RADIO ENGINEERING

BY

FREDERICK EMMONS TERMAN, Sc.D.
Associate Professor of Electrical Engineering,
Stanford University

FIRST EDITION

McGRAW-HILL BOOK COMPANY, INC.
NEW YORK AND LONDON
1932

PRINTED IN THE UNITED STATES OF AMERICA

THE MAPLE PRESS COMPANY, YORK, PA.

PREFACE

The aim of "Radio Engineering" is to present a comprehensive engineering treatment of the more important vacuum tube and radio phenomena. Electrical circuits and vacuum tubes behave according to exact laws, which in the main are simple and easily understood, and which can be used to predict the performance of radio circuits and radio apparatus with the same certainty and accuracy that the performance of other types of electrical equipment, such as transformers, motors, and transmission lines, is analyzed. It is this ability to reduce a problem to quantitative relations that predict with accuracy the performance to be expected or explain the results already obtained that represents a real mastery of the subject such as the radio engineer is expected to possess.

The principal prerequisite for undertaking the study of radio engineering is a good working knowledge of the fundamental concepts of alternating currents, such as reactance, impedance, power factor, phase angle, and vector representation. An elementary idea of complex quantity notation is also desirable but not absolutely essential. This means that radio work as outlined in this volume can be taken up in the senior year of the usual electrical engineering curriculum.

The order of presentation has been intentionally so arranged that the first part of the volume is devoted to the theory of tuned circuits and the fundamental properties of vacuum tubes and vacuum tube applications, all of which are of importance and interest to every electrical engineer. The latter part then takes up more specialized radio topics, such as radio receivers and transmitters, wave propagation, antennas, and direction finding. This makes it possible where desired to arrange a two-semester course, the first term of which will be suitable for all electrical engineers, with the second term continuing for those with definite radio interests.

Particular care has been taken to avoid unnecessary equations in developing the analytical side of the various subjects taken up. It has been the author's experience that the usual student when first coming in contact with a new subject is confused by the presence of numerous equations and, in such circumstances, frequently fails to realize which relations are of real importance. By carrying on the reasoning in terms of words as far as possible, by judicious use of footnotes, and by skipping over purely routine mathematical manipulations, it has been possible to cut down the number of equations appearing in the text to the point where the important mathematical relations stand out by virtue of the fact that they stand nearly alone, free of attention-diverting trivial equations. The result is that while "Radio Engineering" appears to be relatively free from mathematics, yet it actually carries the analysis much deeper than is customary. A typical illustration of this is the treatment of the transformer-coupled amplifier given in Chap. V. As far as the author is aware, this represents the only published analysis that can be used to predict the complete amplification characteristic of the transformer-coupled amplifier without an unreasonable amount of work and with engineering accuracy. At the same time it is almost devoid of mathematics as compared with the incomplete and often incorrect treatment ordinarily found. This result has been achieved by carrying the reasoning along in terms of physical concepts and words and by writing down an equation only when the equation itself is of importance.

A considerable quantity of original material is being published here for the first time. Notable instances of this are the analysis of the transformer-coupled amplifier mentioned above, the universal resonance curve, the Class A power amplifier formulas, the analysis of the Class B (linear) power amplifier, the analysis of regeneration resulting from a common plate impedance, the concept of the effective Q of the tuned amplifier, the analysis of the input admittance of amplifiers, the treatment of the voltage and current relations existing in the screen-grid tube, and the approximate analysis of rectifier-filter systems having a shunt condenser across the filter input.

The footnote references form an integral part of the text and have been carefully selected with a view toward helping the reader who desires more information on a particular subject than is given in this volume. No attempt has been made to compile complete bibliographies, the aim having been rather to cite a limited number of comprehensive articles that are really readable by the average student.

The author wishes to acknowledge the very helpful cooperation which has been received on all sides. Particular mention should be made of Philip G. Caldwell, the late Nathaniel R. Morgan, Paul F. Byrne, Dr. Horace E. Overacker, William R. Triplett, Harry Engwicht, and D. A. Murray, all former students at Stanford University, who assisted in drawing the figures and checking the manuscript and proof. The author is also greatly indebted to the Bell Telephone Laboratories, the American Telephone and Telegraph Company, the General Radio Company, the De Forest Radio Company, the General Electric Company, RCA-Radiotron, Inc., and the Stromberg-Carlson Telephone Manufacturing Company for supplying copy for certain of the illustrations.

FREDERICK EMMONS TERMAN.

STANFORD UNIVERSITY, CALIFORNIA,
August, 1932.

