
2. $u(t) = 0, t < 0$ } unit $u(t) = 1, t > 0$ } step	2. $\dfrac{1}{s}$
3. $e^{-\alpha t}$	3. $\dfrac{1}{s + \alpha}$
4. $\dfrac{1}{\omega} \sin \omega t$	4. $\dfrac{1}{s^2 + \omega^2}$
5. $\dfrac{1}{\omega} \sinh \omega t$	5. $\dfrac{1}{s^2 - \omega^2}$
6. $\cos \omega t$	6. $\dfrac{s}{s^2 + \omega^2}$
7. $\cosh \omega t$	7. $\dfrac{s}{s^2 - \omega^2}$
8. $\dfrac{1}{\omega} e^{-\alpha t} \sin \omega t$	8. $\dfrac{1}{(s + \alpha)^2 + \omega^2}$
9. $e^{-\alpha t} \cos \omega t$	9. $\dfrac{s + \alpha}{(s + \alpha)^2 + \omega^2}$
10. t	10. $\dfrac{1}{s^2}$
11. $\dfrac{1}{(n - 1)!} t^{n-1}$	11. $\dfrac{1}{s^n}$, n a positive integer
12. $te^{-\alpha t}$	12. $\dfrac{1}{(s + \alpha)^2}$
13. $e^{-\alpha t} f(t)$	13. $F(s + \alpha)$
14. $(1 - \alpha t)e^{-\alpha t}$	14. $\dfrac{s}{(s + \alpha)^2}$
15. $\dfrac{e^{-\beta t} - e^{-\alpha t}}{\alpha - \beta}$	15. $\dfrac{1}{(s + \alpha)(s + \beta)}$
16. $\dfrac{\alpha e^{-\alpha t} - \beta e^{-\beta t}}{\alpha - \beta}$	16. $\dfrac{s}{(s + \alpha)(s + \beta)}$
Mathematical operations and identities		

1. Translation in time.............	$f(t - \alpha)$	$e^{-\alpha s} F(s)$
2. Differentiation.................	$\dfrac{df(t)}{dt}$	$sF(s) - f(0+)$
3. Integration*....................	$\int f(t)\, dt$	$\dfrac{F(s)}{s} + \dfrac{f^{-1}(0+)}{s}$
4. Linearity......................	$Kf(t)$ $f_1(t) \pm f_2(t)$	$KF(s)$ $F_1(s) \pm F_2(s)$
5. Final value....................	$\lim_{t \to \infty} f(t)$	$\lim_{s \to 0} sF(s)$
6. Initial value...................	$\lim_{t \to 0} f(t)$	$\lim_{s \to \infty} sF(s)$
7. Steady-state analysis............	$j\omega$	s
8. Scale change...................	$f\left(\dfrac{t}{\alpha}\right)$	$\alpha F(\alpha s)$
9. Complex differentiation..........	$tf(t)$	$-\dfrac{d}{ds} F(s)$
10. Complex integration.............	$\dfrac{1}{t} f(t)$	$\displaystyle\int_{s}^{\infty} F(s)\, ds$

* The factor $f^{-1}(0+)$ is an integration constant obtained by determining the value of the integral of $f(t)$ evaluated for t equal to zero. If a discontinuity exists in $f(t)$ at t equals zero, the value of t on the positive side of zero should be used.

directly as the sum of one or more of the forms $F(s)$ appearing in Table 23.4, the inverse transformation can be written down directly as the corresponding sum of one or more of the forms of $f(t)$ from the table.   In general, however, this fortunate circumstance will not occur, and some algebraic manipulation of the complex form such as partial fraction expansion must be employed to separate the terms of the complex form into identifiable transforms, or direct integration of the complex function $F(s)$ must be performed to obtain $f(t)$.

Discussion of the Response Transform.   In general, the algebraic solution of the applicable set of network equations expressed in complex form will result in an expression that can be separated into the following component parts:

$$\text{Response transform} = \text{system transform} \times \text{excitation transform} \qquad (23.32)$$

The *system function* will usually be a rational transform in s.[1]   The system function depends upon the values of the component parts of the network and determines the characteristic of the transient response.

The *excitation* transform includes the transform of the excitation signal together with any initial voltages or currents within the network at $t = 0$.

The *response transform* is the product of the system transform and the excitation transform.   When the response transform is the ratio two of polynomials in s, the inverse transform can be obtained from the following method of partial fractions.

Partial Fraction Expansion.   The response transform can be represented in many cases by a fraction consisting of two polynomials $A(s)$ and $B(s)$ in s.

$$F(s) \;=\; \frac{A(s)}{B(s)} \;=\; \frac{a_m s^m + a_{m-1} s^{m-1} + \cdots + a_1 s + a_0}{s^n + b_{n-1} s^{n-1} + \cdots + b_1 + b_0} \qquad (23.33)$$

where the a's and b's are constants and $n > m$ for the fraction to be rational.   Equation (23.33) can be expressed as

$$F(s) \;=\; \frac{A(s)}{B(s)} \;=\; \frac{A(s)}{(s - s_1)(s - s_2)(s - s_3) \cdots (s - s_n)} \qquad (23.34)$$

where $s_1, s_2, \ldots, s_n$ are the factors (roots) of the denominator and can be either real or complex.

The *form* of the response of the system is determined entirely by these roots, which are called the *poles* of the expression $F(s)$.

If the poles of Eq. (23.34) are single-order (that is, $s_1, s_2, \ldots, s_n$ are all different), the inverse transform is obtained by breaking the expression into the sum of a number of simple expressions, each of which can be transformed by the use of Table 23.4. This is accomplished by the following partial fraction expansion:

$$\frac{A(s)}{B(s)} \;=\; \frac{k_1}{s - s_1} + \frac{k_2}{s - s_2} + \cdots + \frac{k_n}{s - s_n} \qquad (23.35)$$

The coefficients $k_1, k_2, \ldots, k_n$ can be evaluated as follows:

$$k_n \;=\; \left[\frac{A(s)}{B(s)} (s - s_n) \right]_{s=s_n} \qquad (23.36)$$

---

[1] A rational function is an analytic function containing no essential singularities.   An analytic function is such that the function and its first derivative are finite and single-valued.   A singularity is a point where a function is not analytic.   An essential singularity is a singularity other than a pole of finite order.   For further details see N. W. McLaughlan, *op. cit.*, or E. T. Whittaker and G. N. Watson, "Modern Analysis," The Macmillan Company, New York, 1943.

The common terms in the numerator and the denominator (that is, $s - s_n$) are canceled prior to substituting $s = s_n$. From Table 23.4, it can be seen that

$$\mathcal{L}^{-1}\left(\frac{k_n}{s - s_n}\right) = k_n e^{s_n t} \qquad t > 0 \tag{23.37}$$

The complete inverse transform will then be

$$f(t) = \mathcal{L}^{-1}F(s) = k_1 e^{s_1 t} + \cdots + k_n e^{s_n t} \tag{23.38}$$

It can be seen that expressions such as Eq. (23.38) can lead to sinusoidally varying functions since

$$e^{j\omega t} + e^{-j\omega t} = 2 \cos \omega t \tag{23.39}$$

Higher-order Poles. In the event there are two or more identical poles associated with the response transform, the partial fraction expansion procedure outlined above is not sufficient for a complete solution. In this event, the following general relationship can be used:

Rewriting Eq. (23.34) for the general case including higher-order poles

$$F(s) = \frac{A(s)}{B(s)} = \frac{A(s)}{(s - s_1)^{m_1}(s - s_2)^{m_2} \cdots (s - s_h)^{m_h} \cdots (s - s_n)^{m_n}}$$

The number of coefficients associated with each exponent of $B(s)$ is equal to the value of the exponent. For a given exponent m_h, the coefficients are designated k_{hg} where g has as many values as the value of m_h. For example, if $m_h = 3$, there are three values of g; that is, $g = 1$, $g = 2$, and $g = 3$. The coefficients associated with m_h in this case would be designated k_{h1}, k_{h2}, and k_{h3}. These coefficients can be evaluated from

$$k_{hg} = \frac{1}{(g - 1)!}\left[\frac{d^{g-1}}{ds^{g-1}}\frac{(s - s_h)^{m_h}A(s)}{B(s)}\right]_{s=s_h} \tag{23.40a}$$

After determining the coefficients, $f(t)$ can be determined from Eq. (23.40b).

$$f(t) = \sum_{h=1}^{h}\sum_{g=1}^{m_h}\frac{k_{hg}}{(m_h - g)!}t^{m_h - g}e^{s_h t} \tag{23.40b}$$

In general, the process of obtaining the real function $f(t)$ from the complex function $F(s)$ may be accomplished mathematically by means of Eq. (23.41).

$$f(t) = \frac{1}{2\pi j}\int_{C-j\infty}^{C+j\infty} F(s)e^{st}\, ds \tag{23.41}$$

Equation (23.41) is a *contour integral.* The path of integration in the s plane is along a line generally parallel to the $j\omega$ axis displaced to the right of this axis the distance C. The value of C is chosen so that the Laplace integral will be convergent and so that all singularities of $F(s)$ will be to the left of C.

The use of these methods of obtaining inverse Laplace transforms will be illustrated by the following example.

Example 23.10

Determine the inverse transforms of the following functions.

1. $F(s) = \dfrac{(s + a)}{s(s + b)}$

From Eq. (23.35)

$$F(s) = \frac{k_1}{s} + \frac{k_2}{s + b}$$

From Eq. (23.36)

$$k_1 = \left[\frac{(s + a)s}{s(s + b)} \right]_{s=0} = \frac{a}{b}$$

$$k_2 = \left[\frac{(s + a)}{s(s + b)} (s + b) \right]_{s=-b} = \frac{-b + a}{-b} = 1 - \frac{a}{b}$$

Therefore,

$$F(s) = \frac{a/b}{s} + \frac{1 - a/b}{s + b}$$

and from Table 23.4

$$f(t) = \frac{a}{b} + \left(1 - \frac{a}{b} \right) e^{-bt}$$

2. $F(s) = \dfrac{1}{(s + a)^3}$; therefore $m_1 = 3$, $n = 1$, $h = 1$, $g = 1, 2, 3$.

From Eq. (23.40a)

$$\left.\begin{array}{l} h = 1 \\ g = 1 \end{array}\right\} k_{11} = \frac{1}{0!} [1] = 1$$

$$\left.\begin{array}{l} h = 1 \\ g = 2 \end{array}\right\} k_{12} = \frac{1}{1!} [0] = 0$$

$$\left.\begin{array}{l} h = 1 \\ g = 3 \end{array}\right\} k_{13} = \frac{1}{2!} [0] = 0$$

From Eq. (23.40b)

$$\begin{aligned}
f(t) &= \sum_{h=1}^{1} \sum_{g=1}^{3} \frac{k_{1g}}{(3 - g)!} t^{(3-g)} e^{-at} \\
&= \frac{k_{11}}{(3 - 1)!} t^{(3-1)} e^{-at} + \frac{k_{12}}{(3 - 2)!} t^{(3-2)} e^{-at} + \frac{k_{13}}{(3 - 3)!} t^{(3-3)} e^{-at} \\
&= \tfrac{1}{2} t^2 e^{-at}
\end{aligned}$$

23.6d. Graphical Determination of Inverse Transforms.[1] The use of Eqs. (23.34) to (23.41) are often time-consuming. The graphical procedure described in the following paragraphs is quick, simple, minimizes arithmetical errors, and provides a means of visualizing the effect of zero and pole locations on the time response.

Three steps are required in the graphical determination of inverse transforms:

1. Locate on the complex plane the zeros and poles of the response transform. (Complex-plane coordinates are values of α for abscissa and values of $j\omega$ for ordinate.)

2. Write the mathematical *form* of the inverse transform from an inspection of the pole locations on the complex plane.

3. Evaluate by graphical measurement the numerical values of the unknown constants to obtain a complete solution.

Location of Poles. In general, the response transform will be a rational function similar to Eq. (23.34). The zeros of this function are the roots of the numerator. If there are no zeros, the numerator will be equal to a constant. The *poles* of the response function are the roots of the denominator. If there are no poles, the denominator will be equal to unity. The roots of the numerator and the denominator can be obtained by standard algebraic methods[2] or by use of the graphical techniques described in Sec. 18.4e.

Once the zeros and poles have been determined, they are plotted on the complex plane. A typical case is shown in Fig. 23.15. The zero at α_2 is a negative real zero. The pole at α_1 is a negative real pole. The pole corresponding to the $1/s$ term is a pole at the origin. The pair of zeros $\pm j\omega_1$ are conjugate zeros on the imaginary axis.

[1] Adapted from unpublished papers of N. Patrusky.

[2] See, for example, R. E. Doherty and E. G. Keller, "Mathematical Methods in Engineering," vol. II, John Wiley & Sons, Inc., New York, 1942.

The pair of poles $-\alpha_3 \pm j\omega_2$ are conjugate complex poles. Multiple-order zeros or poles, not shown on the figure, occur when two or more zeros coincide or two or more poles coincide.

Mathematical Form of the Inverse Transform. Once the poles have been located on the complex plane, the mathematical *form* of the inverse transform can be written by inspection, i.e., from an examination of a table of transform pairs, for a majority of the cases encountered in practice. Presence or absence of zeros or the number of zeros does not affect the *form* of the solution—only the amplitudes of the constants.[1] The most useful forms are tabulated in Table 23.5 for convenience. The complex expressions in this table are for the roots of the *denominator*. In a case containing several poles, the corresponding complex expressions in the denominator are *multiplied*, and the corresponding inverse transform forms are added.

$$F(s) = \frac{(s+a_2)(s+j\omega_1)(s-j\omega_1)}{s(s+a_1)(s+a_3-j\omega_2)(s+a_3+j\omega_2)}$$

Fig. 23.15. Plot of zero and pole locations.

Graphical Evaluation of Constants. The unknown constants of the inverse transform (A's and ϕ's) are evaluated by making length and angle measurements on the complex plane plot. Each pole is considered separately, and when measurements are made from the other zeros and poles to the particular "pole of interest," the constants in the expression resulting from that pole are determined. When conjugate poles are the poles of interest, measurements are made only to the upper pole. For simplicity, the general subject of graphical analysis has been divided into four separate cases.

TABLE 23.5. PARTIAL TABLE OF INVERSE TRANSFORMS

Denominator Roots	Form of Inverse Transform
s	A
$(s + \alpha)$	$Ae^{-\alpha t}$
$(s + j\omega)(s - j\omega)$	$A \sin (\omega t + \phi)$
$(s + \alpha + j\omega)(s + \alpha - j\omega)$	$Ae^{-\alpha t} \sin (\omega t + \phi)$

CASE 1. POLES AND ZEROS ON THE NEGATIVE REAL AXIS—SINGLE-ORDER POLES ONLY. The amplitude constants are obtained as follows:

$$A = \frac{\text{product of distances to the pole of interest from each zero}}{\text{product of distances to the pole of interest from each other pole}} \quad (23.42)$$

The numerator is unity for the case of no zeros, and the denominator is unity for the case of a single pole. If a higher-order zero occurs on the negative real axis, the distance to the pole of interest from the higher-order zero is multiplied by the power of the order.

The algebraic sign associated with each length measurement is determined by considering each length measurement to have an associated angle. The angle for a

[1] An exception to this statement occurs when the number of zeros is equal to the number of poles. For this case, the numerator is divided by the denominator. The resulting quotient will contain a constant and a remainder, and it may contain terms involving powers of s. The inverse transform of the remainder is obtained by the methods of this section. The inverse transforms of the remaining factors are added to this result.

particular measurement is the angle between the real axis in the positive direction and the line *to* the pole of interest. For poles and zeros on the real axis, this angle is either $0°$ (+ sign) or $180°$ (− sign). If a higher-order zero occurs, the angle between the higher-order zero and the pole of interest should be multiplied by the power of the order to determine the proper sign. The application of Eq. (23.42) is illustrated by the following example.

Example 23.11

Determine the inverse transform of the following response functions:

$$F(s) = \frac{s+1}{s^2+2s} = \frac{s+1}{s(s+2)}$$

There is a zero at −1, and poles at 0, −2; these are plotted in Fig. 23.16.

The *form* of the inverse transform is written by inspection or from Table 23.5 as follows:

$$f(t) = A_o + A_1 e^{\alpha_1 t}$$

$\alpha_1 = -2$ from the $(s+2)$ term

A_o and A_1 are evaluated from Eq. (23.42) as follows:

To evaluate A_o, the pole of interest is the pole at the origin; lengths are measured from the zero at −1 and the pole at −2 toward the origin (all positive). Therefore,

$$A_o = \frac{+1}{+2} = \frac{1}{2}$$

In a similar manner,

$$A_1 = \frac{-1}{-2} = \frac{1}{2}$$

The final solution is, therefore,

$$f(t) = \tfrac{1}{2}(1 + e^{-2t})$$

Although this is a simple example for purposes of explanation, the same procedure applies for cases containing many poles which would be very time-consuming to solve by other methods. The following quantitative rules will be found to apply for poles on the real axis and are an aid in evaluating the physical effect of changing pole and zero locations.

Fig. 23.16. Plot for Example 23.11.

1. As poles approach each other, the amplitudes of the terms due to these poles will increase. Conversely, for poles far apart, the amplitudes of the terms due to these poles will be small.

2. As a zero approaches a pole, the amplitude of the term due to the pole decreases. Conversely, as a zero is separated from a pole, the amplitude of the term due to the pole increases.

CASE 2. REAL POLES AND CONJUGATE COMPLEX ZEROS—SINGLE-ORDER POLES ONLY. The rules and procedure for case 1 also apply for conjugate complex zeros. An additional rule is that lengths should be taken from *both zeros* of a conjugate complex pair to a pole of interest. The sum of the angles, measured between a line parallel to the real axis and the vector from each zero to the pole of interest, will always be $360°$. Consequently, the algebraic sign of the *product* of the two terms resulting from conjugate complex zeros will always be positive. This is shown in Fig. 23.17.

CASE 3. CONJUGATE COMPLEX POLES—SINGLE-ORDER POLES ONLY. When conjugate-complex poles $(s = -\alpha \pm j\omega)$ are present, sinusoidal terms result, and the phase

angle ϕ must be evaluated in addition to the amplitude constants. In addition, the amplitude factors are multiplied by $1/\omega$ where ω is the imaginary part of the particular complex pole of interest. These rules are given by Eqs. (23.43) and (23.44).

$$A = \frac{1}{\omega} \frac{\text{(product of distances to the pole of interest from each zero)}}{\text{(product of distances to the pole of interest from each pole)}} \quad (23.43)$$

where ω is the imaginary part of the complex pair being evaluated.

The numerator is unity for the case of no zeros. When the constants resulting from a complex pair are being evaluated, only one of the pair is selected as the pole of interest. For the positive sign associated with ϕ in Table 23.5 the upper conjugate pole should be selected as the pole of interest. Length and angles are measured from both complex poles of *other* pairs which may be present.

$\phi = \Sigma$ angles to pole of interest from zeros $- \Sigma$ angles to pole of interest from other
poles (excluding the other conjugate member)   (23.44)

Angles are measured as positive in rotating counterclockwise from a line through the pole or zero parallel to the real axis. The sum of the angles of the vectors from the two poles of a complex pair to another pole of interest which is on the negative real axis is always 360°, and consequently the product of these two vectors is always positive.

FIG. 23.17. Complex zeros.

$$F(s) = \frac{(s+a_4)(s+a_2-j\omega_2)(s+a_2+j\omega_2)}{(s+a_3)(s+j\omega_1)(s-j\omega_1)}$$
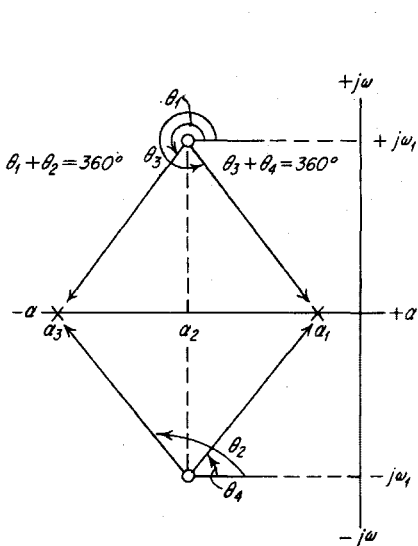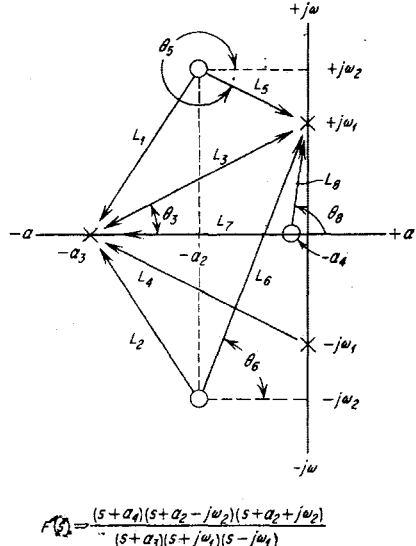
FIG. 23.18. Pole and zero locations for Example 23.12.

Example 23.12

Determine the inverse transform of the transfer function shown in Fig. 23.18.

Solution

The form of the inverse transform can be written by the use of Table 23.5 and an examination of Fig. 23.18 as follows:

$$f(t) = A_1 e^{-\alpha_3 t} + A_2 \sin(\omega_1 t + \phi)$$

A_1 is evaluated from Eq. (23.42) with $-\alpha_3$ as the pole of interest.

$$A_1 = \frac{(-L_7)(L_2)(L_1)}{(L_3)(L_4)}$$

A_2 is evaluated from Eq. (23.43) with the upper conjugate pole selected as the pole of interest.

$$A_2 = \frac{1}{\omega} \frac{(L_5)(L_6)(L_8)}{(L_3)}$$

Finally, ϕ is evaluated from Eq. (23.44) as follows:

$$\phi = \theta_5 + \theta_6 + \theta_8 - \theta_3$$

CASE 4. MULTIPLE-ORDER POLES. When multiple-order poles are encountered in a transfer function, they must be handled in a slightly different manner. The procedure is as follows:

1. Locate on the complex plane the zeros and poles of the response transform as before.

2. Evaluate the constants for each pole of interest, except the coincident poles. This will square the distances and double the angles from the coincident pole position to the pole of interest for a second-order pole.

3. Separate coincident poles along the real axis by a distance δ (move one pole, or set of poles in the case of complex poles, to the left a distance δ).

4. Write the expressions for the inverse transform of the separated poles from Table 23.5, treating each pole as a single-order pole.

5. Evaluate the amplitude and phase constants for each of the separated poles, including the δ factor wherever it appears.

6. Let δ approach zero to evaluate the expressions obtained in step 5 for the amplitude and phase constants. The remainder in each case is the required value. NOTE: To take the limit of indeterminate forms, that is, $0/0$, ∞/∞, differentiate the numerator and the denominator with respect to δ and then let δ approach zero. If the result is still indeterminate, differentiate as many times as are required, letting δ approach zero after each differentiation until a finite result (or zero) is obtained.

Second or higher-order terms in δ may be removed before taking the limit.

The final expression so obtained should be of the form shown in Table 23.6 for second-order poles. For higher-order poles, terms similar to the basic form for a single-order pole, but containing t^2, t^3, . . . , t^{m-1} are added. If m is the order of the pole, $m - 1$ is the highest power of t and there are m terms in the expression.

TABLE 23.6. INVERSE TRANSFORMS FOR SECOND-ORDER POLES

$F(s)$	$f(t)$
s^2	$A_o + A_1 t$
$(s + \alpha)^2$	$A_o e^{-\alpha t} + A_1 t e^{-\alpha t}$
$(s^2 + \omega^2)^2$	$A_o \sin(\omega t + \phi_o) + A_1 t \sin(\omega t + \phi_1)$
$(s + \alpha + j\omega)^2 (s + \alpha - j\omega)^2$	$A_o \sin(\omega t + \phi_o)e^{-\alpha t} + A_1 t \sin(\omega t + \phi_1)e^{-\alpha t}$

As an example, the form of the inverse transform for a third-order pole at $-\alpha_1$ can be written directly from the above discussion as follows:

$$f(t) = A_o e^{-\alpha_1 t} + A_1 t e^{-\alpha_1 t} + A_2 t^2 e^{-\alpha_1 t} \tag{23.45}$$

It will often save effort to shift the imaginary axis to the coincident poles before separating if only one set of coincident poles is present. The inverse transform can then be corrected by multiplying each term by $e^{-\alpha t}$, where α is the distance moved along the negative real axis.

Example 23.13

Obtain the inverse transform of the function shown in Fig. 23.19a.