CONTENTS

CHAPTER XVII

RADIO MEASUREMENTS

CHAPTER XVIII

SOUND AND SOUND EQUIPMENT

APPENDIX A

APPENDIX B

RADIO ENGINEERING

CHAPTER I

THE ELEMENTS OF A SYSTEM OF RADIO COMMUNICATION

1. Radio Waves.—Electrical energy that has escaped into free space exists in the form of electromagnetic waves. These waves, which are commonly called radio waves, travel with the velocity of light and consist of magnetic and electrostatic fields at right angles to each other and also at right angles to the direction of travel. If these electrostatic and magnetic fluxes could be actually seen, the wave would have the appearance indicated in Fig. 1. One half of the electrical energy contained

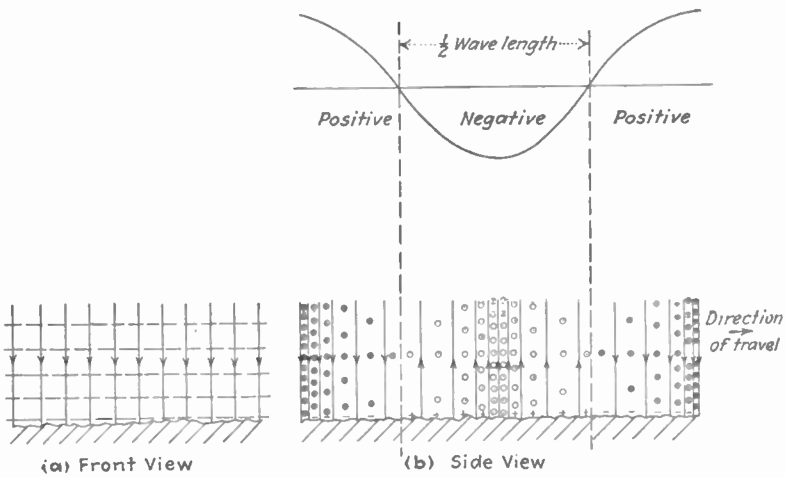


FIG. 1.—Front and side views of a vertically polarized wave. The solid lines represent electrostatic flux, while the dotted lines and the circles indicate magnetic flux.

in the wave exists in the form of electrostatic energy, while the remaining half is in the form of magnetic energy.

The essential properties of a radio wave are the frequency, intensity, direction of travel, and plane of polarization. The radio waves produced by an alternating current will vary in intensity at the frequency of the current and will therefore be alternately positive and negative as shown at *b* in Fig. 1. The distance occupied by one complete cycle of such an alternating wave is equal to the velocity of the wave divided by the number of cycles that are sent out each second and is called the wave length.

The relation between wave length λ in meters and frequency f in cycles per second is therefore

$$\lambda = \frac{300,000,000}{f} \tag{1}$$

The quantity 300,000,000 is the velocity of light in meters per second. The frequency is ordinarily expressed in kilocycles, abbreviated kc; or in megacycles, abbreviated mc. A low-frequency wave is seen from Eq. (1) to have a long wave length, while a high frequency corresponds to a short wave length.

The strength of a radio wave is measured in terms of the voltage stress produced in space by the electrostatic field of the wave and is usually expressed in microvolts stress per meter. Since the actual stress produced at any point by an alternating wave varies sinusoidally from instant to instant, it is customary to consider the intensity of such a wave to be the effective value of the stress, which is 0.707 times the maximum stress in the atmosphere during the cycle. The strength of the wave measured in terms of microvolts per meter of stress in space is exactly the same voltage that the magnetic flux of the wave induces in a conductor 1 meter long when sweeping across this conductor with the velocity of light. Thus the strength of a wave is not only the dielectric stress produced in space by the electrostatic field, but it also represents the voltage that the magnetic field of the wave will induce in cutting across a conductor. In fact the voltage stress produced by the wave can be considered as resulting from the movement of the magnetic flux of the same wave.

The minimum field strength required to give satisfactory reception of the radio wave varies with the amount of interference that is present. Under the most favorable conditions it is possible to obtain intelligible signals from waves having a strength as low as 0.1 μv per meter, but ordinarily interfering waves generated by both man-made and natural sources drown out such weak radio signals and make much greater field strengths necessary. Thus experience has shown that in rural areas it requires a field strength in the order of 100 μv per meter to give what the listener considers satisfactory service from a broadcast station, while in urban locations where the man-made interference is much greater, a field strength of 5000 to 30,000 μv per meter is needed to insure good reception at all times.

A plane parallel to the mutually perpendicular lines of electrostatic and electromagnetic flux is called the wave front. The wave always travels in a direction at right angles to the wave front, but whether it goes forward or backward depends upon the relative direction of the lines of electromagnetic and electrostatic flux. If the direction of either the magnetic or electrostatic flux is reversed the direction of travel is reversed, but reversing both sets of flux has no effect.

The direction of the electrostatic lines of flux is called the direction of polarization of the wave. If the electrostatic flux lines are vertical, as shown in Fig. 1, the wave is vertically polarized, while when the electrostatic flux lines are horizontal and the electromagnetic flux lines vertical, the wave is horizontally polarized.