Solution

1. The zeros and poles are located in Fig. 23.19a.
2. The coincident poles are shown separated in Fig. 23.19b.

(a) POLE AND ZERO LOCATION   (b) SEPARATED POLE LOCATION
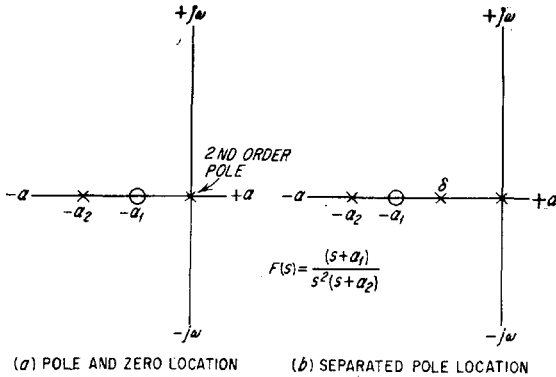
FIG. 23.19. Plot of Example 23.13 with separated poles.

3. From Table 23.5 the expression for this inverse transform is as follows for the separated poles:

$$f(t) = A_o + A_1 e^{-\delta t} + A_2 e^{-\alpha_2 t}$$

4. The pole at $-\alpha_2$ is the only noncoincident pole. The constant A_2 due to this pole is evaluated from Eq. (23.42).

$$A_2 = -\frac{\alpha_2 - \alpha_1}{(-\alpha_2)^2} = -\frac{\alpha_2 - \alpha_1}{\alpha_2^2}$$

5. The expressions for the amplitude constants resulting from the separated poles are as follows:

$$A_o = \frac{\alpha_1}{\delta \alpha_2} \qquad A_1 = \frac{\alpha_1 - \delta}{(-\delta)(\alpha_2 - \delta)}$$

6. The complete expression is, therefore,

$$f(t) = -\frac{(\alpha_2 - \alpha_1)}{\alpha_2^2} e^{-\alpha_2 t} + \frac{\alpha_1}{\delta \alpha_2} + \left[\frac{\alpha_1 - \delta}{(-\delta)(\alpha_2 - \delta)}\right] e^{-\delta t}$$

Combining terms containing δ,

$$\frac{\alpha_1}{\delta \alpha_2} + \frac{(\alpha_1 - \delta)}{(\delta)(\delta - \alpha_2)} e^{-\delta t} = \frac{\alpha_1(\delta - \alpha_2) + \alpha_2(\alpha_1 - \delta)e^{-\delta t}}{\delta \alpha_2(\delta - \alpha_2)}$$

$$= \frac{\delta \alpha_1 - \alpha_1 \alpha_2 + \alpha_2 \alpha_1 e^{-\delta t} - \delta \alpha_2 e^{-\delta t}}{\delta^2 \alpha_2 - \delta \alpha_2^2}$$

Eliminating the δ^2 term and taking the limit

$$\lim_{\delta \to 0} \left[\frac{\delta \alpha_1 - \alpha_1 \alpha_2 + \alpha_1 \alpha_2 e^{-\delta t} - \delta \alpha_2 e^{-\delta t}}{-\delta \alpha_2^2}\right] = \frac{0}{0} \quad \text{and is indeterminant}$$

Differentiating the numerator and the denominator separately with respect to δ and taking the limit again,

$$\lim_{\delta \to 0} \left[\frac{\alpha_1 - t\alpha_1 \alpha_2 e^{-\delta t} - \alpha_2 e^{-\delta t} + \delta t \alpha_2 e^{-\delta t}}{-\alpha_2^2}\right] = \frac{\alpha_1 - t\alpha_1 \alpha_2 - \alpha_2}{-\alpha_2^2}$$

The final result is therefore

$$f(t) = \frac{(\alpha_2 - \alpha_1) + \alpha_1 \alpha_2 t + (\alpha_1 - \alpha_2)e^{-\alpha_2 t}}{\alpha_2^2}$$

Example 23.24

Obtain the inverse transform of the following function:

$$F(s) = \frac{s + \alpha_1}{[(s + \alpha_1)^2 + \omega_1^2]^2}$$

The poles and zero are shown in Fig. 23.20a. The poles and zero are shown shifted in Fig. 23.20b. The shifted poles are shown separated an amount δ in Fig. 23.20c. The

(a) POLES AND ZERO OF $F(s)$ (b) SHIFTED POLES AND ZERO   (c) SHIFTED AND SEPARATED POLES AND ZERO
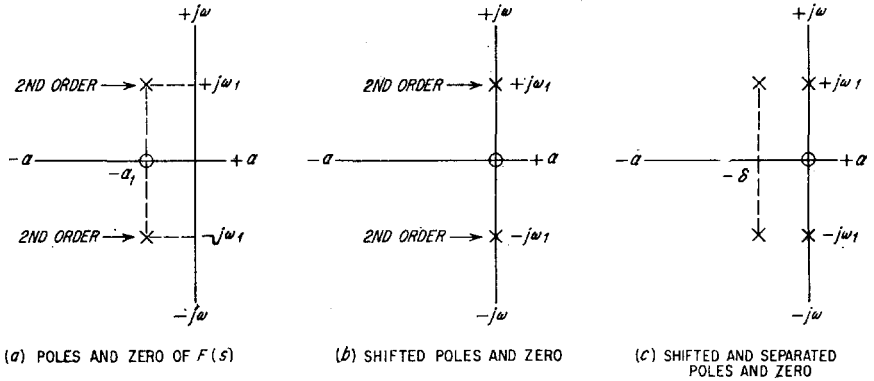
FIG. 23.20. Plot of Example 23.14 with poles shifted and separated.

inverse transform of the shifted separated poles is, from Table 23.5,

$$f(t) = [A_1 \sin (\omega_1 t + \phi_1) + A_2 e^{-\delta t} \sin (\omega_1 t + \phi_2)]e^{-\alpha_1 t}$$

The constants A_1, A_2 are from Eq. (23.43).

$$A_1 = \frac{\omega_1}{\omega_1 \delta (\delta^2 + 4\omega_1^2)^{1/2}} = \frac{1}{\delta(\delta^2 + 4\omega_1^2)^{1/2}}$$

$$A_2 = \frac{(\delta^2 + \omega_1^2)^{1/2}}{\omega_1 \delta (\delta^2 + 4\omega_1^2)^{1/2}}$$

The constants ϕ_1 and ϕ_2 are from Eq. (23.44).

$$\phi_1 = +90° - \tan^{-1}\frac{2\omega_1}{\delta} - 0$$

$$\lim_{\delta \to 0} \phi_1 = +90° - 90° - 0° = 0$$

$$\phi_2 = -\tan^{-1}\frac{\omega_1}{\delta} + \tan^{-1}\frac{2\omega_1}{\delta} - 180°$$

$$\lim_{\delta \to 0} \phi_2 = -180°$$

The expression for $f(t)$ becomes

$$f(t) = \left[\frac{\sin \omega_1 t}{\delta(\delta^2 + 4\omega_1^2)^{1/2}} + \frac{(\delta^2 + \omega_1^2)^{1/2}e^{-\delta t} \sin (\omega_1 t - 180°)}{\omega_1 \delta (\delta^2 + 4\omega_1^2)^{1/2}} \right] e^{-\alpha_1 t}$$

Combining the terms within the bracket over a common denominator and simplifying,

$$f(t) = \left[\frac{\omega_1 \sin \omega_1 t - (\delta^2 + \omega_1^2)^{1/2}e^{-\delta t} \sin \omega_1 t}{\omega_1 \delta (\delta^2 + 4\omega_1^2)^{1/2}} \right] e^{-\alpha_1 t}$$

Since this expression approaches zero over zero as δ approaches zero, it is necessary to differentiate the numerator and denominator with respect to δ. The result, after elimina-

tion of the δ^2 terms and differentiation is

$$f(t) = \lim_{\delta \to 0} \frac{\omega_1 t \sin \omega_1 t}{2\omega_1{}^2} e^{-\alpha_1 t} = \frac{t e^{-\alpha_1 t} \sin \omega_1 t}{2\omega_1}$$

23.6e. Relationship between Pole Location and $f(t)$. From Secs. 23.6c and 23.6d it can be seen that the location of the poles of the system transfer function on the complex plane determines the general time response of the output. Table 23.7 gives a summary of the correspondence between the location of poles on the complex $(\alpha, j\omega)$ plane and the response of the network as a function of the real variable, time.

TABLE 23.7. RELATIONSHIP BETWEEN POLE LOCATION AND $f(t)$

Location of Pole in the Complex Plane	*Corresponding $f(t)$ for $t > 0$*
1. A first-order pole at the origin	1. A constant
2. A first-order pole on the negative real axis	2. A decreasing exponential
3. A first-order pole on the positive real axis	3. An increasing exponential
4. First-order conjugate poles on the imaginary axis	4. Sinusoidal oscillation of constant amplitude whose frequency increases as the separation between the conjugate poles increases
5. First-order conjugate poles in the left half of the complex plane	5. Exponentially decreasing oscillations
6. First-order conjugate poles in the right half of the complex plane	6. Exponentially increasing oscillations
7. A second- or higher-order pole at the origin	7. A linear or higher-order polynomial
8. A second- or higher-order pole on the negative real axis	8. The product of a linear or higher-order polynomial and a decreasing exponential
9. A second- or higher-order pole on the positive real axis	9. The product of a linear or higher-order polynomial and an increasing exponential

23.6f. Summary of the Procedure for the Solution of Networks by Transform Methods. To summarize the use of the Laplace transform in the solution of network problems, the following outline contains the general steps necessary for the complete solution of lumped constant network problems.

1. *Describe the operation of the network by means of the applicable loop or nodal equations.*

2. *Transform the network equations to complex form.* This step consists of obtaining the transform of the applied signals from Table 23.4 or Eq. (23.30).

3. *Solve the resulting equations for the required unknown.* Solve by determinants or other algebraic means the expression for the required unknown voltage, current, impedance, etc. The resulting expression for the unknown quantity will be the response transform discussed previously.

4. *Examine the poles of the response transform to determine the nature of the response of the network as a function of time.* This step requires determination of the roots of the denominator of the response transform and the use of Tables 23.4 to 23.7.

5. *For a complete quantitative solution obtain the inverse transform of the response function.* This can be accomplished for simple cases directly from Table 23.4 and for other cases by means of partial fraction expansion, or the graphical procedure of Sec. 23.6d.

The functions $A(s)/B(s)$, obtained by solving the loop or nodal equations after transformation to complex form, contain poles for each root of $B(s)$. If there are many reactive elements in the system, the degree of this polynomial can be very high. To solve equations of fourth degree or higher, either Graeff's root-squaring method[1] or the root-locus method discussed in Sec. 18.4e is recommended.
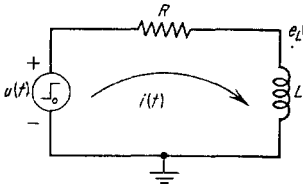
Example 23.15

Determine the voltage across the inductance in Fig. 23.21 by both the loop and nodal analysis for a step-function input of unit amplitude.

1. *Loop Analysis*
   a. The number of loops is 1, from Eq. (23.19),

$$l = e - n + s = 3 - 3 + 1 = 1$$

and the number of unknowns, from Eq. (23.18),

$$u_i = l - i_s = 1 - 0 = 1$$

FIG. 23.21. Network with step excitation, for Example 23.15.

b. Writing the loop equation:

$$Ri(t) + L\frac{di(t)}{dt} = u(t)$$

c. Change to complex variable form. Let the complex form of $i(t)$ be $I(s)$; then, from Table 23.4, the transform of $\left[\dfrac{di(t)}{dt}\right]$ is $sI(s) - (0)$, assuming zero initial current. The transform of $u(t)$ is $1/s$ and

$$RI(s) + LsI(s) = \frac{1}{s}$$

d. Solving this algebraically for $E_L(s)$

$$E_L(s) = LsI(s) = Ls\left[\frac{1}{sL + R}\cdot\frac{1}{s}\right] = \frac{L}{sL + R} = \frac{1}{s + R/L} = \frac{A(s)}{B(s)}$$

In this example, $1/s$ is the excitation function, and $sL/(sL + R)$ is the system function. These can be combined into the form $A(s)/B(s)$, where $A(s) = 1$ and $B(s) = \dfrac{1}{(s + R/L)}$.

From Table 23.7 it can be determined that the form of the response current will be a decreasing exponential. To complete the solution, the inverse transform is taken from Table 23.4.

$$\mathcal{L}^{-1}\left[\frac{A(s)}{B(s)}\right] = e^{-Rt/L}$$

2. *Nodal Analysis*
In a similar manner, from Eqs. (23.24) and (23.25),
a. $n_s = n - s = 3 - 1 = 2$
   $u_v = n_s - e_s = 2 - 1 = 1$
b. Writing the nodal equation.

$$i(t) = \frac{u(t) - e_L(t)}{R} = \frac{1}{L}\int e_L(t)\,dt$$

c. Changing to the complex notation,

$$\frac{1/s - E_L(s)}{R} = \frac{1}{L}\cdot\left[\frac{E_L(s)}{s} + (0)\right]$$

d. Solving for $E_L(s)$

$$E_L(s) = \frac{1}{(s + R/L)} = \frac{A(s)}{B(s)}$$

which is the same result obtained in part 1.

[1] Doherty and Keller, *op. cit.*

23.7. Thévenin's Theorem. This theorem is useful in the solution of network problems where it is necessary to reduce a complex network containing many elements to an equivalent network consisting of a constant-voltage generator in series with a single impedance.

Theorem. Any network containing voltage and/or current generators and linear bilateral impedances can be replaced, with respect to any two external terminals of the network, by an equivalent network. The equivalent network consists of a constant-voltage generator with zero internal impedance whose generated voltage is the open-circuit voltage e_{oc} measured between the terminals of the original network and a series impedance Z_G whose value is the impedance measured between the terminals

(a) GENERAL THÉVENIN NETWORK

$$i = \frac{e_{oc}}{Z_R + Z_G}$$

(b) DC NETWORK EQUIVALENT

$$e_{oc} = 6 \times \frac{5}{6} = 5 \ VOLTS$$

$$Z_G = \frac{1 \times 5}{1+5} = \frac{5}{6} OHM$$

(c) VACUUM TUBE AMPLIFIER EQUIVALENT

$$e_{oc} = \frac{-\mu e_g R_L}{R_L + r_p}$$

$$Z_G = \frac{R_L \ r_p}{R_L + r_p}$$
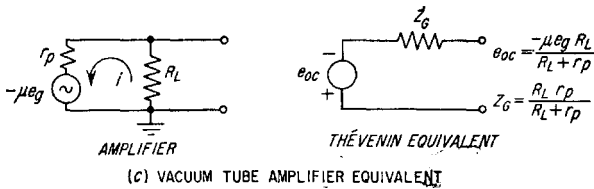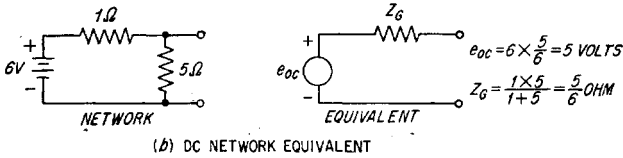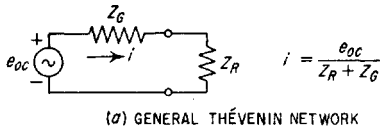
FIG. 23.22. Thévenin equivalents.

of the original network when all the generators in the network are replaced by their internal impedances.

Figure 23.22 illustrates the general case and some examples of this theorem.

23.8. Norton's Theorem. This theorem is useful in the solution of network problems where it is necessary to reduce a complex network containing many elements to an equivalent network consisting of a constant-current generator shunted by a single impedance.

Theorem. Any network containing voltage and/or current generators and linear bilateral impedances can be replaced, with respect to any two external terminals of the network, by an equivalent network. This equivalent network consists of a constant-current generator with infinite internal impedance, whose generated current is equal to the short-circuit current i_{sc} measured between the terminals of the original network, in shunt with the impedance Z_G (or admittance Y_G) seen looking back into the original network from the terminals when all generators are replaced by their internal impedances.

Figure 23.23 illustrates the general case and some examples of this theorem.
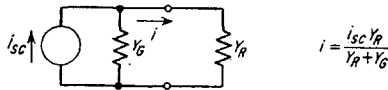
(a) GENERAL NORTON NETWORK

$$i = \frac{i_{sc}\,Y_R}{Y_R + Y_G}$$

(b) D-C NETWORK EQUIVALENT

$$i_{sc} = 6 \times \frac{5}{5+1} = 5 \text{ AMPERES}$$

$$Z_G = 5 + 1 = 6 \text{ OHMS}$$

(c) VACUUM TUBE AMPLIFIER EQUIVALENT

$$i_{sc} = g_m e_s$$

$$Z_G = \frac{R_L\, r_p}{R_L + r_p}$$
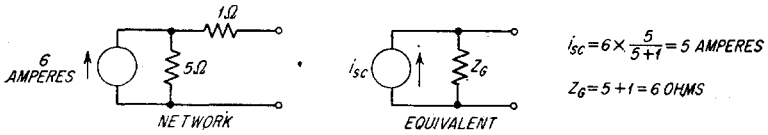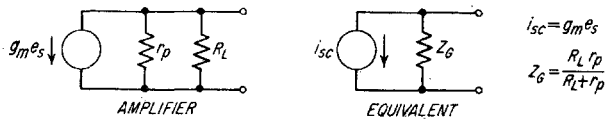
FIG. 23.23. Norton equivalents.

23.9. Superposition Theorem. This theorem is the basis for much network analysis theory. It enables a network, composed of many branches having linear, bilateral elements and containing many generators, to be broken down into a series of simpler networks, each containing only one generator. The voltages and/or currents at the required points in the general network may then be found by adding the voltages and/or currents, determined by the single-generator networks. The generators may be at different frequencies, making the theorem valid where complex waveforms are involved. If transients are important, initial stored energy in capacitors or inductors may be accounted for by adding the required initial voltage or current generators. The theorem is especially useful where voltage or current sources are added to an existing network.

Theorem. In a network composed of linear, bilateral impedances, the current in any branch or the potential difference between any two nodes due to any number of voltage or current generators distributed in any manner throughout the network is the sum of the currents or voltages which would be produced by the individual generators acting alone with all other generators replaced by their internal impedances.

Example 23.16

1. Determine e_o as shown in Fig. 23.24. This network contains two generators and has been divided into two separate networks, each containing only one generator. The voltages
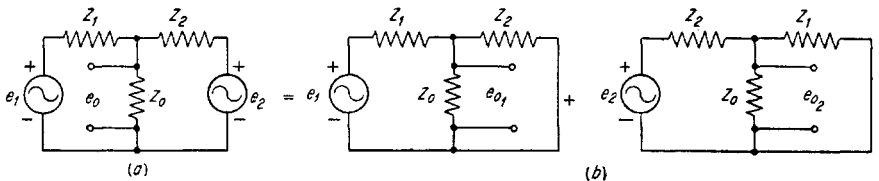
FIG. 23.24. Network for Example 23.16.

across Z_0 for the case of the divided networks, e_{01} and e_{02}, can be determined separately, and then added by means of the superposition theorem (see Fig. 23.24b).

$$e_{o1} = \frac{e_1 \left(\dfrac{Z_o Z_2}{Z_o + Z_2} \right)}{Z_1 + \dfrac{Z_o Z_2}{Z_o + Z_2}}$$

$$e_{o2} = \frac{e_2 \left(\dfrac{Z_o Z_1}{Z_o + Z_1} \right)}{Z_2 + \dfrac{Z_o Z_1}{Z_o + Z_1}}$$

By superposition, $e_o = e_{o1} + e_{o2}$

$$e_o = Z_o \frac{e_1 Z_2 + e_2 Z_1}{Z_1 Z_o + Z_1 Z_2 + Z_o Z_2}$$

Example 23.17

If a third generator e_3 is now added in series with Z_o, as shown in Fig. 23.25, determine the change in e_o.

Solution:

The change in e_o can be quickly obtained by computing e_{o3} across Z_o due to e_3 without regard to e_1 and e_2 as follows:

$$e_{o3} = -e_3 \frac{(Z_1 + Z_2) Z_o}{Z_1 Z_o + Z_1 Z_2 + Z_2 Z_o}$$

This result was obtained considerably faster than would have been the case if the entire network were recalculated and the difference obtained between the two results.
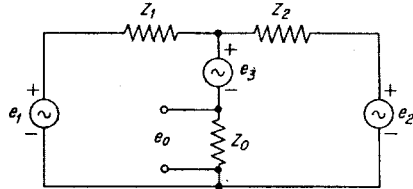
FIG. 23.25. Network for Example 23.17.

23.10. Reciprocity Theorem. This theorem permits the interchange in position of a generator and any resulting current or voltage in another part of the network for purposes of analysis. The theorem applies for both transient and steady-state generators in linear bilateral networks.

Theorem. If the output of a generator produces a current through an element located at any point in a network composed of linear, bilateral elements, the same current will be produced if the positions of the generator and the current measuring point are interchanged.

Only a single generator is considered at one time. Other voltages and currents in the network will not necessarily remain the same.

Example 23.18

This theorem will be demonstrated by means of the network shown in Fig. 23.26.

1. With the generator and ammeter as shown in the figure, the current i_A is determined as follows:

$$i_A = \frac{e}{Z_1 + \dfrac{Z_2 Z_3}{Z_2 + Z_3}} \frac{Z_3}{Z_2 + Z_3}$$
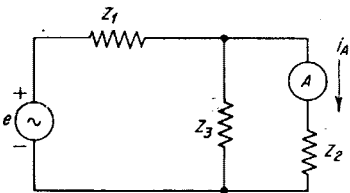
$$= \frac{e Z_3}{Z_1 Z_2 + Z_2 Z_3 + Z_1 Z_3}$$

FIG. 23.26. Network illustrating reciprocity.

2. Now, reversing the positions of the generator and ammeter, as shown in Fig. 23.27, the current through the ammeter is again determined.

$$i_A' = \frac{e}{Z_2 + \dfrac{Z_1 Z_3}{Z_1 + Z_3}} \frac{Z_3}{Z_1 + Z_3}$$

$$= \frac{e Z_3}{Z_1 Z_2 + Z_1 Z_3 + Z_2 Z_3}$$

which is the same as in step 1.

23.11. Compensation Theorem. This theorem is useful in calculating the effect of impedance changes in a network on the currents and/or voltages in other parts of

FIG. 23.27. Network for Example 23.18 with generator and ammeter reversed.

the network. The theorem applies for networks containing bilateral or unilateral, linear or nonlinear impedances.

Theorem. If an impedance in a network changes by an amount ΔZ, the change in current in any other branch of the network is equal to the current which would be produced in that branch by replacing all generators in the circuit by their internal impedances and adding in series with the modified impedance a generator with terminal voltage equal to $-i\,\Delta Z$, where i is the original current in the modified impedance.

Example 23.19

In the network shown in Fig. 23.28a, determine the effect on the current through Z_3 when Z_1 is doubled.

(a) BASIC NETWORK   (b) NETWORK WITH COMPENSATING GENERATOR
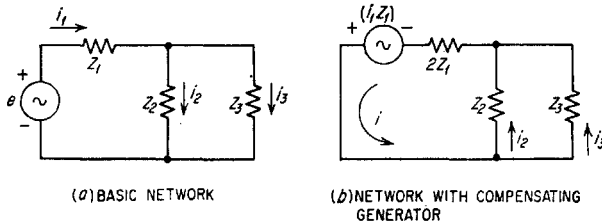
FIG. 23.28. Network illustrating the compensation theorem.

Solution

Before Z_1 is changed, the current through Z_3 is given by

$$i_3 = \frac{e}{Z_1 + Z_2Z_3/(Z_2 + Z_3)} \frac{Z_2}{Z_2 + Z_3} = \frac{eZ_2}{Z_1Z_2 + Z_1Z_3 + Z_2Z_3}$$

From the theorem, the increase in current in any other branch due to increasing Z_1 by an amount ΔZ is equivalent to replacing the existing generators with their internal impedances and adding in series with $Z_1 + \Delta Z_1$ a generator with a terminal voltage equal to the negative product of the original current (i_1) and the change in impedance (ΔZ_1). From this the modified circuit shown in Fig. 23.28b can be drawn to compute the *change* in i_3. This change is seen to be given by

$$\Delta i_3 = \frac{-i_1Z_1}{2Z_1 + Z_2Z_3/(Z_2 + Z_3)} \frac{Z_2}{Z_2 + Z_3} = \frac{-i_1Z_1Z_2}{2Z_1Z_2 + 2Z_1Z_3 + Z_2Z_3}$$

Where the original current

$$i_1 = \frac{e}{Z_1 + Z_2Z_3/(Z_2 + Z_3)}$$

$$\Delta i_3 = -\frac{Z_1Z_2}{2Z_1Z_2 + 2Z_1Z_3 + Z_2Z_3} \frac{e(Z_2 + Z_3)}{Z_1Z_2 + Z_1Z_3 + Z_2Z_3}$$

The new value i_3' is

$$i_3' = i_3 + \Delta i_3 = \frac{eZ_2}{Z_1Z_2 + Z_1Z_3 + Z_2Z_3}\left[1 - \frac{Z_1(Z_2 + Z_3)}{2Z_1Z_2 + 2Z_1Z_3 + Z_2Z_3}\right]$$

The last factor gives the reduction in i_3 due to the increase in Z_1.

23.12. Driving Point and Transfer Impedances and Admittances. Many of the characteristics of networks are described if the driving point and transfer impedances

or admittances are known. The following relationships can be used to determine these quantities in any given network and their effect on the voltages and currents of the network.

23.12a. Definitions of Impedances and Admittances. The loop and nodal equations[1] for a given network may be written as follows:

$$z_{11}i_1 + z_{12}i_2 + \cdots + z_{1n}i_n = e_1$$
$$z_{21}i_1 + z_{22}i_2 + \cdots + z_{2n}i_n = e_2$$

$$\phantom{z_{11}i_1 + z_{12}i_2 + \cdots +}$$

$$z_{n1}i_1 + z_{n2}i_2 + \cdots + z_{nn}i_n = e_n \tag{23.46}$$

and

$$y_{11}e_1 - y_{12}e_2 - \cdots - y_{1n}e_n = i_1$$
$$-y_{21}e_1 + y_{22}e_2 - \cdots - y_{2n}e_n = i_2$$

$$\phantom{y_{11}e_1 - y_{12}e_2 - \cdots -}$$

$$-y_{n1}e_1 - y_{n2}e_2 - \cdots + y_{nn}e_n = i_n \tag{23.47}$$

The factors $z_{11}, z_{22}, \ldots, z_{nn}$ are the *self-impedances* or *mesh impedances* of the network. Each is defined as the impedance around its closed loop or mesh with all other loops open-circuited. The factors $z_{12}, z_{23}, \ldots, z_{mn}$ are the *mutual impedances* of the network. Each is defined as the impedance common to loops m and n.

The factors $y_{11}, y_{22}, \ldots, y_{nn}$ are the *self-admittances* of the network, defined as the total admittance between the subject node and the reference node when all the other nodes of the network are connected to the reference node. The factors $y_{12}, y_{23}, \ldots, y_{mn}$ are the *mutual admittances* of the network, defined as the common admittances between nodes m and n.

By superposition the total current in a given loop is the sum of the currents due to generators in the loop plus the currents in the loop due to generators in other loops of the network. For example, the general solution of the loop equations can be expressed in the following form:

$$i_1 = \frac{\Delta_{11}}{\Delta}e_1 + \frac{\Delta_{21}}{\Delta}e_2 + \cdots + \frac{\Delta_{n1}}{\Delta}e_n$$
$$i_2 = \frac{\Delta_{12}}{\Delta}e_1 + \frac{\Delta_{22}}{\Delta}e_2 + \cdots + \frac{\Delta_{n2}}{\Delta}e_n$$

$$\phantom{i_1 = \frac{\Delta_{11}}{\Delta}e_1}$$

$$i_n = \frac{\Delta_{1n}}{\Delta}e_1 + \frac{\Delta_{2n}}{\Delta}e_2 + \cdots + \frac{\Delta_{nn}}{\Delta}e_n \tag{23.48}$$

where Δ is the impedance determinant of the network formed from Eq. (23.46) as follows:

$$\Delta = \begin{vmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ z_{n1} & z_{n2} & \cdots & z_{nn} \end{vmatrix} \tag{23.49}$$

[1] See Secs. 23.4 and 23.5.

and Δ_{jk} is formed by deleting the jth row and kth column from Δ. The algebraic sign of Δ_{jk} is $(-1)^{j+k}$.

From Eq. (23.48) the terms $(\Delta_{11}/\Delta)e_1$, $(\Delta_{22}/\Delta)e_2$, . . . , $(\Delta_{nn}/\Delta)e_n$ give the components of loop current in loops 1, 2, . . . , n, respectively, caused by a generator $e_1, e_2, \ldots, e_n$, located in the same loop. Further, the other terms, $(\Delta_{jk}/\Delta)e_j$ $(j \neq k)$, give the components of current in loop k caused by generators located in loop j.

In a similar manner, the general solution of Eq. (23.47) by Cramer's rule is

$$e_1 = i_1 \frac{\Delta'_{11}}{\Delta'} + i_2 \frac{\Delta'_{21}}{\Delta'} + \cdots + i_n \frac{\Delta'_{n1}}{\Delta'}$$

$$e_2 = i_1 \frac{\Delta'_{12}}{\Delta'} + i_2 \frac{\Delta'_{22}}{\Delta'} + \cdots + i_n \frac{\Delta'_{n2}}{\Delta'}$$

$$\ (23.50)$$

$$e_n = i_1 \frac{\Delta'_{n1}}{\Delta'} + i_2 \frac{\Delta'_{n2}}{\Delta'} + \cdots + i_n \frac{\Delta'_{nn}}{\Delta'}$$

where Δ' is the *admittance determinant* of the network formed from Eq. (23.47) as follows:

$$\Delta' = \begin{vmatrix} y_{11} & -y_{12} & \cdots & -y_{1n} \\ -y_{21} & y_{22} & \cdots & -y_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ -y_{n1} & -y_{n2} & \cdots & y_{nn} \end{vmatrix} \qquad (23.51)$$

and Δ'_{jk} is formed by deleting the jth row and kth column from Δ'. The algebraic sign of Δ'_{jk} is $(-1)^{j+k}$.

From Eq. (23.50) the terms $(\Delta'_{11}/\Delta')i_1$, $(\Delta'_{22}/\Delta')i_2$, . . . , $(\Delta'_{nn}/\Delta')i_n$ give the components of node voltage e_n at node n because of the current flowing into node n. Further, the other terms $(\Delta'_{jk}/\Delta')i_k$ $(j \neq k)$ give the component of node voltage at node j caused by generators located at node k.

23.12b. Driving-point Impedance. The *driving-point* (input) *impedance* of loop j is given by Eq. (23.52):

$$Z_{jj} = \frac{\Delta}{\Delta_{jj}} = \frac{e_j}{i_j} \qquad (23.52)$$

The driving-point impedance of loop j is the ratio of the voltage applied to loop j to the resultant current flowing around loop j with all other loops normally connected and all other generators replaced by their internal impedances. It should be noted that this definition is correct only when the source voltage e_j is applied in a branch of loop j which is common to no other loop of the network.

23.12c. Transfer Impedance. The *transfer impedance* between loops j and k is given by Eq. (23.53).

$$Z_{jk} = \frac{\Delta}{\Delta_{jk}} = \frac{e_j}{i_k} \qquad (23.53)$$

The transfer impedance between loops j and k is the ratio of the voltage applied to loop j to the resulting current in loop k with all other loops normally connected and all other generators replaced by their internal impedances. It should be noted that this alternate definition is correct only when e_j is applied in a branch of loop j which is common to no other loop of the network.

23.12d. Driving-point Admittance. The *driving-point admittance* of node j is given by Eq. (23.54).

$$Y_{jj} = \frac{\Delta'}{\Delta'_{jj}} = \frac{i_j}{e_j} \qquad (23.54)$$

Y_{jj} may be defined in an alternate manner as follows: The driving-point admittance between node j and the reference node is the ratio between the driving current i_j entering node j, and the resulting voltage e_j between node j and the reference node.

23.12e. Transfer Admittance. The *transfer admittance* between nodes j and k is given by Eq. (23.55).

$$Y_{jk} = \frac{\Delta'}{\Delta'_{jk}} = \frac{i_k}{e_j} \qquad (23.55)$$

Y_{jk} may be defined in an alternate manner as follows: The transfer admittance between nodes j and k is the ratio between the current i_k applied at node k to the resultant voltage e_j appearing at node j.

23.12f. Relationships between Functions. The following relationships exist between the quantities defined above:

$$z_{jj} = Z_{jj} - \frac{\underset{jj}{\Delta}}{\Delta_{jj}} \qquad (23.56)$$

$$z_{jk} = Z_{jk} - \frac{\underset{jk}{\Delta}}{\Delta_{jk}} \qquad (23.57)$$

$$Z_{jj} = \frac{\underset{jj}{\Delta}}{\Delta_{jj}} + z_{jj} \qquad (23.58)$$

$$Z_{jk} = \frac{\underset{jk}{\Delta}}{\Delta_{jk}} + z_{jk} \qquad (23.59)$$

$$y_{jj} = Y_{jj} - \frac{\underset{jj}{\Delta'}}{\Delta'_{jj}} \qquad (23.60)$$

$$y_{jk} = Y_{jk} - \frac{\underset{jk}{\Delta'}}{\Delta'_{jk}} \qquad (23.61)$$

$$Y_{jj} = \frac{\underset{jj}{\Delta'}}{\Delta'_{jj}} + y_{jj} \qquad (23.62)$$

$$Y_{jk} = \frac{\underset{jk}{\Delta}}{\Delta'_{jk}} + y_{jk} \qquad (23.63)$$

where $\underset{jj}{\Delta} = \Delta$ with $z_{jj} = 0$; $\underset{jj}{\Delta'} = \Delta'$ with $y_{jj} = 0$
      $\underset{jk}{\Delta} = \Delta$ with $z_{jk} = 0$; $\underset{jk}{\Delta'} = \Delta'$ with $y_{jk} = 0$

Example 23.20

Determine the following impedances for the network shown in Fig. 23.29.

$$Z_{11} = \frac{\Delta}{\Delta_{11}} = z_{11} + \frac{\underset{11}{\Delta}}{\Delta_{11}}$$

$$Z_{12} = \frac{\Delta}{\Delta_{12}} = z_{12} + \frac{\underset{12}{\Delta}}{\Delta_{12}}$$
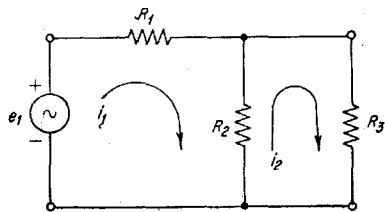
FIG. 23.29. Network for Example 23.20.

From inspection of the network:

$$\Delta = \begin{vmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{vmatrix} = \begin{vmatrix} R_1 + R_2 & -R_2 \\ -R_2 & R_2 + R_3 \end{vmatrix}$$

$$\Delta_{11} = R_2 + R_3; \ \Delta_{12} = +R_2; \ \Delta_{11} = -R_2{}^2; \ \Delta_{12} = (R_1 + R_2)(R_2 + R_3)$$

$$\therefore Z_{11} = \frac{\Delta}{\Delta_{11}} = \frac{(R_1 + R_2)(R_2 + R_3) - R_2{}^2}{R_2 + R_3} = R_1 + R_2 - \frac{R_2{}^2}{R_2 + R_3} \qquad \text{from Eq. (23.52)}$$

$$Z_{11} = z_{11} + \frac{\Delta_{11}}{\Delta_{11}} = R_1 + R_2 - \frac{R_2{}^2}{R_2 + R_3} \qquad \text{from Eq. (23.58)}$$

$$Z_{12} = \frac{\Delta}{\Delta_{12}} = \frac{(R_1 + R_2)(R_2 + R_3) - R_2{}^2}{R_2} = R_1 + R_3 + \frac{R_1 R_3}{R_2} \qquad \text{from Eq. (23.53)}$$

$$Z_{12} = z_{12} + \frac{\Delta_{12}}{\Delta_{12}} = -R_2 + \frac{(R_1 + R_2)(R_2 + R_3)}{+R_2} = R_1 + R_3 + \frac{R_1 R_3}{R_2} \qquad \text{from Eq. (23.59)}$$

23.13. Dual and Inverse Networks. The dual of a given network is another network whose equations have the same mathematical form as those of the given network. The inverse of a network is another network whose impedance is proportional to the reciprocal of the impedance of the given network. The following paragraphs give methods for determining dual and inverse networks.

23.13a. Duality. Two different physical systems whose equations are of the same form are said to be dual. Two expressions are dual if the second is formed by replacing all the variables, constants, and driving forces in the first by their duals. As examples $E = IR$ is the dual of $I = EG$ where E, I, and R are replaced by I, E, and G, and $e = L \, di/dt$ is the dual of $i = C \, de/dt$ where e, L, and i are replaced by i, C, and e. The equation must remain valid when transformed. The following tabulation gives dual expressions for the common electrical functions.

func-tion	voltage	impedance	series	resistance	reactance	inductance	mesh
dual	current	admittance	parallel	con-ductance	susceptance	capacitance	node

The dual of a common branch between two meshes is a common branch between nodes. The dual of a network may be drawn by replacing every series connection with a parallel connection and every element with its dual. The loop equation will have the same form as the nodal equation of the dual network, and vice versa. Dual networks can be drawn from geometrical considerations as follows:

1. Arbitrarily assign a numbered point inside each loop together with an additional point outside the network (1, 2, 3 in Fig. 23.30a). These points will be identified with the nodes of the dual network.

2. The branches in the dual network are established by connecting these points with lines passing through each element between the points. In other words, a point inside loop 1 will be connected to the external point by lines passing through only the elements between loop 1 and the external point. The branches are then the duals of the elements so crossed. Thus, Fig. 23.30b can be derived as the dual of Fig. 23.30a. For a network to possess a dual, it must be capable of being drawn on a flat surface without having any branches cross one another. A valid equivalent network can often be formed by taking the dual of a given network twice in succession. For example, Fig. 23.31 shows the conventional T equivalent of a coupled circuit obtained by taking two successive duals.

Note that the *dual of a network* is not the same as the *dual representation* of a network, the latter being merely the original network with all elements replaced by their duals, but the structure remaining unchanged. The two cases are illustrated in Fig. 23.32.

23.13b. Reciprocal and Inverse Networks. Two elements or networks are said to be inverse if the impedance of one is proportional to the reciprocal of the other for all frequencies. The product of the impedances or admittances of two inverse networks

(a) LOCATION OF POINTS AND LINES   (b) DUAL OF FIG. 23.30a
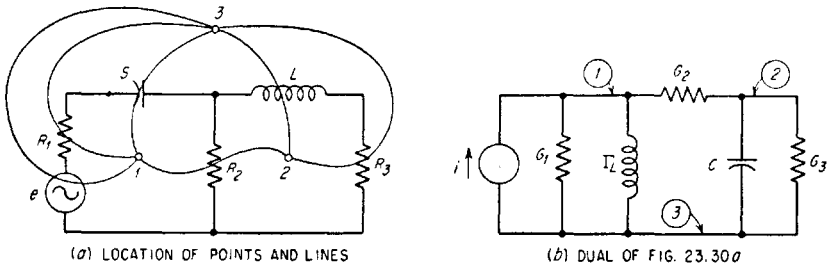
FIG. 23.30. Dual networks.

(a)   (b)   (c)
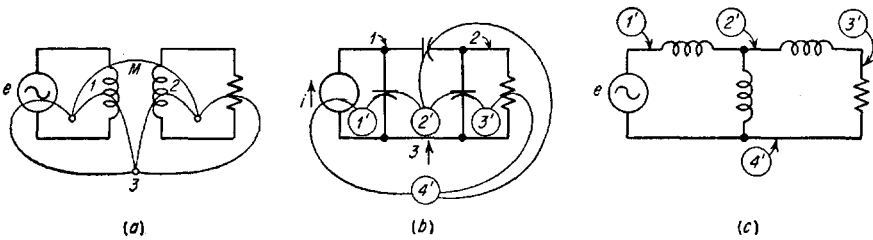
FIG. 23.31. Successive duals of a coupled circuit.

NETWORK A   DUAL OF A   DUAL REPRESENTATION OF A
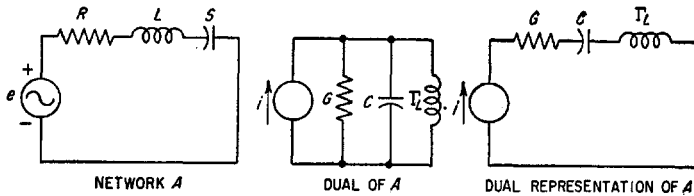
FIG. 23.32. Dual and dual representation.

must, therefore, be a constant. This constant is real and positive and has the dimensions of resistance squared. The networks are said to be inverse with respect to this constant. If the constant is unity, the networks are said to be reciprocal with respect to frequency. The poles and zeros of inverse elements replace each other, i.e., the poles of one element occur at frequencies corresponding to the zeros of the other element. To determine the inverse of a multielement network, the dual of the network is determined, and the magnitudes of the elements of the dual network are given by the inverse of the original element with respect to R_o^2, where R_o is the arbitrary constant noted above, i.e., the impedance of the dual element equals R_o^2 divided by the impedance of the original network element.

Examples of inverse networks are given in Fig. 23.33.

NETWORK	INVERSE NETWORK	RECIPROCAL EQUATIONS
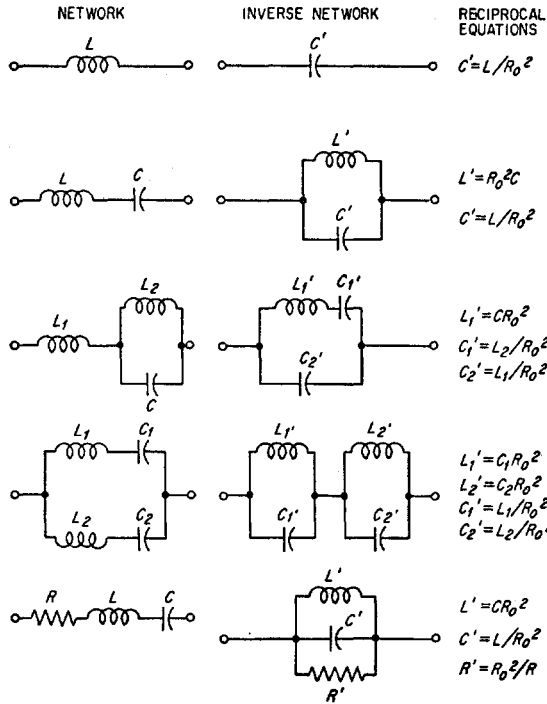
FIG. 23.33. Inverse networks.

23.14. Two-terminal-pair Networks[1]

23.14a. General Relations. In Fig. 23.34, a two-terminal-pair network (a network

FIG. 23.34. General two-terminal-pair network.

containing a pair of input and a pair of output terminals) is shown. The reference directions of input and output currents are chosen for convenience. A change in reference directions changes the signs of some terms in subsequent expressions. The following relationships apply for networks of this type containing linear, bilateral, passive elements.

1. *Voltage-current relationships in terms of network impedances*

$$E_1 = z_{11}I_1 - z_{12}I_2 \qquad (23.64)$$
$$E_2 = z_{21}I_1 - z_{22}I_2 \qquad (23.65)$$

where z_{11} = input impedance between terminals 1-1' with terminals 2-2' open-circuited

$= E_1/I_1$ when $I_2 = 0$

z_{22} = input impedance between terminals 2-2' with terminals 1-1' open-circuited

$= -E_2/I_2$ when $I_1 = 0$

z_{12} = transfer impedance with terminals 1-1 open-circuited

$= -E_1/I_2$ when $I_1 = 0$

z_{21} = transfer impedance with terminals 2-2' open-circuited

$= E_2/I_1$ when $I_2 = 0$

[1] Sometimes referred to as four-terminal networks.

2. *Voltage-current relationships in terms of admittances*

$$I_1 = y_{11}E_1 - y_{12}E_2 \tag{23.66}$$
$$I_2 = y_{21}E_1 - y_{22}E_2 \tag{23.67}$$

where y_{11} = driving-point admittance between terminals 1-1' with terminals 2-2'
short-circuited

$$= I_1/E_1 \text{ when } E_2 = 0$$

y_{22} = driving-point admittance between terminals 2-2' with terminals 1-1'
short-circuited

$$= -I_2/E_2 \text{ when } E_1 = 0$$

y_{12} = transfer admittance with terminals 1-1' shorted

$$= -I_1/E_2 \text{ when } E_1 = 0$$

y_{21} = transfer admittance with terminals 2-2' shorted

$$= I_2/E_1 \text{ when } E_2 = 0$$

3. *Input-output voltage-current relationships*

$$E_1 = AE_2 + BI_2 \tag{23.68}$$
$$I_1 = CE_2 + DI_2 \tag{23.69}$$
$$E_2 = DE_1 - BI_1 \tag{23.70}$$
$$-I_2 = CE_1 - AI_1 \tag{23.71}$$

where A, B, C, and D are "general network parameters" defined in terms of network impedances and admittances in Table 23.8.

Although Eqs. (23.64) to (23.71) apply to a network containing two meshes, they are perfectly general in that any complex network can be reduced to a two-mesh equivalent.

TABLE 23.8. BILATERAL NETWORK COEFFICIENT RELATIONSHIPS

Impedance	Admittance	General coefficients
$z_{11} = \dfrac{y_{22}}{-\Delta_y} = \dfrac{A}{C}$	$y_{11} = \dfrac{z_{22}}{-\Delta_z} = \dfrac{D}{B}$	$A = \dfrac{y_{22}}{y_{12}} = \dfrac{z_{11}}{z_{12}}$
$z_{22} = \dfrac{y_{11}}{-\Delta_y} = \dfrac{D}{C}$	$y_{22} = \dfrac{z_{11}}{-\Delta_z} = \dfrac{A}{B}$	$B = \dfrac{1}{y_{12}} = \dfrac{-\Delta_z}{z_{12}}$
$z_{12} = \dfrac{y_{12}}{-\Delta_y} = \dfrac{1}{C}$	$y_{12} = \dfrac{z_{12}}{-\Delta_z} = \dfrac{1}{B}$	$C = \dfrac{-\Delta_y}{y_{12}} = \dfrac{1}{z_{12}}$
$z_{21} = \dfrac{y_{21}}{-\Delta_y} = \dfrac{1}{C}$	$y_{21} = \dfrac{z_{21}}{-\Delta_z} = \dfrac{1}{B}$	$D = \dfrac{y_{11}}{y_{12}} = \dfrac{z_{22}}{z_{12}}$
		$AD - BC = 1$
$\Delta_z = z_{12}{}^2 - z_{11}z_{22}$	$\Delta_y = y_{12}{}^2 - y_{11}y_{22}$	$\Delta_z = 1/\Delta_y$

Four network coefficients are required to establish the performance of a four-terminal network containing unilateral elements. For linear, passive, bilateral networks, the reciprocity theorem requires the transfer impedances to be equal in both directions and the transfer admittances to be equal in both directions. Consequently, the performance of such networks is completely specified by three independent parameters. This allows an unknown network to be represented by a T or π equivalent.

23.14b. Symmetrical Network. A network is said to be symmetrical when input and output terminals can be interchanged with no change in external behavior. This will occur for $\Delta_{11} = \Delta_{22}$ or $y_{11} = y_{22}$ and $z_{11} = z_{22}$. Only two properly chosen independent parameters are necessary to specify the performance of a symmetrical network.

23.14c. Image Impedance. A four-terminal network is operating on an image impedance basis when the generator impedance is equal to the input impedance of the network with the load connected, and the load impedance is equal to the imped-ance looking into the network from the load with the generator impedance connected. A characteristic which, together with the image impedances, completely specifies the performance of the network is the *image transfer constant* θ. The image transfer constant can be defined as one-half the natural logarithm of the complex ratio of the steady-state volt-amperes entering and leaving the network when the network is terminated in its image impedance. The transfer constant has a real (attenuation) and imaginary (phase-shift) part. The relationships between these image coefficients and other network coefficients are as follows:

$$Z_{I_1} = \sqrt{Z_{oc1}Z_{sc1}} \qquad (23.72)$$

$$Z_{I_2} = \sqrt{Z_{oc2}Z_{sc2}} \qquad (23.73)$$

where Z_{I_1} = image impedance looking into input terminals and is equal to the source
(generator) impedance

Z_{I_2} = image impedance looking into output terminals and is equal to the load impedance

Z_{oc1} = impedance between input terminals with output terminals open-circuited

Z_{sc1} = impedance between input terminals with output terminals short-circuited

Z_{oc2} = impedance between output terminals with input terminals open-circuited

Z_{sc2} = impedance between output terminals with input terminals short-circuited

Further, $$\tanh \theta = \sqrt{\frac{Z_{sc1}}{Z_{oc1}}} = \sqrt{\frac{Z_{sc2}}{Z_{oc2}}} \qquad (23.74)$$

$$\frac{E_2}{E_1} = \sqrt{\frac{Z_{I_2}}{Z_{I_1}}} \, \epsilon^{-\theta}; \frac{I_2}{I_1} = \sqrt{\frac{Z_{I_1}}{Z_{I_2}}} \, \epsilon^{-\theta} \qquad (23.75)$$

$$\theta = \frac{1}{2} \ln \frac{E_1 I_1}{E_2 I_2}$$

where ϵ = base of natural logarithms = 2.718
ln = natural logarithm

$$\theta = \alpha + j\beta \qquad (23.76)$$

where α = image attenuation constant, nepers
β = image phase constant, radians

When several networks are cascaded on an image impedance basis, the impedance looking into the output of a given network must equal the impedance looking into

$$\theta = \theta_1 + \theta_2 + \theta_3$$
$$= \alpha_1 + \alpha_2 + \alpha_3 + j(\beta_1 + \beta_2 + \beta_3)$$
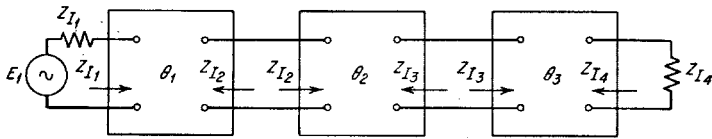
FIG. 23.35. Networks cascaded on an image basis.

the input of the next successive network. The over-all network may then be regarded as having the input image impedance of the first network, the output image impedance of the last network, and an over-all image transfer constant equal to the sum of the transfer constants of each section. This is illustrated in Fig. 23.35.

23.14d. Iterative Impedance. A four-terminal network is operating on an iterative impedance basis when the input impedance with the load impedance connected is

equal to the load impedance, and the impedance looking into the output terminals with the generator impedance connected is equal to the generator impedance. Another characteristic which, together with the iterative impedance, completely specifies the performance of the network is the *iterative transfer constant P*. The iterative transfer constant can be defined as the natural logarithm of the ratio of input and output currents of a network terminated on an iterative basis at both ends. P, is in

$$P = P_1 + P_2 + P_3$$
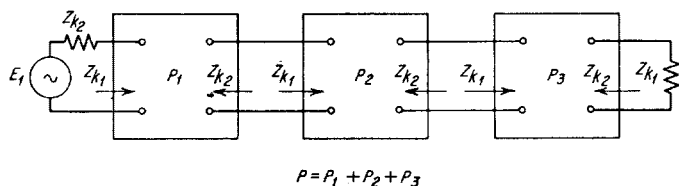
Fig. 23.36. Networks cascaded on an iterative-impedance basis.

general, complex, containing a real part (attenuation in nepers) and an imaginary part (phase shift in radians).

$$P = \ln \frac{I_1}{I_2} \qquad (23.77)$$

When successive networks are cascaded on an iterative-impedance basis there will be a mismatch at each junction, but the over-all transfer constant will be the sum of the transfer constants of each section. This is illustrated in Fig. 23.36.

Example 23.21

Determine the relations required between elements in the network shown in Fig. 23.37 for it to be symmetrical.

Solution

By inspection, it is seen that the network is symmetrical, that is, $z_{11} = z_{22}$ under the conditions that $Z_1 = Z_3$ and $Z_4 = Z_5$. It is of interest, however, to investigate the possibility of other solutions. By definition, for a symmetrical network,

$$\Delta_{11} = \Delta_{22}$$

From Fig. 23.37

$$\Delta = \begin{vmatrix} (Z_1 + Z_2 + Z_4)(Z_2) \\ (Z_2)(Z_2 + Z_3 + Z_5) \end{vmatrix}$$
$$\Delta_{11} = Z_2 + Z_3 + Z_5$$
$$\Delta_{22} = Z_1 + Z_2 + Z_4$$

Therefore, for symmetry,

$$Z_3 + Z_5 = Z_1 + Z_4$$

This is more general than the solution obtained by inspection. If

$$Z_s = Z_3 + Z_5 = Z_1 + Z_4$$

the network can be represented completely by only two variables, Z_s and Z_2.
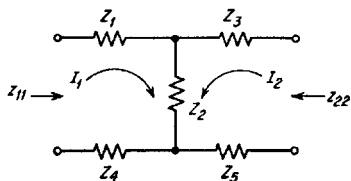
Fig. 23.37. Network for Example 23.21.
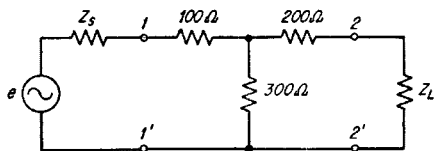
Fig. 23.38. Network for Example 23.22.

Example 23.22

Calculate the image impedances and image transfer constant for the network shown in Fig. 23.38.

Solution

For image impedance operation

$$Z_{I_1} = Z_s$$
$$Z_{I_2} = Z_L$$

From Eqs. (23.72) and (23.73)

$$Z_{I_1} = \sqrt{Z_{oc1}Z_{sc1}}$$
$$Z_{I_2} = \sqrt{Z_{oc2}Z_{sc2}}$$

By inspection of **Fig. 23.38**

$$Z_{oc1} = 100 + 300 = 400\,\Omega$$
$$Z_{sc1} = 100 + \frac{200 \times 300}{200 + 300} = 220\,\Omega$$
$$Z_{oc2} = 200 + 300 = 500\,\Omega$$
$$Z_{sc2} = 200 + \frac{100 \times 300}{100 + 300} = 275\,\Omega$$

Therefore,

$$Z_{I_1} = \sqrt{400 \times 220} = 297\,\Omega = Z_s$$
$$Z_{I_2} = \sqrt{500 \times 275} = 371\,\Omega = Z_L$$

From Eq. (23.74)

$$\tanh \theta = \sqrt{\frac{Z_{sc1}}{Z_{oc1}}} = \sqrt{\frac{220}{400}} = 0.74$$
$$\tanh \theta = \frac{e^\theta - e^{-\theta}}{e^\theta + e^{-\theta}}$$
$$\therefore\ e^\theta - e^{-\theta} = 0.74e^\theta + 0.74e^{-\theta}$$
$$0.26e^\theta = 1.74e^{-\theta}$$
$$e^{2\theta} = \frac{1.74}{0.26} = 6.7$$

Therefore,

$$2\theta = \ln 6.7 = 1.90$$
$$\theta = 0.95 \text{ neper}$$

23.15. Matrix Methods. Matrix algebra is a system of mathematical shorthand which can be used to write simultaneous equations in a very compact form and to operate upon them in an orderly manner. A matrix is an orderly array of coefficients into rows and columns. When applied to networks, matrices can describe the physical configuration of the network by means of the arrangement and type of the matrix coefficients. It is not a determinant in that the matrix merely represents the arrangement of the coefficients of a systematically written set of equations and does not represent the combination of the equations as in a determinant. A matrix has no "value" and cannot be solved by determinant methods. The number of rows and columns need not be the same in a matrix.

The matrix of a system of linear equations is formed by the array of coefficients of the system. Consider the following set of equations:

$$a_{11}x_1 + a_{12}x_2 = y_1 \qquad (23.78)$$
$$a_{21}x_1 + a_{22}x_2 = y_2$$

The matrix of the coefficients of x_1 and x_2 in this set of equations is formed as follows:

$$a = \begin{Vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{Vmatrix} \qquad (23.79)$$

The double line distinguishes a matrix.

23.15a. Matrix Multiplication. Two matrices can be multiplied to form a third matrix as follows:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \times \begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix} = \begin{vmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{vmatrix} \tag{23.80}$$

where $c_{11} = a_{11}b_{11} + a_{12}b_{21}$
$c_{12} = a_{11}b_{12} + a_{12}b_{22}$
$c_{21} = a_{21}b_{11} + a_{22}b_{21}$
$c_{22} = a_{21}b_{12} + a_{22}b_{22}$

This multiplication is achieved by the application of the following general rule:

The c_{jk} coefficient in the product matrix is formed by multiplying the jth row of the first matrix by the kth column of the second matrix. Rows are multiplied by columns by forming the sum of the products of the first element of the row and the first element of the column, the second element of the row and the second element of the column, etc.

Matrix multiplication is indicated by

$$a \times b = c \tag{23.81}$$

where a_{jk} indicates the matrix

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{j1} & a_{j2} & \cdots & a_{jk} \end{vmatrix}$$

Matrix multiplication is, in general, not commutative; i.e.,

$$a \times b \neq b \times a \tag{23.82}$$

Matrix addition is commutative; i.e.,

$$a + b = b + a \tag{23.83}$$

Matrix multiplication is associative; i.e.,

$$a(b \times c) = (a \times b)c \tag{23.84}$$

Matrix addition is associative; i.e.,

$$a + (b + c) = (a + b) + c \tag{23.85}$$

The basis for determining the order of multiplication will be discussed later. To multiply a matrix a by a constant factor k, each coefficient of the matrix is multiplied by k, and the matrix is denoted by ka.

23.15b. Matrix Equations. Equation (23.78) can be expressed in the following matrix form:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \times \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = \begin{vmatrix} y_1 \\ y_2 \end{vmatrix} \tag{23.86}$$

Applying the multiplication rules to Eq. (23.86),

$$\begin{vmatrix} (a_{11}x_1 + a_{12}x_2) \\ (a_{21}x_1 + a_{22}x_2) \end{vmatrix} = \begin{vmatrix} y_1 \\ y_2 \end{vmatrix} \tag{23.87}$$

The equality sign in a matrix equation indicates the matrices on each side of the sign are equal term for term. Therefore,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= y_1 \\ a_{21}x_1 + a_{22}x_2 &= y_2 \end{aligned} \tag{23.88}$$

which is again Eq. (23.78), indicating that this interpretation of matrix equations (23.86) and (23.87) is valid.

Consider the following two sets of linear equations and their corresponding matrix equations.

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = y_1$$
$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = y_2 \qquad (23.89)$$
$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = y_3$$

or
$$\boldsymbol{a} \times \boldsymbol{x} = \boldsymbol{y}$$

$$b_{11}y_1 + b_{12}y_2 + b_{13}y_3 = z_1$$
$$b_{21}y_1 + b_{22}y_2 + b_{23}y_3 = z_2 \qquad (23.90)$$
$$b_{31}y_1 + b_{32}y_2 + b_{33}y_3 = z_3$$

or
$$\boldsymbol{b} \times \boldsymbol{y} = \boldsymbol{z}$$

If it is required to determine by simple algebraic methods the relationship between the x terms of Eq. (23.89) and the z terms of Eq. (23.90), Eq. (23.89) must be substituted in Eq. (23.90), requiring considerable effort. The same result can be achieved in matrix form by substituting for the y matrix of Eq. (23.90) the equivalent expression given by Eq. (23.89).

$$\begin{Vmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{Vmatrix} \times \begin{Vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{Vmatrix} \times \begin{Vmatrix} x_1 \\ x_2 \\ x_3 \end{Vmatrix} = \begin{Vmatrix} z_1 \\ z_2 \\ z_3 \end{Vmatrix} \qquad (23.91)$$

or
$$\boldsymbol{b} \times \boldsymbol{a} \times \boldsymbol{x} = \boldsymbol{z}$$

Matrix operations organize the effort required to solve systems of linear equations.

23.15c. Inverse of a Matrix (Transposition). If in Eq. (23.89) it is required to determine the expression for the x's in terms of the y's, the result may be obtained by rewriting Eq. (23.89) as follows:

$$\begin{Vmatrix} x_1 \\ x_2 \\ x_3 \end{Vmatrix} = \begin{Vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{Vmatrix}^{-1} \times \begin{Vmatrix} y_1 \\ y_2 \\ y_3 \end{Vmatrix} \qquad (23.92)$$

or
$$\boldsymbol{x} = \boldsymbol{a}^{-1} \times \boldsymbol{y}$$

The $\boldsymbol{a}^{-1}$ matrix is called the inverse $\boldsymbol{a}$ matrix. An inverse matrix can only be obtained for square matrices (number of rows equal number of columns) with the additional restriction that the value of the determinant formed from the coefficients of the matrix must not equal zero. An inverse matrix is obtained as follows:

1. Replace each element of $\boldsymbol{a}$ by its minor in the determinant $|a|$ formed by the corresponding coefficients of $\boldsymbol{a}$. A minor of the element a_{jk} is a determinant formed by deleting the jth row and kth column.

2. Divide each minor by the determinant $|a|$.

3. Transpose the result (interchange rows and columns).

As an example, consider the $\boldsymbol{z}$ matrix equation obtained from the loop equations of a network, $\boldsymbol{e} = \boldsymbol{z} \times \boldsymbol{i}$. To obtain the solution for the currents in terms of the voltages, the matrix equation $\boldsymbol{i} = \boldsymbol{z}^{-1} \times \boldsymbol{e}$ must be solved. The inverse of the $\boldsymbol{z}$ matrix for a three-loop network is obtained as follows:

$$\boldsymbol{z} = \begin{vmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ z_{31} & z_{32} & z_{33} \end{vmatrix} \qquad (23.93)$$

After carrying out steps 1 and 2, the result is

$$
\begin{vmatrix}
\dfrac{\begin{vmatrix} z_{22} & z_{23} \\ z_{32} & z_{33} \end{vmatrix}}{|z|} & -\dfrac{\begin{vmatrix} z_{21} & z_{23} \\ z_{31} & z_{33} \end{vmatrix}}{|z|} & \dfrac{\begin{vmatrix} z_{21} & z_{22} \\ z_{31} & z_{32} \end{vmatrix}}{|z|} \\[4mm]
-\dfrac{\begin{vmatrix} z_{12} & z_{13} \\ z_{32} & z_{33} \end{vmatrix}}{|z|} & \dfrac{\begin{vmatrix} z_{11} & z_{13} \\ z_{31} & z_{33} \end{vmatrix}}{|z|} & -\dfrac{\begin{vmatrix} z_{11} & z_{12} \\ z_{31} & z_{32} \end{vmatrix}}{|z|} \\[4mm]
\dfrac{\begin{vmatrix} z_{12} & z_{13} \\ z_{22} & z_{23} \end{vmatrix}}{|z|} & -\dfrac{\begin{vmatrix} z_{11} & z_{13} \\ z_{21} & z_{23} \end{vmatrix}}{|z|} & \dfrac{\begin{vmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{vmatrix}}{|z|}
\end{vmatrix}
\tag{23.94}
$$

or

$$
\begin{vmatrix}
\dfrac{\Delta_{11}}{\Delta} & \dfrac{\Delta_{12}}{\Delta} & \dfrac{\Delta_{13}}{\Delta} \\[3mm]
\dfrac{\Delta_{21}}{\Delta} & \dfrac{\Delta_{22}}{\Delta} & \dfrac{\Delta_{23}}{\Delta} \\[3mm]
\dfrac{\Delta_{31}}{\Delta} & \dfrac{\Delta_{32}}{\Delta} & \dfrac{\Delta_{33}}{\Delta}
\end{vmatrix}
\tag{23.95}
$$

where $\Delta = |z|$ and Δ_{jk} is the minor of Δ with respect to element jk
Interchanging rows and columns:

$$
z^{-1} =
\begin{vmatrix}
\dfrac{\Delta_{11}}{\Delta} & \dfrac{\Delta_{21}}{\Delta} & \dfrac{\Delta_{31}}{\Delta} \\[3mm]
\dfrac{\Delta_{12}}{\Delta} & \dfrac{\Delta_{22}}{\Delta} & \dfrac{\Delta_{32}}{\Delta} \\[3mm]
\dfrac{\Delta_{13}}{\Delta} & \dfrac{\Delta_{23}}{\Delta} & \dfrac{\Delta_{33}}{\Delta}
\end{vmatrix}
= y
\tag{23.96}
$$

the final result is

$$
\begin{vmatrix} i_1 \\ i_2 \\ i_3 \end{vmatrix}
=
\begin{vmatrix}
\dfrac{\Delta_{11}}{\Delta} & \dfrac{\Delta_{21}}{\Delta} & \dfrac{\Delta_{31}}{\Delta} \\[3mm]
\dfrac{\Delta_{12}}{\Delta} & \dfrac{\Delta_{22}}{\Delta} & \dfrac{\Delta_{32}}{\Delta} \\[3mm]
\dfrac{\Delta_{13}}{\Delta} & \dfrac{\Delta_{23}}{\Delta} & \dfrac{\Delta_{33}}{\Delta}
\end{vmatrix}
\times
\begin{vmatrix} e_1 \\ e_2 \\ e_3 \end{vmatrix}
\tag{23.97}
$$

Note that the impedance and admittance matrices of a given network are inverse.
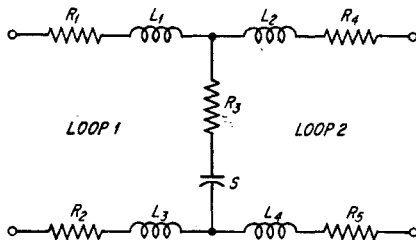
FIG. 23.39. Network for Eq. (23.98).

Figure 23.39 and Eq. (23.98) illustrate a network and the corresponding impedance matrix,

$$
z =
\begin{Vmatrix}
\left[(R_1 + R_2 + R_3) + s(L_1 + L_3) + \dfrac{S}{s} \right] & \left[R_3 + \dfrac{S}{s} \right] \\[4mm]
\left[R_3 + \dfrac{S}{s} \right] & \left[(R_3 + R_4 + R_5) + s(L_2 + L_4) + \dfrac{S}{s} \right]
\end{Vmatrix}
\tag{23.98}
$$

A *unit matrix* u is defined as a square matrix with the value of the diagonal elements equal to unity and all other elements equal to zero. For example, for a fourth-order unit matrix,

$$u = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix} \tag{23.99}$$

The product of a matrix and its inverse is a unit matrix.

23.15d. Matrix Addition and Subtraction. Matrices are added or subtracted by adding or subtracting corresponding coefficients. For example:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \pm \begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix} = \begin{vmatrix} (a_{11} \pm b_{11})(a_{12} \pm b_{12}) \\ (a_{21} \pm b_{21})(a_{22} \pm b_{22}) \end{vmatrix} \tag{23.100}$$

23.15e. Matrix Differentiation and Integration. Matrices can be differentiated or integrated by differentiating or integrating each coefficient.

23.15f. Matrices of Simple Networks. To aid in understanding the method by which matrices are set up, the simple four-terminal networks shown in Fig. 23.40 will
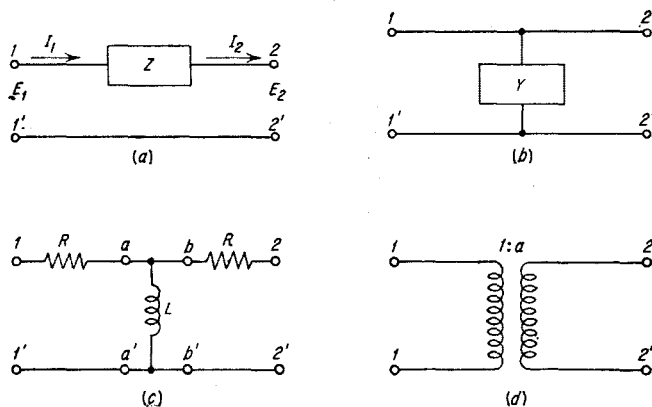
Fig. 23.40. Simple four-terminal networks.

be considered. The equations describing the performance of these networks can be written in several ways. One general method of describing all four-terminal networks is given by Eq. (23.101) (see Sec. 23.14):

$$E_1 = AE_2 + BI_2$$
$$I_1 = CE_2 + DI_2 \tag{23.101}$$

These may be expressed in matrix form as follows:

$$\begin{vmatrix} E_1 \\ I_1 \end{vmatrix} = \begin{vmatrix} A & B \\ C & D \end{vmatrix} \times \begin{vmatrix} E_2 \\ I_2 \end{vmatrix} \tag{23.102}$$

To determine the coefficients of the matrices of the networks of Fig. 23.40 consider the following example. From Fig. 23.40a note that when $I_2 = 0$, $E_1 = E_2$, and therefore $A = 1$. When $E_2 = 0$, $E_1 = ZI_2$, and therefore $B = Z$. When $I_2 = 0$, $I_1 = 0$, and therefore $C = 0$. When $E_2 = 0$, $I_1 = I_2$, and therefore $D = 1$. The matrix for the network of Fig. 23.40a is therefore

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = \begin{vmatrix} 1 & Z \\ 0 & 1 \end{vmatrix} \tag{23.103}$$

and Eq. (23.102) becomes

$$\begin{vmatrix} E_1 \\ I_1 \end{vmatrix} = \begin{vmatrix} 1 & Z \\ 0 & 1 \end{vmatrix} \times \begin{vmatrix} E_2 \\ I_2 \end{vmatrix} \tag{23.104}$$

which can be expanded into

$$E_1 = E_2 + ZI_2$$
$$I_1 = I_2 \tag{23.105}$$

By similar reasoning, the matrix of the network shown in Fig. 23.40b is as follows:

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ Y & 1 \end{vmatrix} \tag{23.106}$$

The network of Fig. 23.40c can be considered as three simple networks similar to the series and shunt cases described above in cascade.   The over-all matrix is therefore equal to the product:

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = \begin{vmatrix} 1 & R \\ 0 & 1 \end{vmatrix} \times \begin{vmatrix} 1 & 0 \\ \dfrac{1}{j\omega L} & 1 \end{vmatrix} \times \begin{vmatrix} 1 & R \\ 0 & 1 \end{vmatrix} \tag{23.107}$$

The product of the first two terms is

$$\begin{vmatrix} 1 & R \\ 0 & 1 \end{vmatrix} \times \begin{vmatrix} 1 & 0 \\ \dfrac{1}{j\omega L} & 1 \end{vmatrix} = \begin{vmatrix} \left(1 + \dfrac{R}{j\omega L}\right) & R \\ \dfrac{1}{j\omega L} & 1 \end{vmatrix} \tag{23.108}$$

The product of (23.108) and the last term is

$$\begin{vmatrix} \left(1 + \dfrac{R}{j\omega L}\right) & R \\ \dfrac{1}{j\omega L} & 1 \end{vmatrix} \times \begin{vmatrix} 1 & R \\ 0 & 1 \end{vmatrix} = \begin{vmatrix} \left(1 + \dfrac{R}{j\omega L}\right) & \left[R\left(1 + \dfrac{R}{j\omega L}\right) + R\right] \\ \dfrac{1}{j\omega L} & \left[\dfrac{R}{j\omega L} + 1\right] \end{vmatrix} \tag{23.109}$$

This reduces to

$$\frac{1}{j\omega L}\begin{vmatrix} (R + j\omega L) & R(2j\omega L + R) \\ 1 & j\omega L + R \end{vmatrix} \tag{23.110}$$

An ideal transformer with a turns ratio of $1{:}a$ between the primary and secondary windings is shown in Fig. 23.40d.   The matrix for this network is

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = \begin{vmatrix} \dfrac{1}{a} & 0 \\ 0 & a \end{vmatrix} \tag{23.111}$$

Matrix methods afford a very convenient method of determining the over-all transfer function, that is E_0/E_i of many types of networks.   The network is divided into a series of simple cascaded elements.   The matrix of each element can usually be written from inspection or by use of the previous examples. The individual matrices are multiplied to form the complete product matrix.   The over-all transfer function is then obtained by taking the reciprocal of the A coefficient in the product matrix (this assumes no additional terminal loading).   Applying this
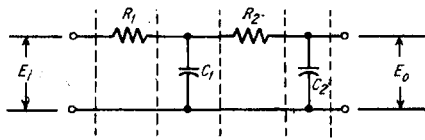
FIG. 23.41. Ladder network.

procedure to the network shown in Fig. 23.41,

$$\begin{vmatrix} A_1 & B_1 \\ C & D_1 \end{vmatrix} \times \begin{vmatrix} A_2 & B_2 \\ C_2 & D_2 \end{vmatrix} \times \begin{vmatrix} A_3 & B_3 \\ C_3 & D_3 \end{vmatrix} \times \begin{vmatrix} A_4 & B_4 \\ C_4 & D_4 \end{vmatrix} = \begin{vmatrix} A & B \\ C & D \end{vmatrix} \quad (23.112)$$

Substituting values for the individual matrices,

$$\begin{vmatrix} 1 & R_1 \\ 0 & 1 \end{vmatrix} \times \begin{vmatrix} 1 & 0 \\ sC_1 & 1 \end{vmatrix} \times \begin{vmatrix} 1 & R_2 \\ 0 & 1 \end{vmatrix} \times \begin{vmatrix} 1 & 0 \\ sC_2 & 1 \end{vmatrix} = \begin{vmatrix} A & B \\ C & D \end{vmatrix} \quad (23.113)$$

The result of multiplying the individual matrices is

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = \begin{vmatrix} [(1 + sR_1C_1)(1 + sR_2C_2) + sR_1C_2] & [R_2(1 + sR_1C_1) + R_1] \\ [sC_1(1 + sR_2C_2) + s^2C_1C_2] & [sR_2C_1 + 1] \end{vmatrix} \quad (23.114)$$

The transfer function is therefore

$$\frac{E_0}{E_i} = \frac{1}{A} = \frac{1}{s^2(R_1R_2C_1C_2) + s(R_1C_1 + R_2C_2 + R_1C_2) + 1}$$

(See Table 1.4, #9.)

23.16. Network Transformations

23.16a. Equivalent T and π Relationships. Any unsymmetrical network in the form of a T or a π can be replaced by an equivalent π or T network, respectively. Although the transformation from T to π or vice versa is completely general, the transformation will not necessarily yield a network which will duplicate the performance of the original network over a band of frequencies. A physically realizable equivalent network can always be obtained at a single frequency, however, and often one can be obtained which is valid over the frequency band of interest.
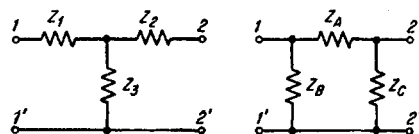
FIG. 23.42. T and π networks.

The T and π networks are shown in Fig. 23.42. It is frequently desirable to convert a T network into its π equivalent, and vice versa. The required relationships are given below:

π to T transformation,

$$Z_1 = \frac{Z_A Z_B}{Z_A + Z_B + Z_C} \quad (23.115)$$

$$Z_2 = \frac{Z_A Z_C}{Z_A + Z_B + Z_C} \quad (23.116)$$

$$Z_3 = \frac{Z_B Z_C}{Z_A + Z_B + Z_C} \quad (23.117)$$

T to π transformation,

$$Z_A = \frac{Z_1 Z_2 + Z_2 Z_3 + Z_1 Z_3}{Z_3} \quad (23.118)$$

$$Z_B = \frac{Z_1 Z_2 + Z_2 Z_3 + Z_1 Z_3}{Z_2} \quad (23.119)$$

$$Z_C = \frac{Z_1 Z_2 + Z_2 Z_3 + Z_1 Z_3}{Z_1} \quad (23.120)$$

If the networks are symmetrical, the impedance looking into each end of the network is the same and the equivalent T or π can be expressed in terms of only two independent variables. When two networks are equivalent, the addition of an identical impedance between input and output terminals of each network does not

destroy the equivalence.   Further, the addition of identical terminating impedances does not alter the equivalence of unterminated networks.

The elements of T and π networks can be expressed in terms of external measurements of open- and short-circuit impedances or admittances.   These expressions are as follows:

For the T network:

$$Z_{oc1} = Z_1 + Z_3 \tag{23.121}$$

$$Z_{sc1} = Z_1 + \frac{Z_2 Z_3}{Z_2 + Z_3} \tag{23.122}$$

$$Z_{oc2} = Z_2 + Z_3 \tag{23.123}$$

$$Z_{sc2} = \frac{Z_1 Z_3}{Z_1 + Z_3} + Z_2 \tag{23.124}$$

where Z_{oc1} = impedance across terminals 1-1' with terminals 2-2' open-circuited
$\quad\;\; Z_{sc1}$ = impedance across terminals 1-1' with terminals 2-2' short-circuited
$\quad\;\; Z_{oc2}$ = impedance across terminals 2-2' with terminals 1-1' open-circuited
$\quad\;\; Z_{sc2}$ = impedance across terminals 2-2' with terminals 1-1' short-circuited

$$Z_1 = Z_{oc1} - \sqrt{Z_{oc2}(Z_{oc1} - Z_{sc1})} \tag{23.125}$$

$$Z_2 = Z_{oc2} - \sqrt{Z_{oc2}(Z_{oc1} - Z_{sc1})} \tag{23.126}$$

$$Z_3 = \sqrt{Z_{oc2}(Z_{oc1} - Z_{sc1})} \tag{23.127}$$

For the π network,

$$Z_{oc1} = \frac{Z_B(Z_A + Z_C)}{Z_A + Z_B + Z_C} \tag{23.128}$$

$$Z_{sc1} = \frac{Z_B Z_A}{Z_B + Z_A} \tag{23.129}$$

$$Z_{oc2} = \frac{Z_C(Z_A + Z_B)}{Z_A + Z_B + Z_C} \tag{23.130}$$

$$Z_{sc2} = \frac{Z_C Z_A}{Z_C + Z_A} \tag{23.131}$$

$$Z_A = \frac{Z_{oc2} Z_{sc1}}{\sqrt{Z_{oc2}(Z_{oc1} - Z_{sc1})}} \tag{23.132}$$

$$Z_B = \frac{Z_{oc2} Z_{sc1}}{Z_{oc2} - \sqrt{Z_{oc2}(Z_{oc1} - Z_{sc1})}} \tag{23.133}$$

$$Z_C = \frac{Z_{oc2} Z_{sc1}}{Z_{oc1} - \sqrt{Z_{oc2}(Z_{oc1} - Z_{sc1})}} \tag{23.134}$$

23.16b. Balanced T and π Networks.   If balanced, i.e., symmetrical with respect to ground, networks are required, the balanced T or π configurations shown in Fig.
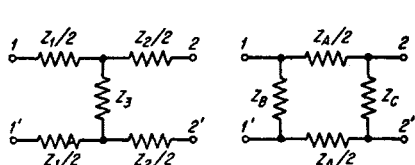
FIG. 23.43. Balanced T and π networks.   FIG. 23.44. General lattice network.
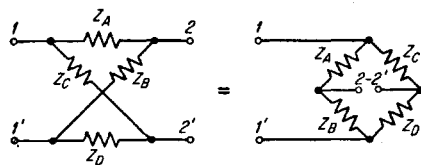
23.43 can be used.   Equations (23.115) to (23.134) can also be used for these balanced structures.

23.16c. Lattice Structures.[1]   A lattice network is shown in Fig. 23.44.   Lattices can

---

[1] See Sec. 16.2 for transmission properties of bridged T and parallel T networks.

be reduced to T or π equivalents. The relationships are given below for the T equivalent.

$$Z_1 = \frac{Z_A Z_C + 2Z_A Z_D + Z_B Z_D}{Z_A + Z_B + Z_C + Z_D} \qquad (23.135)$$

$$Z_2 = \frac{Z_A Z_B + 2Z_A Z_D + Z_C Z_D}{Z_A + Z_B + Z_C + Z_D} \qquad (23.136)$$

$$Z_3 = \frac{Z_B Z_C - Z_A Z_D}{Z_A + Z_B + Z_C + Z_D} \qquad (23.137)$$

Lattice structures are symmetrical and balanced when $Z_A = Z_D$, $Z_B = Z_C$ as shown in Fig. 23.45. Symmetrical lattices can be converted into symmetrical T and π networks, and vice versa. The following transformation equations apply to T networks.
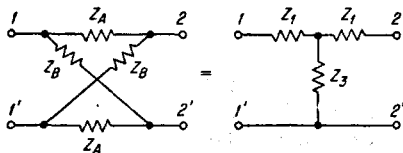
FIG. 23.45. Symmetrical lattice and T networks.

Symmetrical lattice to symmetrical T:

$$Z_1 = Z_A \qquad (23.138)$$

$$Z_3 = \frac{Z_B - Z_A}{2} \qquad (23.139)$$

The quantity $Z_B - Z_A$ must be physically realizable.
Symmetrical T to symmetrical lattice:

$$Z_A = Z_1 \qquad (23.140)$$

$$Z_B = Z_1 + 2Z_3 \qquad (23.141)$$

When transforming from a symmetrical lattice to a symmetrical π, the shunt arms of the π are equal to the lattice impedance Z_B and the series arm is equal to the impedance $2Z_A Z_B/(Z_B - Z_A)$ formed by the lattice elements. The latter term must be physically realizable.

Example 23.23

Obtain the equivalent T and π networks for the inductively coupled circuit shown in Fig. 23.46. Assume positive mutual coupling.
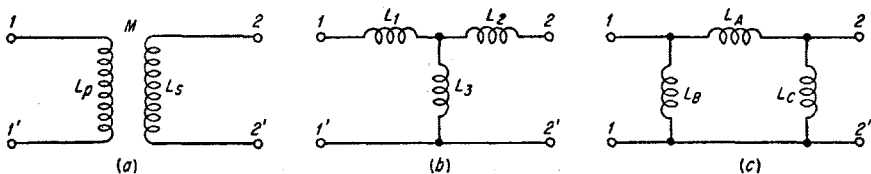
FIG. 23.46. Equivalent coupled circuits.

Solution

Although the equivalence equations are expressed in terms of impedance, the only circuit elements entering into this problem are inductances, and consequently, the equations can be interpreted directly in terms of inductance values as follows:
For the T network:

From Eq. (23.125) $L_1 = L_{oc1} - \sqrt{L_{oc2}(L_{oc1} - L_{sc1})}$

From Eq. (23.126) $L_2 = L_{oc2} - \sqrt{L_{oc2}(L_{oc1} - L_{sc1})}$

From Eq. (23.127) $L_3 = \sqrt{L_{oc2}(L_{oc1} - L_{sc1})}$

By inspection of Fig. 23.46a,

$$L_{oc1} = L_p$$
$$L_{oc2} = L_s$$
$$L_{sc1} = L_p - \frac{M^2}{L_s}$$

Therefore,
$$L_3 = \sqrt{L_s\left(L_p - L_p + \frac{M^2}{L_s}\right)} = M$$
$$L_2 = L_s - M$$
$$L_1 = L_p - M$$

For the π network:

From Eq. (23.118) $L_A = \dfrac{L_1L_2 + L_2L_3 + L_3L_1}{L_3}$

$$L_A = \frac{(L_p - M)(L_s - M) + (L_s - M)M + (L_p - M)M}{M}$$

$$= \frac{L_pL_s - M^2}{M}$$

From Eq. (23.119) $L_B = \dfrac{L_1L_2 + L_2L_3 + L_3L_1}{L_2}$

$$= \frac{L_pL_s - M^2}{L_s - M}$$

From Eq. (23.120) $L_C = \dfrac{L_1L_2 + L_2L_3 + L_3L_1}{L_1}$

$$= \frac{L_pL_s - M^2}{L_p - M}$$

23.17. Dimensional Analysis. All[1] the quantities used in electrical engineering can be expressed in terms of four "fundamental" quantities. These are often taken

TABLE 23.9. DIMENSIONAL EQUIVALENTS

Quantity	Symbol	Dimensions	Quantity	Symbol	Dimensions
Capacitance	C	$R^{-1}T$	Frequency	f,ω	T^{-1}
Charge	Q	IT	Force	F	$RI^2L^{-1}T$
Linear charge density	l	$IL^{-1}T$	Inductance	L,M	RT
Surface charge density	σ	$IL^{-2}T$	Magnetic-field intensity	H	IL^{-1}
Volume charge density	ρ	$IL^{-3}T$	Magnetic flux	Φ	RIT
Conductance	G	R^{-1}	Magnetic flux density	B	$RIL^{-2}T$
Conductivity	γ	$R^{-1}L^{-1}$	Mass	M	$RI^2L^{-2}T^3$
Current	I	I	Magnetomotive force	MMF	I
Current density	J	IL^{-2}	Permeance	P	RT
Dielectric constant	ϵ	$R^{-1}L^{-1}T$	Permeability	μ	$RL^{-1}T$
Displacement (electric flux density)	D	$IL^{-2}T$	Power	P	RI^2
Elastance	S	RT^{-1}	Reluctance	$\mathfrak{R}$	$R^{-1}T^{-1}$
Electric-field intensity	$\mathcal{E}$	RIL^{-1}	Resistance	R	R
Electric flux	ψ	IT	Resistivity	ρ	RL
			Time	T	T
Energy (work)	W	RI^2T	Voltage (potential difference)	E	RI

as the dimensional units of mechanical physics, i.e., mass, length, time, and permeability; however, four quantities more applicable to electrical systems are resistance R, current I, length L, and time T. Four other quantities could have been chosen, in which case resistance, current, length, and time would be defined in terms of these quantities. There is some reason for this choice, however, since these quantities

[1] Excluding temperature.

are subject to measurement with extremely precise standards.   Table 23.9 lists several electrical engineering quantities and their dimensional formulas.

Example 23.24

Find the dimensional expressions for the following:

Quantity	Defining equation	Dimensions
Amplification factor...........	$\mu = \left(\dfrac{\partial e_p}{\partial e_g}\right)_{i_p = c}$	$\dfrac{RI}{RI} =$ dimensionless
Transconductance..............	$g_m = \left(\dfrac{\partial i_p}{\partial e_g}\right)_{e_p = c}$	$\dfrac{I}{RI} = R^{-1}$
Phase........................	$\theta = \omega t$	$\dfrac{T}{T} =$ dimensionless
Inductive reactance............	$X_L = j\omega L$	$\dfrac{RT}{T} = R$
Capacitive reactance...........	$X_C = \dfrac{-j}{\omega C}$	$\dfrac{-TR}{T} = -R$

Example 23.25

Check the validity of the expression given below by dimensional analysis.

1. $f = \dfrac{1}{2\pi \sqrt{LC}}$; dimensionally,

$$T^{-1} = (RTR^{-1}T)^{-\frac{1}{2}}, \text{ or } T^{-1} = T^{-1}$$

The equation is dimensionally correct.

23.18. Buckingham π Theorem.   The Buckingham π theorem is useful in determining the form of the mathematical relationship between a number of physical quantities.   This is explained by the following combined procedure and illustrative example.

1. *State the problem.*   Determine the expression for the resonant frequency of an LC circuit.

2. *Determine from observation or previous knowledge the parameters upon which the unknown quantity depends.*   Assume f depends upon L and C.

3. *Express all factors in dimensional form from Table* 23.9.

$$f = T^{-1}$$
$$L = RT$$
$$C = R^{-1}T$$

4. *Let m equal the number of different fundamental units (resistance, length, time, and current) required in step 3.*

$$m = 2 \text{ } (R \text{ and } T)$$

5. *Let n equal the total number of factors in the problem.*

$$n = 3 \text{ } (f, L, \text{ and } C)$$

6. *Form the product of the n factors, assigning an exponent to each factor.*   If the exponents for f, L, and C are selected to be u, v, and w, respectively, the product of the factors with their exponents is

$$k = (f)^u (L)^v (C)^w$$

where k signifies the product and is to be dimensionless.

7. *Substitute the dimensional form of the factors from step 3.*

$$k = (T^{-1})^u (RT)^v (R^{-1}T)^w$$

If k is to be dimensionless, the sum of the exponents of each dimensional unit must equal zero.

  8. *Write an equation for each fundamental unit as follows:*
     a. Multiply the exponent of the fundamental unit in each term by the letter exponent assigned to that term.
     b. Add the resulting products from each term.
  For the fundamental unit T,

$$-u + v + w = 0$$

and for R,

$$0 + v - w = 0$$

  9. *A number of exponents given by $(n - m)$ may be assigned arbitrary values because there are fewer equations than unknowns. Care must be taken so that the assigned values do not make the determinant of the equations equal to zero. Solve for the remaining unknowns.*

$$n - m = 3 - 2 = 1$$

Therefore, it is possible to assign a value to u, v, or w. Since an expression for f is desired, let $u = -1$. Solving for v and w,

$$\begin{aligned} v &= w \\ 1 + 2v &= 0 \\ v &= -\tfrac{1}{2} \\ w &= -\tfrac{1}{2} \end{aligned}$$

  10. *Rewrite the equation formed in step 6 with numerical values for exponents as determined in step 9 and solve for the required unknown.*

$$(f)^{-1}L^{-\frac{1}{2}}C^{-\frac{1}{2}} = \text{a constant } k$$

Therefore, $f = \dfrac{k^{-1}}{\sqrt{LC}}$, which is the familiar result with $k = 2\pi$.

23.19. Network Algebra. A considerable portion of the algebra encountered in performing computations for electrical networks arises from the fact that many excitations are either sinusoidal in form or can be resolved into a series of harmonic sinusoidal components. The significant quantities defining a sinusoidal voltage are shown in Fig. 23.47a. The variation in amplitude can be expressed as a function of an angle θ as follows:

$$x = A \cos \theta \qquad (23.142)$$
$$y = A \sin \theta \qquad (23.143)$$
$$\theta = \omega t = \frac{2\pi t}{T}$$
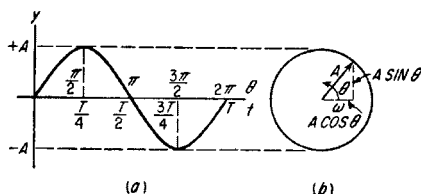
Fig. 23.47. Sinusoidal function.

where T is the period of the sinusoid

It can be seen from Eqs. (23.142) and (23.143) that the quantities x and y can be expressed in an alternate graphical form as the magnitude of the projection on the horizontal and vertical axes, respectively, of a line of length A, rotated an angle θ from the horizontal axis. This representation is shown in Fig. 23.47b.

23.19a. Mathematical Forms of Complex Quantities. In representing sinusoidal electrical quantities by the vector A (line of length A and angular position θ) in Fig. 23.47b, an analytical means is required to rotate the vector around the origin. The mathematical operator j is used for this purpose. If a vector is multiplied by $\pm j$, the vector is rotated $\pm 90°$ and the magnitude of the vector remains unchanged. If a vector C is multiplied twice by j, the resultant vector $j^2 C$ will be rotated $180°$ from the original position of C. Since the negative of a vector quantity is that same quantity reversed in direction,

$$j^2 = -1$$

or
$$j = \sqrt{-1} \qquad (23.144)$$

A third multiplication by j will rotate C $270°$ from the reference axis, denoted by $-jC$.

Multiplication by $+j$ rotates vector quantities counterclockwise. Multiplication by $-j$ rotates clockwise.

Cartesian Form.   A vector quantity can be expressed as the sum of two vectors displaced by an angle of 90° where one component lies along the x axis and the other component lies along the y axis. This may be expressed by

$$A = x + jy \tag{23.145}$$

$$|A| = \sqrt{x^2 + y^2} \tag{23.146}$$

$$\theta_A = \tan^{-1}\frac{y}{x} \tag{23.147}$$

When Eqs. (23.146) and (23.147) are solved for x and y, the following relationships are obtained:

$$x = |A| \cos \theta \tag{23.148}$$

$$y = |A| \sin \theta \tag{23.149}$$

Therefore, since the direction of the vector y is rotated 90° counterclockwise with respect to the vector x, A can be rewritten as

$$A = |A|(\cos \theta + j \sin \theta) \tag{23.150}$$

Where the plus sign rotates A counterclockwise.

Exponential Form.   Vector quantities can be expressed in exponential form by the following relation

$$e^{j\theta} = \cos \theta + j \sin \theta \tag{23.151}$$

therefore,

$$A = |A|(\cos \theta + j \sin \theta) = |A|e^{j\theta} \tag{23.152}$$

Polar Form.   Equation (23.150) may also be expressed in the polar form as follows:

$$A = |A|e^{j\theta} = |A|\underline{/\theta_A} \tag{23.153}$$

23.19b. Mathematical Operations with Complex Quantities.   In performing network calculations, the conventional mathematical operations of addition, subtraction, multiplication, division, taking powers, roots, or logarithms must be performed. The facility with which these operations can be accomplished is largely dependent upon the form chosen for the operation. This is illustrated by the following example: If

$$A = a + ja' = |A|e^{j\theta_A} = |A|\underline{/\theta_A}$$
$$B = b + jb' = |B|e^{j\theta_B} = |B|\underline{/\theta_B}$$
$$C = c + jc' = |C|e^{j\theta_C} = |C|\underline{/\theta_C}$$

then

$$A + B + C = (a + b + c) + j(a' + b' + c') \tag{23.154}$$

$$= |\sqrt{(a + b + c)^2 + (a' + b' + c')^2}|e^{j\theta} \tag{23.155}$$

$$= |\sqrt{(a + b + c)^2 + (a' + b' + c')^2}|\underline{/\theta} \tag{23.156}$$

where $\theta = \tan^{-1}[(a' + b' + c')/(a + b + c)]$

Subtraction

$$A - B = (a - b) + j(a' - b') \tag{23.157}$$

$$= |\sqrt{(a - b)^2 + (a' - b')^2}|e^{j\theta} \tag{23.158}$$

$$= |\sqrt{(a - b)^2 + (a' - b')^2}|\underline{/\theta} \tag{23.159}$$

where $\theta = \tan^{-1}[(a' - b')/(a - b)]$

Multiplication

$$A \cdot B = (ab - a'b') + j(a'b + ab') \tag{23.160}$$
$$= |AB|e^{j(\theta_A + \theta_B)} \tag{23.161}$$
$$= |AB|\underline{/(\theta_A + \theta_B)} \tag{23.162}$$

Division

$$\frac{A}{B} = \frac{ab + a'b'}{b^2 + b'^2} + j\frac{(a'b - b'a)}{b^2 + b'^2} \tag{23.163}$$

$$= \left|\frac{A}{B}\right| e^{j(\theta_A - \theta_B)} \tag{23.164}$$

$$= \left|\frac{A}{B}\right| \underline{/(\theta_A - \theta_B)} \tag{23.165}$$

Powers

$$A^n = (a + ja')^n \tag{23.166}$$
$$= |A|^n e^{jn\theta_A} \tag{23.167}$$
$$= |A|^n \underline{/n\theta_A} \tag{23.168}$$

and

$$A^n B^n = |AB|^n \underline{/n(\theta_A + \theta_B)} \tag{23.169}$$

Roots

$$A^{1/n} = (a + ja')^{1/n} \tag{23.170}$$
$$= |A|^{1/n} e^{i(\theta_A + 2\pi q)/n} \tag{23.171}$$
$$= |A|^{1/n} \underline{/(\theta_A + 2\pi q)/n} \tag{23.172}$$

where $q = 0, 1, \ldots, n - 1$

Logarithms

$$\ln A = \ln |A| + j\theta_A \tag{23.173}$$

The angle θ_A can be increased or divided by integral multiples of 2π.

23.19c. Complex Impedance. In general, the impedance (or admittance) of a network is composed of resistance (or conductance) and reactance (or susceptance). Since a reactance (or susceptance) introduces a $\pm 90°$ phase relationship between the voltage across it and the current through it, it is convenient to express impedance or
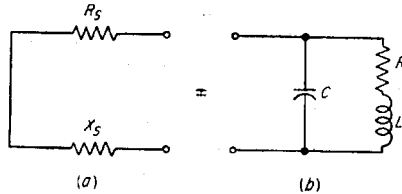
Fig. 23.48. Equivalent networks.

admittance in complex form (that is, $Z = R + jX$). The operations described in the previous section are necessary when network calculations are required involving complex quantities.

To illustrate manipulations with complex quantities, let it be required to determine the equivalent series network (Fig. 23.48a) of the series-parallel network shown in Fig. 23.48b. This may be accomplished by the following steps:

1. Equate the impedance across the terminals of the network shown in Fig. 23.48a to the impedance across the terminals of the network shown in Fig. 23.48b.

$$\frac{(R + j\omega L)(-j/\omega C)}{(R + j\omega L) + (-j/\omega C)} = R_s + jX_s$$

2. Reduce the left-hand expression to real and imaginary components using Eq. (23.163). Equate R_s to the real part and X_s to the imaginary part.

$$\frac{(R + j\omega L)(-j/\omega C)}{(R + j\omega L) + (-j/\omega C)} = \frac{(L/C) - j(R/\omega C)}{R + j(\omega L - 1/\omega C)}$$

$$a = L/C \qquad b = R$$

$$a' = -R/\omega C \qquad b' = (\omega L - 1/\omega C)$$

$$R_e + jX_e = \frac{\left(\dfrac{L}{C}\right)(R) + \left(\dfrac{-R}{\omega C}\right)\left(\omega L - \dfrac{1}{\omega C}\right)}{R^2 + \left(\omega L - \dfrac{1}{\omega C}\right)^2} + j\left[\frac{\left(-\dfrac{R}{\omega C}\right)(R) - \left(\dfrac{L}{C}\right)\left(\omega L - \dfrac{1}{\omega C}\right)}{R^2 + \left(\omega L - \dfrac{1}{\omega C}\right)^2}\right]$$

$$R_e = \frac{R/\omega^2 C^2}{R^2 + \left(\omega L - \dfrac{1}{\omega C}\right)^2}$$

$$X_e = -\left[\frac{\dfrac{R^2}{\omega C} + \left(\dfrac{L}{C}\right)\left(\omega L - \dfrac{1}{\omega C}\right)}{R^2 + \left(\omega L - \dfrac{1}{\omega C}\right)^2}\right]$$

23.20. Fundamentals of Probability and Statistics. Probability theory applied to engineering is the study of the average behavior of physical phenomena where it is not possible or practical to predict the precise behavior of the phenomena. The study of probability and statistics can be separated into two major divisions, viz., the study of the laws of probability where the average behavior of the phenomenon under consideration is assumed to be known, and the study of the methods by which the average behavior of some phenomenon can be determined from observations. The first division is loosely called "probability theory," and the second is frequently referred to as "statistics."

23.20a. Frequency Definition of Probability. In describing many phenomena of engineering and physics, it is not practical or possible to predict exactly the behavior of particular physical events. However, it is often possible to describe certain properties of such phenomena in terms of "average behavior" or in terms of *probabilities*. For example, in the tossing of a coin, it is theoretically possible to predict whether the coin will land "heads up" or "tails up" if enough is known about the conditions under which the coin is tossed, viz., initial position data, initial velocity data, atmospheric data, mass distribution of the coin, etc. However, without this data and without formidable calculations, it is possible to state (assuming the coin is balanced about its axis of rotation) that a coin tossed in a "random manner" will, on the average, land "heads up" in one-half of the tosses, or that the probability of its landing "heads up" is one-half. It is difficult to state exactly what is meant by "a random manner," but it implies that the various conditions of each coin tossing are different and are either unknown and/or unpredictable.

Probability theory is based on the premise that the *frequency of occurrence* of a particular outcome of an experiment tends to a definite limit as the number of repetitions of the experiment increases. Thus, the probability $P(A)$ of a particular outcome A is, in the limit, equal to the frequency of occurrence of A.

$$P(A) = \lim_{N \to \infty} \frac{n_A(N)}{N} \qquad (23.174)$$

where N = total number of experiments or observations

$n_A(N)$ = number of occurrences of A in N experiments or observations

No probability can be determined empirically by making use of this formula since to do so would require the completion of an infinite number of experiments or observations. However, probabilities can be calculated based on the *expected frequencies* of various outcomes.

An approach to constructing a mathematical model for probabilities of various

outcomes is to establish a *sample space* of all possible outcomes. Consider an experiment which has n *basic outcomes* such that none of them can occur simultaneously as a result of performing the experiment. To each of these basic outcomes there is assigned a *sample point* within the sample space. Further, to each outcome there corresponds a certain probability so that to each sample point there also corresponds a probability. Thus, *a sample space consists of a set of sample points each of which is assigned a probability which is equal to the expected frequency of occurrence of the basic outcome corresponding to the sample point.* If each of n basic outcomes is equally likely to occur, the probability assigned to each sample point is $1/n$.

The utility of describing probability processes in terms of sample spaces arises chiefly in computing the probability of any of a number of outcomes occurring in an experiment. If there are n basic outcomes, the probability that any one of m of these $(m \leq n)$ will occur in an experiment is equal to the sum of the probabilities assigned to the m sample points (see Sec. 23.20b).

In computing probabilities of events where there are a finite number of possible outcomes, it is often convenient to make use of either of two formulas. One is a formula giving the number of possible *permutations* of n events, i.e., the number of possible ways in which n events can be *ordered* (arranged). Its value is

$$n! = n(n - 1) \cdots 3 \times 2 \times 1$$

The other formula computes the number of combinations of n events taken r at a time without regard to the order of the r events. For example, for four events a, b, c, d, taken three at a time, the combination of a, c, d is counted only once even though there are six permutations of the three letters, viz., a, c, d; a, d, c; c, a, d; c, d, a; d, a, c; and d, c, a. The number of combinations $\binom{n}{r}$ of n events taken r at a time is

$$\binom{n}{r} = \frac{n!}{r!(n - r)!} \tag{23.175}$$

Example 23.26

In drawing two cards at random from a deck of 52 cards, what is the probability of obtaining an ace and either a face card or a ten?

Solution

1. Each point of the sample space is taken to be a different combination of two cards. The total number of sample points is, therefore, given by the formula for $\binom{n}{r}$, the combination of n things taken r at a time, where $n = 52$ and $r = 2$.

$$\binom{n}{r} = \frac{n!}{r!(n - r)!}$$
$$= \frac{52!}{2!50!} = \frac{52 \times 51}{2 \times 1} = 1{,}326$$

Since each outcome is equally probable, the probability assigned to each sample point is $1/1{,}326$.

2. The number of sample points corresponding to the possible combinations of an ace and either a face card or a ten is $4 \times 16 = 64$ since there are 4 aces and a total of 16 face cards and tens in a deck of 52 cards.

3. The probability of obtaining an ace and either a face card or a ten is the sum of the probabilities of the sample points corresponding to these combinations, or

$$64 \times \frac{1}{1{,}326} \simeq 0.04827$$

23.20b. Basic Rules for Combinations of Events. If the outcomes of an experiment are "basic outcomes" as defined for a sample space representation (Sec. 23.20a), the probability of one of several basic outcomes occurring is the sum of the probabilities for each of these outcomes. The probability of the joint occurrence of more than one basic outcome is zero. Although it is always possible to state a probability problem in terms of basic outcomes, it is often more convenient to make use of the *addition* and *multiplication theorems.*

To compute the probability of one or more events occurring, the *addition theorem* of probability is useful. For two events A and B (which are not necessarily "basic outcomes" and, therefore, may occur jointly), the probability $P(A$ or $B)$ of either A occurring or B occurring or both A and B occurring is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \qquad (23.176)$$

where $P(A)$ = probability of A occurring disregarding the occurrence of B

$\quad\;\; P(B)$ = probability of B occurring disregarding the occurrence of A

$P(A$ and $B)$ = probability of *both* A and B occurring jointly

Equation (23.176) is derived by considering a sample space to be made up of basic outcomes which are all possible pairs of events occurring of which the combination A and B is one pair. $P(A)$ is the sum of the probabilities assigned to sample points representing the event A combined with any other event. $P(B)$ is defined similarly. $P(A$ and $B)$ is the probability assigned to the sample point representing the combination of events A and B. Both $P(A)$ and $P(B)$ include the probability $P(A$ and $B)$ so that it must be subtracted from the sum of $P(A)$ and $P(B)$ in order to obtain Eq. (23.176). In the case where $P(A$ and $B) = 0$, the events A and B are said to be *mutually exclusive.*

For three events A, B, and C,

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C) - P(A \text{ and } B) - P(B \text{ and } C)$$
$$- P(C \text{ and } A) + P(A \text{ and } B \text{ and } C) \quad (23.177)$$

Conditional probability is the probability of one event occurring after it is known that another event has occurred. For two events A and B, $P(B|A)$ denotes the probability of B occurring after it is known that A has occurred, or simply, "B after A." Using this definition leads to the *multiplication theorem* of probability, viz.

$$P(A \text{ and } B) = P(A)P(B|A)$$
or $$\qquad\qquad\qquad = P(B)P(A|B) \qquad\qquad (23.178)$$

For three events A, B, C,

$$P(A \text{ and } B \text{ and } C) = P(A)P(B|A)P(C|A \text{ and } B)$$
$$= P(B)P(C|B)P(A|B \text{ and } C), \text{ etc.} \qquad (23.179)$$

If $P(B|A) = P(B)$ or if $P(A|B) = P(A)$, the two events A and B are said to be *mutually independent.*

Example 23.27

In a random drawing of two cards from a deck of 52 cards, what is the probability of drawing either the queen of spades or the ace of hearts or both?

Solution

1. In the drawing of two cards, each has the same probability of being the queen of spades, viz., $\frac{1}{52}$. The probability of drawing it on either the first or second draw is, by Eq. (23.176),

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

where A = occurrence of queen on first draw

$\quad\;\; B$ = occurrence of queen on second draw

Since A and B are mutually exclusive, i.e., the probability of drawing it on both draws $P(A \text{ and } B) = 0$, it follows that $P(A \text{ or } B) = \frac{1}{52} + \frac{1}{52} = \frac{1}{26}$.

2. Likewise, the probability of drawing the ace of hearts as one of the two cards is $\frac{1}{26}$.

3. The probability $P(C \text{ and } D)$ of drawing both the queen and the ace is, by Eq. (23.178),

$$P(C \text{ and } D) = P(C)P(D|C)$$

where C = drawing of queen
      D = drawing of ace

The probability $P(D|C)$ of one of the two cards being the ace if the other is known to be the queen is $\frac{1}{51}$. Therefore,

$$P(C \text{ and } D) = \frac{1}{26} \times \frac{1}{51} = \frac{1}{1{,}326}$$

4. The probability $P(C \text{ or } D)$ of drawing either the ace or the queen or both is then

$$P(C \text{ or } D) = P(C) + P(D) - P(C \text{ and } D)$$
$$= \frac{1}{26} + \frac{1}{26} - \frac{1}{1{,}326} = \frac{101}{1{,}326} \simeq 0.07617$$

23.20c. Binomial Probability Distribution; Probability Density Functions. In many engineering applications, the question arises as to the probability of an event A occurring exactly r times if an experiment is repeated n times where $n \geq r$. When the results of each experiment are independent, this probability is given by the *binomial distribution.* Assume that the probability of a certain outcome A of an experiment is defined as $P(A) = p$. Then, the probability of A not occurring can be defined as $1 - P(A) = 1 - p = q$. The binomial distribution defines the probability $P(\nu = r)$ of A occurring exactly r times in n experiments to be

$$P(\nu = r) = \binom{n}{r} p^r q^{n-r} \qquad (23.180)$$

where ν = number of occurrences of A

$\binom{n}{r}$ = number of combinations of n things taken r at a time [see Eq. (23.175)]

The name "binomial distribution" arises from the fact that the coefficient $\binom{n}{r}$ corresponds to the coefficient of the rth term in the binomial expansion of $(p + q)^n$. *The binomial distribution applies to any set of repeated independent trials where for each trial there are two possible outcomes of probability p and $1 - p$.* Such trials are called *Bernoulli trials.*

A probability distribution, or more precisely, a (discrete) *probability density function*, associated with this binomial process arises when not just one value of ν, but all possible values of ν, are considered for a given number of samples n. For example, a graphical representation of the binomial probability density function for $n = 5$ and $p = 0.6$ is shown in Fig. 23.49. The density function attains its maximum value at or near $r = np$.
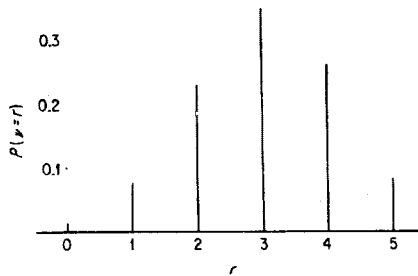
Fig. 23.49. Binomial-distribution density function for $n = 5$ and $p = 0.6$.

Probability density functions are a convenient way (but not the only way) to represent statistical processes. They represent the probabilities associated with all possible outcomes of a particular experiment. Thus, regardless of the (discrete) distribution where there are $n + 1$ possible outcomes for ν,

$$\sum_{r=0}^{n} P(\nu = r) = 1 \tag{23.181}$$

Example 23.28

A manufacturer has found that an electronic assembly employing 10 diodes will not pass production inspection if the cathode emission of six or more of the diodes is below a certain critical value even though the diodes meet a vendor's specification. The probability of a diode being below this critical value has been determined to be 0.1. What is the probability that the manufacturer will have to reject a unit if he does not select the diodes before installation?

Solution

1. The probability p of a diode being defective is 0.1; therefore, $q = 0.9$.
2. The probability $P(R)$ of rejecting an assembly (by the addition theorem of Sec. 23.20b) is the sum of the probabilities associated with the occurrence of 6 defective diodes in a sample of 10, 7 in 10, . . . , 10 in 10. Therefore,

$$P(R) = \sum_{r=6}^{10} P(\nu = r)$$

where ν = number of defective diodes [(see Eq. (23.180)]

$$P(R) = \binom{10}{6} (0.1)^6 (0.9)^4 + \binom{10}{7} (0.1)^7 (0.9)^3 + \binom{10}{8} (0.1)^8 (0.9)^2$$
$$+ \binom{10}{9} (0.1)^9 (0.9)^1 + \binom{10}{10} (0.1)^{10} (0.9)^0$$
$$= (210)(0.1)^6 (0.9)^4 + (120)(0.1)^7 (0.9)^3 + (45)(0.1)^8 (0.9)^2 + (10)(0.1)^9 (0.9)^1$$
$$+ (0.1)^{10}$$
$$\simeq 0.000,146,9$$

On the average, approximately one out of every 6,807 units will be rejected.

Note that the statement of the problem tacitly assumes that the unit will pass inspection if five diodes just meet the vendor's specification and the other five diodes are just above the critical value even though it will not pass inspection if six of the diodes have emission just below the critical value. If this assumption is not satisfied, the problem can be considerably more complicated.

23.20d. Poisson Distribution. The Poisson probability distribution is chiefly used in two different applications. One use is as an approximation to the binomial distribution for certain cases. The other use is an important set of problems where the Poisson distribution is an exact probability representation for certain phenomena.

1. *The Poisson Distribution as an Approximation to the Binomial Distribution.* As stated in Sec. 23.20c, the binomial distribution deals with the probability of a certain number of events occurring in a sample of given size. The *Poisson distribution* deals with the same problem *in an approximate manner* when the sample size is much larger than the number of events for which the probability is being calculated $(n \gg r)$ and the probability p of each event is small. It is easily derived from the binomial distribution [Eq. (23.180)] by making suitable approximations.

$$P(\nu = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{(n-r)} \tag{23.180}$$
$$\simeq \frac{n^r}{r!} p^r (1-p)^n \qquad (n \gg r)$$
$$\simeq \frac{n^r}{r!} p^r e^{-pn} \qquad (p \ll 1)$$
$$\simeq \frac{(pn)^r}{r!} e^{-pn} \tag{23.182}$$

Example 23.29

Compute the probability of finding not more than one defective assembly due to substandard diodes for a production run of 200 of the assemblies described in Example 23.28. Use both the exact binomial distribution and the approximate Poisson distribution to compute the result.

Solution

1. *Binomial distribution*

a. From Example 23.28, the probability p of an assembly being defective is 0.000,146,9 so that $q = 0.999,853,1$.

b. The probability $P(G)$ of not more than one assembly out of 200 being defective is, by Eq. (23.180),

$$P(G) = \sum_{r=0}^{1} P(\nu = r)$$

$$= \binom{200}{0}(0.0001469)^0(0.9998531)^{200} + \binom{200}{1}(0.0001469)^1(0.9998531)^{199}$$

$$= 0.999579$$

2. *Poisson distribution approximation*

$$pn = 0.000,146,9 \times 200 = 0.02938$$

$$\therefore P(G) \simeq \frac{(0.02938)^0}{0!}e^{-0.02938} + \frac{(0.02938)}{1!}e^{-0.02938}$$

$$\simeq e^{-0.02938}(1 + 0.02938)$$

$$\simeq 0.999577$$

For such a small probability and a large ratio between the number of samples (200) compared with the number of events (2) for which the probability was calculated, the Poisson approximation is very close.

2. *The Poisson Distribution as an Exact Distribution.* Note that in Eq. (23.182) both the probability p and the number of samples n appear only in the product pn. As is discussed in Sec. 23.20h, *this product is the average or expected number of events with probability p which occur in n samples.* Thus, Eq. (23.182) gives the probability of r events occurring when the average number of events in n trials is known, assuming that n is large compared to r and p is small.

This interpretation of the Poisson distribution can be extended to an important set of problems. In particular, consider the problem of computing the probability

Fig. 23.50. Poisson-distribution density function for $\lambda T = 2.5$.

of an event occurring in an interval of time T when it is known that, on the average, λ events occur per second. In this case, the total interval T may be thought of as broken up into a very large number n of subintervals each of length Δt so that $n = T/\Delta t$. Then if $p_{\Delta t}$ is the probability of the event occurring in an interval of length Δt, the expected number of occurrences of the event in n samples, i.e., in an interval of length T, is $p_{\Delta t}n = p_{\Delta t}T/\Delta t$ which is equal to λT according to the definition of λ if Δt is small enough. Allowing Δt to approach zero results in n approaching ∞ and $p_{\Delta t}$ approaching zero. Since the only approximations made in deriving Eq. (23.182) were that $n \gg r$ and $p \ll 1$, it follows that the formula is exact in the limit as $n \to \infty$ and $p_{\Delta t} \to 0$. In other words, Eq. (23.182) becomes exact with λT replacing pn as $\Delta t \to 0$. Thus, *the probability of exactly r events occurring in an interval*

of T seconds, where it is known that λ events occur per second on the average, is

$$P(\nu = r) = \frac{(\lambda T)^r}{r!} e^{-\lambda T} \qquad (23.183)$$

A plot of the Poisson distribution for $\lambda T = 2.5$ is shown in Fig. 23.50.

Example 23.30

If it has been established that an electronic computer has an equipment failure which invalidates the answers from the machine at an approximate average rate of once every 10 hr, what is the probability of obtaining invalid answers from the machine for a problem which takes 1 hr to run?   What is the probability for a problem which takes 10 hr to run?

Solution

Note: The probability of at least one failure is equal to one minus the probability of no failure.

1. $\lambda = 0.1$ failure per hour
       $T_1 = 1$ hr, $T_2 = 10$ hr

2. From Eq. (23.183),

$$P_1 \text{ (no failure in 1 hr)} = \frac{(0.1)^0}{0!} e^{-0.1} = 0.9048$$

$$P_2 \text{ (no failure in 10 hr)} = \frac{(1)^0}{0!} e^{-1} = 0.3679$$

3. Therefore, the probability P_1 of a failure which invalidates the answers in 1 hr is

$$P_1 \text{ (failure in 1 hr)} = 1 - 0.9048 = 0.0952$$

For 10-hr operation, the probability P_2 of invalid answers is

$$P_2 \text{ (failure in 10 hr)} = 1 - 0.3679 = 0.6321$$

23.20e. The Normal Distribution as an Approximation to the Binomial Distribution. The Poisson distribution is a good approximation to the binomial distribution when the sample size is much larger than the number of events for which the probability is being calculated and the probability of each event occurring is small. *The normal distribution, on the other hand, is a good approximation to the binomial distribution when the sample size is large regardless of the number of events being considered.* The approximation is best when the probabilities p and q are 0.5.

The normal distribution approximation can be derived from the binomial distribution and is found to be

$$P(\nu = r) = \binom{n}{r} p^r q^{n-r} \simeq \frac{\phi(x)}{\sqrt{npq}} \qquad (23.184)$$

where
$$\phi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

$$x = \frac{r - np}{\sqrt{npq}}$$

The function $\phi(x)$ is the *normalized normal distribution density function.* A plot comparing $\binom{n}{r} p^r q^{n-r}$ and $\phi(x)/\sqrt{npq}$ (Fig. 23.51a and b) shows how close the normal approximation is for $n = 10$, $p = 0.1$ and for $n = 10$, $p = 0.6$. For a fixed sample size the normal approximation is better for p closer to 0.5 than to either 0 or 1. In the latter cases, the binomial distribution is unsymmetrical whereas the normal distribution is always symmetrical. The maximum value of the normal density function always occurs at $x = 0$, that is, $r = np$.

The usefulness of the normal approximation comes about when, as is often the case, a probability is to be computed which consists of a summation of terms of the binomial

distribution. This application of the normal distribution may be seen by plotting a binomial distribution density function as a series of step or jump functions where the jumps occur at $-\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, \ldots, n - \frac{1}{2}, n + \frac{1}{2}$, and the normal distribution density function as shown in Fig. 23.51b. The area included by two successive jumps is then equal to a term of the binomial distribution. Moreover, each of these areas is approximated by the area under the normal distribution density function between the abscissa
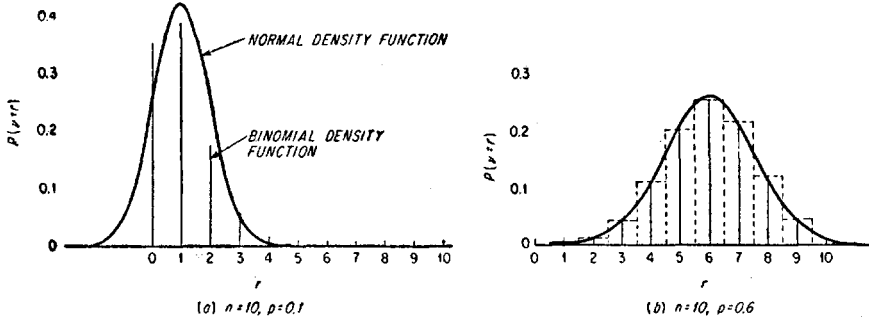
FIG. 23.51. Normal distribution density functions compared to binomial distribution density functions.

values $(-\frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, \frac{3}{2}), \ldots, (n - \frac{1}{2}, n + \frac{1}{2})$. Thus, the normal approximation to the binomial distribution on the basis of area included between these jump points is

$$P(\nu = r) = \binom{n}{r} p^r q^{n-r} \simeq \frac{1}{\sqrt{npq}} \int_{r-\frac{1}{2}}^{r+\frac{1}{2}} \phi(x)\, dr$$

Since

$$dx = \frac{dr}{\sqrt{npq}}$$

$$P(\nu = r) \simeq \int_{x_r^-}^{x_r^+} \phi(x)\, dx \qquad (23.185)$$

where

$$x_r^+ = \frac{(r + \frac{1}{2}) - np}{\sqrt{npq}}$$

$$x_r^- = \frac{(r - \frac{1}{2}) - np}{\sqrt{npq}}$$

Numerical tables for the normal distribution generally give computed values of $\int_{-\infty}^{a} \phi(x)\, dx$ so that to compute results from Eq. (23.185) it is only necessary to make use of the identity

$$\int_{b}^{a} \phi(x)\, dx = \int_{-\infty}^{a} \phi(x)\, dx - \int_{-\infty}^{b} \phi(x)\, dx \qquad (23.186)$$

From Eq. (23.185) it follows that the summation of consecutive terms of a binomial distribution is approximated by

$$\sum_{r=r_1}^{r_2} \binom{n}{r} p^r q^{n-r} \simeq \int_{x_{r_1}^-}^{x_{r_2}^+} \phi(x)\, dx \qquad (23.187)$$

where

$$x_{r_2}^+ = \frac{(r_2 + \frac{1}{2}) - np}{\sqrt{npq}}$$

$$x_{r_1}^- = \frac{(r_1 - \frac{1}{2}) - np}{\sqrt{npq}}$$

Example 23.31

In a factory producing transistors, it has been found over a long period of time that 20 per cent of the transistors do not meet acceptance specifications and must be rejected. Assuming that the rejections are completely random, how many acceptable transistors must be on hand at the beginning of a week to provide a probability of 0.999 of being able to deliver at least 2,000 transistors at the end of the week if the number of transistors going into production is 2,500 for the week?

Solution

1. $n = 2,500$, $p = 0.8$, $q = 0.2$.
2. The problem is to find r_1 such that for $r_2 = 2,500$,

$$\int_{x_{r_1}^-}^{x_{r_2}^+} \phi(x)\,dx = 0.999$$

3. $x_{r_1}^+ = \dfrac{2,500 + \frac{1}{2} - (2,500)(0.8)}{\sqrt{2,500(0.8)(0.2)}} = \dfrac{500.5}{20} = 25.025$

$$0.999 = \int_{x_{r_1}^-}^{25.025} \phi(x)\,dx = \int_{-\infty}^{25.025} \phi(x)\,dx - \int_{-\infty}^{x_{r_1}^-} \phi(x)\,dx = 1.0000 - \int_{-\infty}^{x_{r_1}^-} \phi(x)\,dx$$

Therefore,

$$\int_{-\infty}^{x_{r_1}^-} \phi(x)\,dx = 0.001$$

From tables of the normal distribution, it is found that

$$x_{r_1}^- = -3.09$$

or

$$-3.09 = \dfrac{r_1 - \frac{1}{2} - np}{\sqrt{npq}} = \dfrac{r_1 - \frac{1}{2} - 2,000}{20}$$

$$r_1 = 1938.7$$

Therefore, if there are $2,000 - 1,938 = 62$ transistors on hand at the beginning of the week, the probability will be greater than 0.999 of being able to deliver 2,000 transistors at the end of the week. Note that this example tacitly assumes that the probability of a major production breakdown, i.e., one which will not allow the building of 2,500 transistors, is $\ll 0.001$.

23.20f. Discrete versus Continuous (Cumulative) Distribution Functions. The probability distributions which have been discussed in the preceding sections have all been of the *discrete* type; i.e., the distributions have applicability where there are either a finite number or an *enumerable* number of possible outcomes. (An enumerable number of possible outcomes is an infinite number of alternatives each of which is assigned a positive integer. The exact Poisson distribution is an example of a distribution involving an enumerable number of possible outcomes.) A great many probability problems are not directly solvable by the use of discrete distributions, but are solvable by using *continuous distribution* models.

A discrete distribution can be represented as either a series of spikes as shown in Fig. 23.52a or as a series of steps as shown in Fig. 23.52b. Instead of showing the probability of ν equaling r, $P(\nu = r)$, Fig. 23.52b shows the probability of ν being less than r, that is $P(\nu < r)$. This representation is possible since the sum of $P(\nu = r)$ for all values of r is equal to 1. Figure 23.52a represents a (discrete) probability density function; Fig. 23.52b represents a (discrete) *probability distribution function* or (discrete) *distribution function*. Density functions are often called *frequency functions;* distribution functions are often called *cumulative distribution functions.* The probability $P(\nu = r)$ in terms of the distribution function representation is

$$P(\nu = r) = P(\nu < r + \epsilon) - P(\nu < r - \epsilon)$$

where ϵ is a small quantity (smaller than the distance between successive jumps of the distribution function). A discrete distribution function is a discontinuous function since for some values of r, $\lim_{\epsilon \to 0} P(\nu < r + \epsilon) \neq \lim_{\epsilon \to 0} P(\nu < r - \epsilon)$.

The idea of a probability distribution function is easily extended to continuous functions by assuming the probability $P(\nu < r)$ varies continuously as a function of r instead of in discrete jumps. For example, consider the problem of determining the
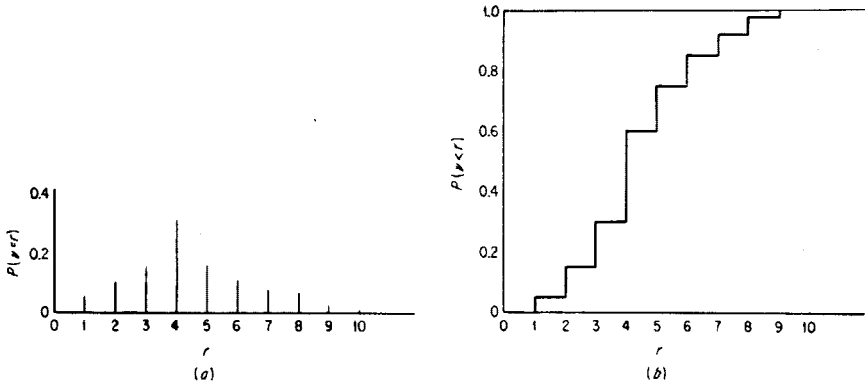
FIG. 23.52. Discrete probability density and distribution function representation.

distribution function describing the probability of the output of a single-turn continuously rotatable linear potentiometer being less than a specified voltage if the potentiometer is rotated in a random manner and is connected across a 10-volt supply. At any particular time, there is no reason to consider any particular output being more probable than any other if nothing is known about the manner in which the potentiometer is rotated. Therefore (neglecting the gap between the two ends of the potentiometer which must inevitably occur), the distribution function

$$P(\xi < x) = F(x)$$

is a straight line as shown in Fig. 23.53a. Alternatively, such a distribution could be shown as a (continuous) probability density function $f(x)$ (Fig. 23.53b) where
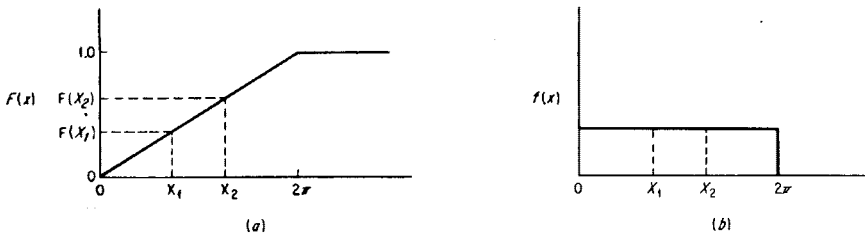
FIG. 23.53. Probability distribution function $F(x)$ and density function $f(x)$ for a uniformly distributed angle of rotation of a potentiometer.

$f(x) = dF(x)/dx$. For a probability density function, the area under the curve between $x = x_1$ and $x = x_2$, represents the probability that the output voltage will correspond to an angle between x_1 and x_2. For the distribution function representation, the same probability is $F(x_2) - F(x_1)$.

Both discrete and continuous distribution functions vary between 0 and 1. Moreover, all distribution functions (discrete as well as continuous) are nondecreasing

functions of x. Represented mathematically, the following relations involving a distribution function $F(x)$ and a corresponding density function $f(x)$ are obtained.

$$F(x) = P(\xi < x) = \int_{-\infty}^{x} f(\xi)\,d\xi \qquad (23.188)$$

$$f(x)(=)\frac{dF(x)}{dx} \qquad (23.189)$$

$$0 \leq F(x) \leq 1 \qquad (23.190)$$

$$f(x) \geq 0 \qquad (23.191)$$

$$F(\infty) = \int_{-\infty}^{\infty} f(x)\,dx = 1 \qquad (23.192)$$

The parentheses around the equality in Eq. (23.189) indicate that if $F(x)$ possesses a discontinuity, i.e., a jump at some value of x, the equation no longer has a direct significance since $f(x)$ is infinite at that value of x.

23.20g. The Normal Distribution as an Exact Distribution. By far the most useful of all continuous distributions is the *normal distribution* (often called the *Gaussian distribution*). It may or may not arise directly from the binomial distribution. One reason that it has such general applicability follows from the central limit theorem which is discussed in Sec. 23.20m.

The *normalized normal probability density function* $\phi(x)$ and the *normalized normal probability distribution function* $\Phi(x)$ are defined by the following equations:

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\,e^{-x^2/2} \qquad (23.193)$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\xi^2/2}\,d\xi = \int_{-\infty}^{x} \phi(\xi)\,d\xi \qquad (23.194)$$

In the equations, x is a normalized variable. To make the distribution generally applicable, x can be written in terms of another variable y by a translation of the origin of y and a change of scale factor, i.e.,

$$x = \frac{y - m}{\sigma} \qquad (23.195)$$

where m = *mean* or *expected value* of y, that is, value about which normal distribution
    is centered (see Sec. 23.20h)

  σ = *standard deviation* of y or roughly a measure of how peaked is the distribution of y (for example, the probability that is between $m - \sigma$ and $m + \sigma$ is approximately 0.683)

When using tables of the normal distribution to obtain numerical results, it is necessary to make use of Eq. (23.195) to obtain the normalized variable found in the tables. It is easily shown that the integral of $\phi(x)$ between the limits of $-\infty$ and ∞ is equal to 1, that is, $0 \leq \Phi(x) \leq 1$.

One of the applications of the normal distribution is in describing the probable amplitude of thermal noise in electronic equipment. The noise amplitude distribution is accurately described by Eqs. (23.193) to (23.195), where m is zero and σ is the rms value of noise voltage, that is, $\sigma = \sqrt{4KT\,\Delta fR}$, where $K = 1.38 \times 10^{-23}$ joule/°K (Boltzmann's constant), T is the absolute temperature of the circuit producing the noise in degrees Kelvin, R is the resistive component of the circuit impedance in ohms, and Δf is the bandwidth of the circuit in cycles per second.

Other important applications of the normal distribution include the analyses of many types of errors in experimental measurements and in evaluating the performance of equipment and systems.

Example 23.32

The rms output voltage of a radar receiver i-f amplifier due to thermal noise is equal to 0.5 volt. Suppose that a "signal present" detector is employed to indicate the presence of a bipolar voltage at the output of the receiver greater than 2.0 volts or less than -2.0 volts. What percentage of the time, on the average, will the detector indicate the presence of a voltage if only noise is present?

Solution

1. Assume a normal distribution with $\sigma = 0.5$ volts. The problem is to calculate $\Phi(\infty)$ − $\Phi(x_2) + \Phi(x_1) - \Phi(-\infty) = 1 - \Phi(x_2) + \Phi(x_1)$, where $x_2 = (2.0 - 0)/0.5 = 4.0$ and $x_1 = (-2.0 - 0)/0.5 = -4.0$.

2. From a table of $\Phi(x)$

$$1 - \Phi(x_2) + \Phi(x_1) = 1 - 0.99997 + 0.00003 = 0.00006$$

3. According to the frequency definition of probability, the average percentage of the time the detector will indicate a voltage with only noise being present will be 100×0.00006, or 0.006 per cent of the time.

Example 23.33

Two receivers tuned to the same frequency contain "signal present" detectors with detection levels set at 2.7 times the value of rms noise. What is the probability at any specific instant of both receivers indicating output voltages due to thermal noise which is greater than the threshold detection levels if the sources of noise in the two receivers are mutually independent?

Solution

1. By Eq. (23.178), the probability of both receivers indicating above-threshold voltages is the product of the individual probabilities for each receiver since the noise sources are assumed independent.

2. At any particular time, the probability of one receiver indicating an above-threshold voltage is

$$1 - \Phi(2.7) + \Phi(-2.7) = 1 - 0.99653 + 0.00347 = 0.00694$$

3. The probability of both receiver outputs simultaneously exceeding the threshold is then $(0.00694)(0.00694) = 0.000048$.

23.20h. Random Variables and Expectations. Any quantity which may assume different values, each of which is assigned a probability of occurrence, is called a *random variable.* Thus, a random variable is a function of a particular sample space since, to each point in a sample space, a value of the random variable, and hence a probability, is assigned. For example, in selecting five resistors from a production run of resistors, any resistor could be either inside or outside a set of specifications. A sample space could be made up of 2^5 points, each point of which could represent a particular combination of good and bad resistors, i.e., a particular "state" of good and bad resistors. (For example, R_1 good, R_2 good, R_3 bad, R_4 good, R_5 bad would be one possible "state" of the resistors.) A random variable could be defined on this sample space as the number of out-of-tolerance resistors represented by each sample point. Alternatively, another random variable could be defined by assigning the number 1 to each good resistor and the number 0 to each defective resistor in the sample space. In this case, the random variable would be five-dimensional. [For example, one value for the random variable would be (0, 1, 0, 0, 0).] In the first case, the random variable would take on one of the six values 0, 1, 2, 3, 4, 5 for each point of the sample space. In the second case, there are 2^5 values for the random variable, a different value for each point of the sample space. Since, by definition, each point in a sample space is assigned a certain probability, it is possible to calculate a probability for each of the values of either of these random variables or any other random variable which might be defined. Random variables are defined on sample spaces which are continuous (i.e., where continuous probability distributions apply) in the same way that they are defined for discrete sample spaces. They can again be one-dimensional or multidimensional.

One of the most fundamental ideas of probability theory is that of the *average value* of a random variable. It is also called the *mean*, or the *expected value*, of a random variable. Going back to the frequency definition of probability, each point r ($r = 0, 1, \ldots, N$) of a sample space is assigned a probability P_r in accordance with its expected frequency of occurrence. If, then, each sample point r is assigned a value of a random variable x_r, it follows that the *expected value* of a random variable $\mathbf{E}(x)$, or $\bar{x}$, is the sum of the products of the random variable values x_r and probabilities $P_r = P(x_r)$ assigned to each point.

$$\mathbf{E}(x) = \bar{x} = \sum_{r=0}^{N} x_r P(x_r) \tag{23.196}$$

or, in terms of continuous distributions,

$$\mathbf{E}(x) = \bar{x} = \int_{-\infty}^{\infty} x f(x) \, dx \tag{23.197}$$

where $f(x)$ is the probability density function for the random variable x.

Example 23.34

A Geiger counter counts an average of 50,000 counts per second from a radioactive sample. What is the expected average rate of gamma-ray production if the counter has a dead time of 1 μsec?

Solution

1. Compute the probability of one or more gamma rays being produced in a 1-μsec period after another has been produced. The production of one gamma ray is assumed to be independent of another (Poisson distribution), so that the probability of one or more gamma rays being produced in a 1-μsec period following another is the same as the probability of one or more gamma rays being produced in any 1-μsec period. According to the Poisson distribution, the probability $P(r)$ of r gamma rays occurring in a period T where the rate is λ is

$$P(\nu = r) = P(r) = \frac{(\lambda T)^r}{r!} e^{-\lambda T}$$

where λ = number of gamma rays per second
$T = 10^{-6}$

2. The sample space for this problem can be thought of as an infinite (enumerable) number of points; for each point r ($r = 0, 1, 2, \ldots, \infty$) there corresponds r gamma rays occurring in 1 μsec. The random variable x_r assigned to the point r is then the value 50,000r missed gamma rays per second. The expected number of missed gamma rays $\mathbf{E}(x_r)$ is then

$$\mathbf{E}(x_r) = \sum_{r=0}^{\infty} x_r P(r)$$

$$= \sum_{r=0}^{\infty} 50{,}000r \frac{(\lambda T)^r}{r!} e^{-\lambda T}$$

$$= 0 + \sum_{r=1}^{\infty} 50{,}000r \frac{(\lambda T)^r}{r!} e^{-\lambda T}$$

$$= (50{,}000)(\lambda T) e^{-\lambda T} \sum_{r=1}^{\infty} \frac{(\lambda T)^{r-1}}{(r-1)!}$$

$$= (50{,}000)(\lambda T) e^{-\lambda T} \cdot e^{\lambda T} = 50{,}000 \lambda T = 0.0\ell \lambda$$

$$\therefore \lambda = 50{,}000 + 0.05\lambda; \quad \lambda = 52{,}632$$

The expected rate of gamma-ray production is approximately 52,632 per second. Notice that this is the same as the answer obtained by observing that the average missing period per second is 0.05 sec so that gamma rays are produced at a rate of 50,000 per 0.95 sec or 50,000/0.95 $\simeq$ 52,632/sec.

Example 23.35

Neglecting losses, what is the expected value of the d-c component of output voltage from a full-wave diode detector if the thermal noise input voltage is 1.0 volt rms?

Solution

1. The distribution of the input noise is normal. At the output, the distribution is unilaterally normal, i.e., the density function $f(x)$ is

$$f(x) = \begin{cases} 0 & (x < 0) \\ 2 \times \dfrac{1}{\sigma \sqrt{2\pi}} e^{-x^2/2\sigma^2} & (x \geq 0) \end{cases}$$

where σ is the rms value of input noise voltage, viz., 1.0 volt. The reason for the 2 is that the integral of the density function between $= -\infty$ and ∞ must be equal to 1.

2. From Eq. (23.197), the expected value of the distribution is

$$\begin{aligned} E(x) &= \int_{-\infty}^{\infty} x f(x)\, dx \\ &= 2 \int_{0}^{\infty} \frac{x}{\sigma \sqrt{2\pi}} e^{-x^2/2\sigma^2}\, dx \\ &= 2 \times \frac{\sigma}{\sqrt{2\pi}} = \sigma \sqrt{\frac{2}{\pi}} \simeq 0.7979\sigma \end{aligned}$$

Therefore, for 1-volt-rms thermal noise into a full-wave diode detector, the d-c component of the output is 0.7979 volt. This relation is true regardless of the spectral distribution of the input noise as long as the mean, i.e., expected value, of the input noise is zero.

23.20i. Moments of Distribution and Moment Generating Functions. A direct analogy can be made between many properties of probability density functions and mass density functions. Considering a one-dimensional probability density function, the analogy is generated by assuming that a unit value of mass is distributed along a line in such a way that distance along the line represents a value of a random variable and the mass density at any point is equal to the probability density for the value of the random variable corresponding to that point. For a discrete distribution, the mass is divided into "mass points"; for a continuous distribution, the mass is spread out and not lumped at discrete points. This analogy may be expanded to multi-dimensional probability distributions by assuming that coordinates of a mass distribution are analogous to coordinate values of a random variable.

Using this analogy, *the expected value of a random variable is analogous to the center of gravity of a mass distribution.* The analogy can be extended by investigating the analog in probability distributions of higher-order moments in mass distributions, e.g., moments of inertia. For a one-dimensional mass distribution $M(x)$ in the continuous case, the moment of inertia I about an axis through a point c on the line is

$$I = \int_{-\infty}^{\infty} (x - c)^2 M(x)\, dx \qquad (23.198)$$

For a one-dimensional mass distribution $M(x_r)$ (where $r = 0, 1, 2, \ldots, N$) in the discrete case, the moment of inertia I about an axis through a point c on the line is

$$I = \sum_{r=0}^{N} (x_r - c)^2 M(x_r) \qquad (23.199)$$

The analogous equations for probability distributions are

$$\mathbf{E}[(x - c)^2] = \int_{-\infty}^{\infty} (x - c)^2 f(x) \, dx \qquad (23.200)$$

and
$$\mathbf{E}[(x - c)^2] = \sum_{r=0}^{N} (x_r - c)^2 P(x_r) \qquad (23.201)$$

where $f(x)$ and $P(x_r)$ are continuous and discrete probability density functions, respectively.

Equations (23.200) and (23.201) give the *expected values* of $(x - c)^2$ and are called second moments of the probability distributions $f(x)$ and $P(x_r)$ *centered about c.* If $c = 0$, these expected values often are simply termed *second moments.*

In mechanics, it is well known that the moment of inertia for a given mass distribution is a minimum if the axis of rotation is placed through the center of gravity of the mass aggregate. Analogously, in probability theory, the second moment has special significance and is minimized if it is centered about the expected value of the random variable. The second moment centered about the expected value $\mathbf{E}(x) = m$ is called the *variance*, or *second central moment.* For continuous and discrete probability distributions the variance σ^2 is defined by the following equations:

$$\sigma^2 = \int_{-\infty}^{\infty} (x - m)^2 f(x) \, dx \qquad (23.202)$$

$$\sigma^2 = \sum_{r=0}^{N} (x_r - m)^2 P(x_r) \qquad (23.203)$$

where
$$m = \mathbf{E}(x) = \int_{-\infty}^{\infty} x f(x) \, dx$$

$$m = \mathbf{E}(x) = \sum_{r=0}^{N} x_r P(x_r)$$

The square root of the variance is called the *standard deviation σ* of the random variable (see Sec. 23.20g). It is a good measure of the relative concentration of the distribution about the mean.

Moments of higher order than the second can be defined in the same way as the first- and second-order moments. For a continuous distribution, the nth *moment* α_n is given by Eq. (23.204). The nth *central moment* μ_n is given by Eq. (23.205). The corresponding moments for discrete distributions are similar in form to Eqs. (23.201) and (23.203).

$$\alpha_n = \int_{-\infty}^{\infty} x^n f(x) \, dx \qquad (23.204)$$

$$\mu_n = \int_{-\infty}^{\infty} (x - m)^n f(x) \, dx \qquad (23.205)$$

From these equations, it follows that $\alpha_0 = \mu_0 = 1$ and $\alpha_1 = m$, $\mu_1 = 0$. It is frequently convenient to compute the variance $\mu_2 = \sigma^2$ in terms of the α moments. By expanding Eq. (23.205), it is found that

$$\sigma^2 = \mu_2 = \int_{-\infty}^{\infty} x^2 f(x) \, dx - 2m \int_{-\infty}^{\infty} x f(x) \, dx + m^2 \int_{-\infty}^{\infty} f(x) \, dx$$

$$= \alpha_2 - \alpha_1^2 = \mathbf{E}(x^2) - \mathbf{E}^2(x) \qquad (23.206)$$

This equation holds for discrete as well as continuous distributions.

A useful representation of probability distributions is obtained if the moments α_n are written as coefficients of an arbitrary variable $(-s)^n/n!$ in an infinite series, viz.,

$$\chi_f(s) = \alpha_0 - \alpha_1 \frac{s}{1!} + \alpha_2 \frac{s^2}{2!} - \alpha_3 \frac{s^3}{3!} + \cdots \qquad (23.207)$$

Assuming that the series converges uniformly in the interval $-\epsilon < s < \epsilon$, $(\epsilon > 0)$, a formal representation of $\chi_f(s)$ is then

$$\chi_f(s) = \mathbf{E}(e^{-sx}) = \int_{-\infty}^{\infty} e^{-sx} f(x) \, dx \qquad (23.208)$$

The function $\chi_f(s)$ is called the *moment generating function*, or *characteristic function*, of the random variable x. As Eq. (23.208) shows, it is equal to the expected value of the random variable e^{-sx}. Assuming that s is a complex variable $s = \gamma + j\omega$, $\chi_f(s)$ is the Fourier transform of $f(x)$ if $\gamma = 0$; it is the *bilateral* Laplace transform of $f(x)$ if $\gamma \neq 0$. These transforms are defined only if the integral in Eq. (23.208) is absolutely convergent.

One use of moment generating functions is to compute moments themselves. Based on Eq. (23.207), it is seen that α_n is equal to $(-1)^n$ times the nth derivative of $\chi_f(s)$ with $s = 0$.

$$\alpha_n = \left[(-1)^n \frac{d^n \chi_f(s)}{ds^n} \right]_{s=0} \qquad (23.209)$$

The inverse of Eq. (23.208) is defined (subject to certain convergence restrictions) by Eq. (23.210).

$$f(x) = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} e^{sx} \chi_f(s) \, ds \qquad (23.210)$$

The equation corresponding to the continuous-probability-distribution definition of the moment generating function [Eq. (23.208)] is given by Eq. (23.211).

$$\chi_P(s) = \sum_{r=0}^{N} e^{-sx_r} P(x_r) \qquad (23.211)$$

Table 23.10 gives the expected value m, standard deviation σ, and the moment generating function $\chi_f(s)$ and $\chi_P(s)$ for several of the commonest probability distributions.

23.20j. Other Common Probability Distributions. In Table 23.10 are included several probability distributions not mentioned thus far.

Single-point Distribution. The simplest distribution is the *single-point distribution* which applies when it can be assumed that it is "practically certain" that the value of a random value x will be x_k, that is, x takes on the value x_k with unity probability. This distribution can be considered to be the limit of other distributions when the standard deviation approaches zero.

Geometric Distribution. This distribution applies to problems involving the probability of a number of failures occurring before a success in Bernoulli trials since the probability of obtaining r failures and one success is pq^r if p is the probability of success at each trial. From Table 23.10, it is seen that the expected number of trials before a success is q/p.

Rectangular Distribution. In a rectangular distribution, the probability density function is a constant over a given interval and zero outside. An application for this distribution occurs in problems involving the probable positions of rotating devices where any angle of rotation is equally probable.

Exponential Distribution. This distribution is useful when the random variable

TABLE 23.10. EXPECTED VALUES, STANDARD DEVIATIONS, AND MOMENT GENERATING FUNCTIONS FOR SOME COMMON PROBABILITY DISTRIBUTIONS

Name	Probability density function	Expected value	Standard deviation	Moment generating function
Discrete distribution...............	$P(x_r)$	m	σ	$\chi_P(s)$
Single-point distribution............	$P(x_r) = \begin{cases} 1 & r = k \\ 0 & r \neq k \end{cases}$	x_k	0	e^{-sx_k}
Binomial distribution............	$\binom{n}{r} p^r q^{n-r} \quad (r = 0,1,2, \ldots , n)$ $0 < p < 1; q = 1 - p$	np	$\sqrt{npq}$	$\sum\limits_{r=0}^{n} \binom{n}{r} p^r q^{n-r} e^{-sr}$
Poisson distribution............	$\dfrac{(\lambda T)^r}{r!} e^{-\lambda T}$	λT	$\sqrt{\lambda T}$	$e^{\lambda T(e^{-s}-1)}$
Geometric distribution............	$p\, q^r \quad (r = 0,1,2, \ldots)$ $0 < p < 1; q = 1 - p$	$\dfrac{q}{p}$	$\dfrac{\sqrt{q}}{p}$	$\dfrac{p}{1 - qe^{-s}}$
Continuous distribution............	$f(x)$	m	σ	$\chi_f(s)$
Rectangular distribution............	$f(x) = \begin{cases} 0 & x < a - h \\ \dfrac{1}{2h} & a - h < x < a + h \\ 0 & a + h < x \end{cases}$	a	$\dfrac{h}{\sqrt{3}}$	$\dfrac{\sinh hs}{hse^{as}}$
Normal distribution............	$f(x) = \phi\left(\dfrac{x - m}{\sigma}\right) = \dfrac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$	m	σ	$e^{-ms+\frac{\sigma^2 s^2}{2}}$
Exponential distribution............	$f(x) = \begin{cases} 0 & x < 0 \\ \dfrac{1}{\lambda} e^{-\lambda x} & 0 < x \end{cases}$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda}$	$\dfrac{\lambda}{\lambda + s}$
Rayleigh distribution	$f(x) = \begin{cases} 0 & x < 0 \\ \dfrac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} & 0 < x \end{cases}$	$\sigma \sqrt{\dfrac{\pi}{2}}$	$\sigma \sqrt{2 - \dfrac{\pi}{2}}$	$1 - \sigma s \sqrt{2\pi}\, e^{\frac{\sigma^2 s^2}{2}} \Phi(-\sigma s)$
Normal amplitude − squared distribution............	$f(x) = \begin{cases} 0 & x < 0 \\ \dfrac{x}{\sqrt{2\pi}\sigma} e^{-x/2\sigma^2} & 0 < x \end{cases}$	σ^2	$\sigma^2 \sqrt{3}$	$\dfrac{1}{(2\sigma^2 s + 1)^{1/2}}$

is the length of an interval without the occurrence of a particular event where the average rate of occurrence of the event is λ (Poisson distribution with $r = 0$; see Sec. 23.20d).

Rayleigh Distribution. When thermal noise is placed at the input of a filter whose transfer function is symmetrical about the center frequency of the filter and which has a bandwidth small compared to its center frequency (e.g., an i-f amplifier), the noise at the output of the filter appears as a randomly amplitude-modulated sinewave with random phase changes. The amplitude distribution of the sinewavelike output is still normal (Gaussian) since any Gaussian process results in another Gaussian process when the first process has been transformed by a linear device such as a filter. However, the distribution of the peaks of the sinewaves is not normal. Instead, the amplitudes of the peaks follow the *Rayleigh distribution* shown in Table 23.10. If σ is the rms filter output noise voltage, the expected value of voltage or d-c component of the peaks is $\sigma \sqrt{\pi/2}$. This is the voltage which would appear at the output of an ideal linear detector. For an ideal square-law detector, i.e., a device which squares the input voltage, the output d-c component would be

$$\sigma^2(2 - \pi/2 + \pi/2) = 2\sigma^2$$

Normal Amplitude-squared Distribution. This distribution applies in a Gaussian process where the random variable is taken to be amplitude-squared, for example, in a thermal noise process where the random variable is taken to be power. It also is related to the chi-square distribution used in testing the significance of a set of observations of a random process[1] (see Example 23.37).

23.20k. Joint Probability Distributions; Multiplication of Random Variables. In Sec. 23.20b, Eq. (23.178) may be written in the form of a *joint probability density function* for two random variables. For two discrete random variables x_r and y_t, ($r = 0$, 1, 2, . . . , N; $t = 0, 1, 2, . . . , M$), the joint probability density function $P(x_r, y_t)$ is

$$P(x_r, y_t) = P_1(x_r)P_1(y_t|x_r) = P_2(y_t)P_2(x_r|y_t) \qquad (23.212)$$

In the continuous case, for two random variables x and y the joint density function $f(x,y)$ is

$$f(x,y) = f_1(x)f_1(y|x) = f_2(y)f_2(x|y) \qquad (23.213)$$

The quantities $P_1(y_t|x_r)$, $P_2(x_r|y_t)$, $f_1(y|x)$, $f_2(x|y)$ are the *conditional probability density functions* of the random variables "y_t occurs after it is known x_r occurs," "x_r after y_t," "y after x," and "x after y." The probability density functions of x_r, y_t, x, y are denoted by $P_1(x_r)$, $P_2(y_t)$, $f_1(x)$, $f_2(y)$, respectively. The probability of x_r occurring for all possible value of y_t occurring is

$$P_1(x_r) = \sum_{t=0}^{M} P(x_r, y_t) \qquad (23.214)$$

Similarly,

$$P_2(y_t) = \sum_{r=0}^{N} P(x_r, y_t) \qquad (23.215)$$

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y) \, dy \qquad (23.216)$$

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y) \, dx \qquad (23.217)$$

---

[1] See H. Cramer, *Mathematical Methods of Statistics*, pp. 416ff, Princeton University Press, Princeton, N.J., 1951.

If $P_1(y_t|x_r) = P_2(y_t)$, y_t is said to be *independent* of x_r; or x_r and y_t are said to be *mutually independent* (except for values of random variables of zero probability). Likewise, if $f_1(y|x) = f_2(y)$, y is independent of x, and x is independent of y. Thus, if two random variables are independent,

$$P(x_r,y_t) = P_1(x_r)P_2(y_t) \qquad (23.218)$$
$$f(x,y) = f_1(x)f_2(y) \qquad (23.219)$$

Further, if two random variables x, y are independent, the expected value of their product $\mathbf{E}(xy)$ is $\mathbf{E}(x)\mathbf{E}(y)$.

$$\mathbf{E}(xy) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x,y)\ dx\ dy$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf_1(x)f_2(y)\ dx\ dy$$
$$\mathbf{E}(xy) = \mathbf{E}(x)\mathbf{E}(y) \qquad (x,y \text{ independent}) \qquad (23.220)$$

In fact, for any functions of the random variables $g(x)$, $h(y)$,

$$\mathbf{E}[g(x)h(y)] = \mathbf{E}[g(x)]\mathbf{E}[h(y)] \qquad (x,y \text{ independent}) \qquad (23.221)$$

The corresponding expected value relations for discrete random variables also hold.

Example 23.36

A radar is capable of detecting the presence of any aircraft within a radius of 100 nautical miles. Suppose that a study has indicated that the distribution of aircraft velocities v is approximately triangular with parameters given by $f(v)$ below where v is in units of knots.

$$f(v) = \begin{cases} 0 & v \leq 150 \\ \dfrac{v - 150}{33{,}750} & 150 \leq v \leq 300 \\ \dfrac{600 - v}{67{,}500} & 300 \leq v \leq 600 \\ 0 & 600 \leq v \end{cases}$$

Further, suppose that the aircraft are equally likely to be flying in any direction through the 100-mile-radius circle, but they always travel in straight lines and do not change their speed as they travel. What is the average time which they spend in the circle, and what is the standard deviation of the time?

Solution

1. The problem calls for the computation of the expected value $\mathbf{E}(t)$ of time an aircraft is within the circle which is equal to $\mathbf{E}(L/v)$, where L is the path (chord) length, and v is the velocity of an aircraft through the circle. Since L and v are independent,

$$\mathbf{E}\left(\frac{L}{v}\right) = \mathbf{E}(L)\mathbf{E}\left(\frac{1}{v}\right)$$
and
$$\mathbf{E}\left(\frac{L}{v}\right)^2 = \mathbf{E}(L^2)\mathbf{E}\left(\frac{1}{v^2}\right)$$

2. Compute the expected value and standard deviation of the path length through the circle. By the conditions of the problem, it may be assumed that the minimum distance r between the aircraft flight path and the center of the circle is uniformly distributed between 0 and R, the radius of the circle; that is, the probability function for r is

$$f_1(r) = \begin{cases} 0 & r < 0 \\ 1/R & 0 < r < R \\ 0 & R < r \end{cases}$$

The length of a chord in terms of the minimum radius of the chord is given by

$$L = 2\sqrt{R^2 - r^2}$$

Therefore,

$$\mathbf{E}(L) = \int_{-\infty}^{\infty} Lf_1(r) \, dr$$

$$= \int_0^R \frac{2\sqrt{R^2 - r^2}}{R} \, dr = \frac{\pi}{2} R$$

$$\mathbf{E}(L^2) = \int_0^R \frac{4(R^2 - r^2)}{R} \, dr = \frac{8}{3} R^2$$

For $R = 100$ miles,

$$\mathbf{E}(L) = 157.1 \text{ miles}$$
$$\mathbf{E}(L^2) = \tfrac{8}{3} \times (100)^2 = 26{,}667 \text{ miles}^2$$

3. Compute $\mathbf{E}(1/v)$ and $\mathbf{E}(1/v^2)$.

$$\mathbf{E}\left(\frac{1}{v}\right) = \int_{-\infty}^{\infty} \frac{1}{v} f(v) \, dv$$

$$= \int_a^b \frac{2(v - a)}{v(b - a)(c - a)} \, dv + \int_b^c \frac{2(c - v)}{v(c - b)(c - a)} \, dv$$

where $a = 150$, $b = 300$, $c = 600$ knots

$$\mathbf{E}\left(\frac{1}{v}\right) = \frac{2}{c - a} \left[\frac{c \ln (c/b)}{c - b} - \frac{a \ln (b/a)}{b - a} \right]$$

$$\mathbf{E}\left(\frac{1}{v^2}\right) = \int_a^b \frac{2(v - a)}{v^2(b - a)(c - a)} \, dv + \int_b^c \frac{2(c - v)}{v^2(c - b)(c - a)} \, dv$$

$$= \frac{2}{c - a} \left[\frac{\ln (b/a)}{(b - a)} - \frac{\ln (c/b)}{(c - b)} \right]$$

Thus, for the particular values of a, b, and c,

$$\mathbf{E}\left(\frac{1}{v}\right) = \frac{2}{450} \left[\frac{600 \ln 2}{300} - \frac{150 \ln 2}{150} \right] = \frac{\ln 2}{225} \approx 0.003{,}080{,}7 \text{ per knot}$$

$$\mathbf{E}\left(\frac{1}{v^2}\right) = \frac{2}{450} \left[\frac{\ln 2}{150} - \frac{\ln 2}{300} \right] = \frac{\ln 2}{67{,}500} \approx 0.000{,}010{,}269 \text{ per knot}^2$$

4. The expected value $\mathbf{E}(t)$ of time an aircraft is within the circle is

$$\mathbf{E}(t) = \mathbf{E}(L)\mathbf{E}\left(\frac{1}{v}\right) = (151.7)(0.003{,}080{,}7)$$

$$= 0.48391 \text{ hr}$$
$$= 29.03 \text{ min}$$

$$\sigma_t{}^2 = \mathbf{E}(t^2) - \mathbf{E}^2(t) = \mathbf{E}(L^2)\mathbf{E}\left(\frac{1}{v^2}\right) - \mathbf{E}^2(t)$$

$$= (26{,}667)(0.000{,}010{,}269)^2 - (0.48391)^2$$
$$= 0.039669$$
$$\sigma_t = 0.1992 \text{ hr} = 11.95 \text{ min}$$

23.20l. Addition of Random Variables. For two random variables x and y with a joint probability density function $f(x,y)$, the expected value of their sum $x + y$ is

$$\mathbf{E}(x + y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y)f(x,y) \, dx \, dy$$

$$= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f(x,y) \, dy \, dx + \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f(x,y) \, dx \, dy$$

$$= \int_{-\infty}^{\infty} xf_1(x) \, dx + \int_{-\infty}^{\infty} yf_2(y) \, dy$$

$$= \mathbf{E}(x) + \mathbf{E}(y) \tag{23.222}$$

Thus, *the expected value of the sum of two random variables is equal to the sum of the expected values of the random variables.* The same statement is true for more than two random variables in both the continuous and discrete cases. (Note that the random variables are *not* necessarily independent.)

When two random variables are independent, it is possible to compute sum formulas for moments of order higher than the first in terms of the moments of the individual random variables. From Sec. 23.20i, the moment generating function $\chi(s)$ for the sum of two independent random variables x and y is

$$\chi(s) = \mathbf{E}(e^{-s(x+y)}) = \mathbf{E}(e^{-sx}e^{-sy})$$

Since x and y are independent Eq. (23.221) gives

$$\chi(s) = \mathbf{E}(e^{-sx})\mathbf{E}(e^{-sy}) = \chi_1(s)\chi_2(s) \qquad (x,y \text{ independent}) \qquad (23.223)$$

where $\chi_1(x)$ and $\chi_2(y)$ are the moment generating functions of x and y. Expanding Eq. (23.223) and solving for the moments gives the following formulas:

$$\mathbf{E}[(x + y)^2] = \mathbf{E}(x^2) + \mathbf{E}(y^2) + 2\mathbf{E}(x)\mathbf{E}(y) \qquad (x,y \text{ independent}) \qquad (23.224)$$
$$\sigma_{x+y}^2 = \mathbf{E}[(x + y)^2] - \mathbf{E}^2(x + y)$$
$$= \mathbf{E}(x^2) + \mathbf{E}(y^2) + 2\mathbf{E}(x)\mathbf{E}(y) - \mathbf{E}^2(x) - \mathbf{E}^2(y) - 2\mathbf{E}(x)\mathbf{E}(y)$$
or $\qquad \sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2 \qquad (x,y \text{ independent}) \qquad (23.225)$

For two independent random variables, the variance of the sum is equal to the sum of the variances of each random variable. This formula applies to more than two random variables. However, for central moments greater than third order, the simple form of Eq. (23.225) no longer holds.

Since moment generating functions are the (bilateral) Laplace or Fourier transforms of probability density functions, it is possible to obtain density functions from moment generating functions subject to certain convergence restrictions imposed on the integrals involved. Thus, the probability density function $f(z)$ for the sum of two random variables $z = x + y$ with individual moment generating functions $\chi_1(s)$ and $\chi_2(s)$ is, from Eq. (23.223),

$$f(z) = \int_{-\infty}^{\infty} \chi(s)e^{sz}\,ds = \int_{-\infty}^{\infty} \chi_1(s)\chi_2(s)e^{sz}\,ds \qquad (23.226)$$

An equivalent formula may be derived from Eq. (23.226).

$$f(z) = \int_{-\infty}^{\infty} f_1(z - \xi)f_2(\xi)\,d\xi = \int_{-\infty}^{\infty} f_1(\xi)f_2(z - \xi)\,d\xi \qquad (23.227)$$

where $f_1(x)$ and $f_2(y)$ are the density functions for the random variables x and y. Either form of Eq. (23.227) is called a *convolution integral.* Equivalent equations for distribution functions are

$$F(z) = \int_{-\infty}^{\infty} F_1(z - \xi)f_2(\xi)\,d\xi = \int_{-\infty}^{\infty} F_2(z - \xi)f_1(\xi)\,d\xi \qquad (23.228)$$

where $F_1(x)$, $F_2(y)$, and $F(z)$ are distribution functions for x, y, and $z = x + y$, respectively.

Example 23.37

Compute the number of independent samples which must be measured of a Gaussian noise source with an expected value equal to zero and standard deviation σ to attain a probability of 0.95 that the rms value σ^* of the samples is within 50 per cent of the standard deviation.

Solution

1. Compute the moment generating function for the sum of n Gaussian noise samples $x_1^2 + x_2^2 + \cdots + x_n^2$. The individual random variables $x_1, x_2, \ldots, x_n$ are all assumed to follow the same normal distribution, viz.,

$$f_i(x_i) = \frac{1}{\sigma \sqrt{2\pi}} e^{-x_i^2/2\sigma^2} \qquad -\infty < x_i < \infty$$

so that the probability of x_i being between x_i and $x_i + dx_i$ is

$$f_i(x_i) \, dx_i = \frac{1}{\sigma \sqrt{2\pi}} e^{-x_i^2/2\sigma^2} \, dx_i \qquad -\infty < x_i < \infty$$

Letting $y_i = x_i^2$, the probability of y_i being between y_i and $y_i + dy_i$ is

$$g_i(y_i) \, dy_i = 2f_i(x_i) \, dx_i = \frac{1}{\sigma \sqrt{2\pi}} \frac{e^{-y_i/2\sigma^2}}{\sqrt{y_i}} \, dy_i \qquad -\infty < x_i < \infty \,;\, 0 < y_i < \infty$$

The moment generating function for y_i, $X_{g_i}(s)$ is

$$X_{g_i}(s) = \int_0^\infty \frac{1}{\sigma \sqrt{2\pi}} \frac{e^{-y_i/2\sigma^2}}{\sqrt{y_i}} e^{-sy_i} \, dy_i$$

$$= \frac{1}{(2\sigma^2 s + 1)^{\frac{1}{2}}}$$

Since the random variables $x_1, x_2, \ldots, x_n$ are independent, the moment generating function $X_g(s)$ for $x_1^2 + x_2^2 + \ldots + x_n^2 = y_1 + y_2 + \cdots + y_n$ is

$$X_g(s) = \frac{1}{(2\sigma^2 s + 1)^{n/2}}$$

2. Find the probability density function for $x_1^2 + x_2^2 + \cdots + x_n^2$. The moment generating function for the sum is that of the chi-square distribution.[1] The probability density function $g(y)$ for the sum is

$$g(y) = \frac{y^{(n-2)/2} e^{-y/2\sigma^2}}{(2\sigma^2)^{n/2} \Gamma(n/2)} \qquad 0 < y < \infty$$

3. Compute the number of samples required to achieve a probability $P = 0.95$ that the sum $\sigma^* = \sqrt{(x_1^2 + x_2^2 + \cdots + x_n^2)/n}$ is within 50 per cent of σ. In other words, compute n such that the probability is 0.95 that

$$0.25\sigma^2 < \frac{x_1^2 + x_2^2 \cdots + x_n^2}{n} < 2.25\sigma^2$$

or
$$0.25n\sigma^2 < y_1 + y_2 + \cdots + y_n < 2.25n\sigma^2$$

Thus, it is required to find n such that

$$P = \int_{0.25n\sigma^2}^{2.25n\sigma^2} g(y) \, dy = \int_{0.25n\sigma^2}^{2.25n\sigma^2} \frac{y^{n/2-1} e^{-y/2\sigma^2}}{(2\sigma^2)^{n/2} \Gamma(n/2)} \, dy = 0.95$$

Letting $z = \sigma^2 y$, this integral becomes

$$P = \int_{0.25n\sigma^2}^{2.25n\sigma^2} g(y) \, dy = \int_{0.25n}^{2.25n} \frac{z^{n/2-1} e^{-z/2}}{(2)^{n/2} \Gamma(n/2)} \, dz$$

$$= \int_{0.25n}^{2.25n} h(z) \, dz = 0.95$$

where
$$h(z) = \frac{z^{n/2-1} e^{-z/2}}{(2)^{n/2} \Gamma(n/2)}$$

---

[1] *Op. cit.*

Tables for the chi-square distribution integral are generally written in terms of $\int_a^\infty h(z)\ dz$ so that it is necessary to rewrite the integral above.

$$P = \int_{0.25n}^{2.25n} h(z)\ dz = \int_{0.25n}^{\infty} h(z)\ dz - \int_{2.25n}^{\infty} h(z)\ dz = 0.95$$

From a table for the chi-square distribution, the following table can be prepared.

n	$\int_{0.25n}^{\infty} h(z)\ dz$	$\int_{2.25n}^{\infty} h(z)\ dz$	P
6	0.96	0.04	0.92
7	0.97	0.03	0.94
8	0.98	0.02	0.96
9	0.985	0.015	0.97

From this table, it is seen that eight samples are required to assure a probability of 0.95 that the measured value of standard deviation is within ±50 per cent of the actual standard deviation.

23.20m. Central Limit Theorem. The probability density function for the sum of many independent random variables is usually difficult to obtain, particularly when the density function for each random variable of the sum is different. However, under very general conditions, it is often possible to use a reasonable approximation to the density function which results from the *central limit theorem.* Roughly stated, this theorem postulates that the density function for a sum of many independent random variables is approximately normal. If each of the random variables x_1, x_2, . . . , x_n has corresponding expected values and standard deviations m_1, m_2, . . . , m_n and σ_1, σ_2, . . . , σ_m, the expected value m and standard deviation σ of the normal distribution approximating the sum are

$$m = m_1 + m_2 + \cdots + m_n$$
$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2}$$

The approximation is particularly good when no random variable predominates in the sum of the random variables. Schott noise within a vacuum tube and thermal noise within resistors are good examples of the applicability of the central limit theorem.

23.20n. Correlation of Random Variables. When adding two random variables x and y which are not independent, Eq. (23.222) holds for the expected value of their sum. However, when computing the variance of the sum of two dependent random variables, Eq. (23.225) does not hold.

$$\begin{aligned}
\sigma_{x+y}^2 &= E[(x + y)^2] - E^2(x + y) \\
&= E(x^2) + E(y^2) + 2E(xy) - E^2(x) - E^2(y) - 2E(x)E(y) \\
&= \sigma_x^2 + \sigma_y^2 + 2[E(xy) - E(x)E(y)] & (23.229) \\
&= \sigma_x^2 + \sigma_y^2 + 2\rho\sigma_x\sigma_y & (23.230)
\end{aligned}$$

where ρ is called the *normalized correlation coefficient* and is defined by

$$\rho = \frac{E(xy) - E(x)E(y)}{\sigma_x\sigma_y} \qquad (23.231)$$

The coefficient ρ varies between -1 and $+1$ since it is normalized by the factor $\sigma_x\sigma_y$. Its name implies that it is a measure of how independent one random variable is on

another—an implication which is not strictly true. It is the case that two independent random variables have a zero correlation coefficient. However, a zero correlation coefficient does not necessarily imply independence.

23.20o. Correlation Functions Applied to Time Series; Stationary Time Series and the Ergodic Property. The most important present-day application of correlation coefficients in engineering arises in the study of *time series*, i.e., a continuous or discrete series of events which are connected statistically so that successive values of the series are not statistically independent.

If a continuous time series $x(t)$ is given, then a normalized correlation coefficient relating the expected values of the series at two times t_1 and t_2 may be calculated by Eq. (23.231).

$$\rho(t_1,t_2) = \frac{\mathbf{E}[x(t_1)x(t_2)] - \mathbf{E}[x(t_1)]\mathbf{E}[x(t_2)]}{\sigma_x(t_1)\sigma_x(t_2)}$$

where $x(t_1)$ and $x(t_2)$ are the values of the time series at t_1 and t_2, respectively, and $\sigma_{x(t_1)}$ and $\sigma_{x(t_2)}$ are the standard deviations of the time series at times t_1 and t_2. It is more common practice to deal not with the normalized correlation coefficient but rather with the (unnormalized) *correlation function* $\phi(t_1,t_2)$ defined by

$$\phi(t_1,t_2) = \mathbf{E}[x(t_1)x(t_2)]$$
$$= \int_{-\infty}^{\infty} x(t_1)x(t_2)f[x(t_1),x(t_2)]\, dx(t_1)\, dx(t_2) \qquad (23.232)$$

where $f[x(t_1),x(t_2)]$ is the joint probability density function for the value of $x(t)$ at times t_1 and t_2. Note that if two random variables do not have zero mean, the correlation function defined in this manner will not be zero if the random variables are independent.

It is often the case that the probability density function for $x(t)$ is independent of t. In such a case, the time series is said to be *stationary*. The correlation function of a stationary time series depends only on the difference between t_1 and t_2 and not on the actual value of either. For a difference τ between times, the correlation function of a stationary time series is

$$\phi(\tau) = \int_{-\infty}^{\infty} x(t)x(t + \tau)f[x(t),x(t + \tau)]\, dx(t)\, dx(t + \tau) \qquad (23.233)$$

Another function can be defined which is often equal to the correlation function for a stationary time series. It is given by

$$\psi(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t)x(t + \tau)\, dt \qquad (23.234)$$

The function $\psi(\tau)$ is equal to $\phi(\tau)$ if the stationary time series possesses the *ergodic property*. A stationary time series is defined to be ergodic if

$$\mathbf{E}[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)\, dx = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} g[x(t + u)]\, du \qquad (23.235)$$

for all values of t except for time series of zero probability. The first integral is termed an *ensemble average* of the function $g(x)$; the second integral is called a *time average*. Loosely speaking, an ensemble of stationary time series possesses the ergodic property except for time series of zero probability if, in a collection of all possible time series pertaining to a particular problem, it is highly improbable that one time series is selected as the outcome of an experiment where the ensemble and time averages are not equal. For example, thermal noise can be assumed to be ergodic. It is highly improbable that a sample of thermal noise taken over a long time will be a

function from which the autocorrelation function cannot be calculated by time averages.

When the ergodic property can be assumed to hold, the correlation function for a continuous stationary time series is

$$\phi(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t)x(t + \tau) \, dt \tag{23.236}$$

Analogously, for a discrete stationary time series,

$$\phi_i = \lim_{N \to \infty} \frac{1}{2N + 1} \sum_{i = -N}^{N} x_i x_{i+j} \tag{23.237}$$

From these formulas, it is seen that when the ergodic property holds, the expected squared value of the time series $\mathbf{E}[x^2(t)]$ is

$$\mathbf{E}[x^2(t)] = \phi(0) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x^2(t) \, dt \tag{23.238}$$

and

$$\mathbf{E}[x_i{}^2] = \phi_0 = \lim_{N \to \infty} \frac{1}{2N + 1} \sum_{i = -N}^{N} x_i{}^2 \tag{23.239}$$

Further, when a stationary time series is ergodic, the mean or expected value of x is

$$\mathbf{E}[x(t)] = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t) \, dt \tag{23.240}$$

and

$$\mathbf{E}[x_i] = \lim_{N \to \infty} \frac{1}{2N + 1} \sum_{i = -N}^{N} x_i \tag{23.241}$$

When a time series can be assumed to be stationary and possess the ergodic property, the preceding formulas provide a basis for deriving important statistical information from observed values of the time series. Further, it is possible to determine the reasonableness of the assumption that a time series either is stationary or possesses the ergodic property by applying tests to observations made on the time series, although the derivations of such tests involve fairly complicated statistical questions.

23.20p. Spectra of Stationary Time Series. Subject to certain convergence restrictions applied to the functions involved, a time series (single-value function of time) possesses a Laplace transform or, what is practically equivalent, a Fourier transform. It is usually the practice to utilize the Fourier transform in the discussion of spectra of stationary time series.

For a function of time $\xi(t)$, the Fourier transform of $\xi(t)$ is defined to be

$$H(\omega) = \int_{-\infty}^{\infty} \xi(t)e^{-j\omega t} \, dt \tag{23.242}$$

If $\xi(t)$ is a function proportional to voltage or current in an electrical network, then $H(\omega)$ is called the Fourier spectrum of $\xi(t)$. The square of the absolute value of $H(\omega)$ is called the *energy spectral density* $E(\omega)$ of $\xi(t)$.

$$E(\omega) = |H(\omega)|^2 = H(\omega)H^*(\omega) \tag{23.243}$$

where $H^*(\omega)$ = complex conjugate of $H(\omega)$

From the mathematics of the Fourier integral (Parseval's theorem)

$$\int_0^\infty E(\omega)\,d\omega \;=\; \pi \int_{-\infty}^\infty \xi^2(t)\,dt \tag{23.244}$$

When considering time series, it is usually the case that the energy spectrum of the time function is infinite since the time series is assumed to extend indefinitely in time. The *power spectrum* is then of more interest than the energy spectrum which would, of necessity, be limited to only a finite portion of the time series. For a time series $x(t)$, the *power spectral* density is defined by

$$Y(\omega) \;=\; \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^T x(t) e^{-j\omega t}\,dt \tag{23.245}$$

The *power spectrum* $G(\omega)$ is then

$$G(\omega) \;=\; |Y(\omega)|^2 \;=\; Y(\omega) Y^*(\omega) \tag{23.246}$$

Further,

$$\int_0^\infty G(\omega)\,d\omega \;=\; \lim_{T\to\infty} \frac{\pi}{2T} \int_{-T}^T x^2(t)\,dt \tag{23.247}$$

These equations hold for nonstationary as well as stationary time series.

In terms of the function $\psi(\tau)$, defined by Eq. (23.234), it can be shown that

$$G(\omega) \;=\; \int_{-\infty}^\infty \psi(\tau)\, e^{-j\omega\tau}\,d\tau \tag{23.248}$$

Thus, if a time series is both stationary and ergodic, the power spectrum $G_\phi(\omega)$ is

$$G_\phi(\omega) \;=\; \int_{-\infty}^\infty \phi(\tau) e^{-j\omega\tau}\,d\tau \tag{23.249}$$

Therefore, *the spectrum of a stationary time series which possesses the ergodic property can be computed in terms of purely statistical parameters of the time series.* Further,

$$\phi(\tau) \;=\; \frac{1}{2\pi} \int_{-\infty}^\infty G_\phi(\omega) e^{jwt}\,d\omega \tag{23.250}$$

Thus, *for a stationary time series which possesses the ergodic property, statistical data can be determined from spectral information concerning the time series.* Equations (23.249) and (23.250) form the basis for a large portion of modern filter theory devoted to problems of separating signals and noise.

Index

1