Propagation of Radio Waves of Different Frequencies.—As radio waves travel away from their point of origin, they become attenuated as a result of spreading out and because of energy lost in travel. The amount of this attenuation depends upon the frequency of the wave, the time of day, and the season of the year.

Waves having frequencies below about 100 kc are called low-frequency waves, and travel with an attenuation that is small and relatively independent of time of day and season. These frequencies are therefore well suited for carrying on continuous radio communication over distances as great as 5000 miles.

Frequencies ranging from 100 to 1500 kc are referred to as medium radio frequencies. The distinguishing characteristic of these waves is high received energy at night time, particularly in the winter, and high attenuation during the day. This effect becomes more pronounced as the frequency is increased toward 1500 kc and is well illustrated by the characteristics of broadcast waves. Medium-frequency waves are suitable for covering distances up to several thousand miles at night but only a few hundred miles in the daytime.

As the frequency is increased from 1500 toward 6000 kc, the attenuation, particularly during the day, becomes less, and these waves, which are said to be of medium-high frequency, are therefore suitable for carrying on communication over distances in the order of several thousand miles. The attenuation is greater in the day than at night, and in summer than in winter, but is sufficiently small even under unfavorable conditions to permit communication over distances of one or more thousands of miles.

Waves of frequencies ranging from 6000 to about 30,000 kc are called high-frequency waves. They are capable of traveling great distances with small attenuation, but whether or not they reach a particular destination depends upon the conditions existing in the ionized regions in the upper atmosphere. In general the frequency best suited for reaching a very distant receiving point is highest in a summer day, somewhat lower in a winter day or summer night, and lowest during a winter night. In order to maintain reasonably continuous communication over great distances using high frequencies, it is necessary to change the frequency of transmission as conditions warrant.

Radio waves having frequencies less than about 12 kc have not been found useful for commercial radio communication because these waves require such large radiators for efficient radiation as to be impracticable from an economic point of view. Waves having frequencies higher than about 30,000 kc are likewise of limited use because they travel along straight lines and are not reflected back to earth by the ionized region in the upper atmosphere and therefore can be used only over distances so short that the earth's curvature permits a substantially straight-line path between transmitting and receiving points.

As a result of the different characteristics of propagation possessed by radio waves of different frequencies, each particular range of frequencies is best adapted for a particular type of communication service. The outstanding properties of the different classes of radio waves are

TABLE I.—CLASSIFICATION OF RADIO WAVES

Class	Frequency range, kilocycles	Wave-length range, meters	Outstanding characteristics	Principal uses
Low frequency....	Below 100	Over 3,000	Low attenuation at all times of day and of year	Long-distance trans-oceanic service requiring continuous operation
Medium frequency	100 to 1,500	3,000 to 200	Attenuation low at night and high in the daytime; greater in summer than winter	Range 100 to 500 kc used for marine communication, airplane radio, direction finding, etc. Range 550 to 1500 kc employed for broadcasting
Medium high frequency.	1,500 to 6,000	200 to 50	Attenuation low at night and moderate in the daytime	Moderate-distance communication of all types
High frequency...	6,000 to 30,000	50 to 10	Transmission depends solely upon conditions in Kennelly-Heaviside layer, so varies greatly with the time of day and season. Attenuation extremely small under favorable conditions	Long-distance communication of all kinds; airplane radio
Very high frequency.	Above 30,000	Below 10	Waves travel in straight lines and are not reflected by Kennelly-Heaviside layer, so can only travel between points in sight of each other	Little commercial use at present

tabulated in Table I, as well as the uses to which each class has been found best suited.

2. Radiation of Electrical Energy.—Every electrical circuit carrying alternating current radiates a certain amount of electrical energy in the form of electromagnetic waves, but the amount of energy thus radiated is extremely small unless all the dimensions of the circuit approach the order of magnitude of a wave length. Thus a power line carrying 60-cycle current with 20-ft. spacing between conductors will radiate practically no energy because a wave length at 60 cycles is more than 3000 miles, and 20 ft. is negligible in comparison. On the other hand a coil 20 ft. in diameter and carrying a 2000-kc current will radiate a considerable amount of energy because 20 ft. is comparable with the 150-meter wave length of the radio wave. The common radio antenna consisting of a vertical wire with a flat-top structure as shown in Fig. 2 is essentially a condenser in which one plate is the ground while the other plate is the flat top. Such an arrangement will be a good radiator of electrical energy when the ratio of height to wave length is appreciable, that is, at least 1:100, and preferably 1:10. Similarly a coil will be a good radiator of electrical energy provided the size of the coil is sufficiently great. The usual loop antenna consists of a coil and will be an efficient radiator to the extent that the ratio of loop diameter to wave length is appreciable.

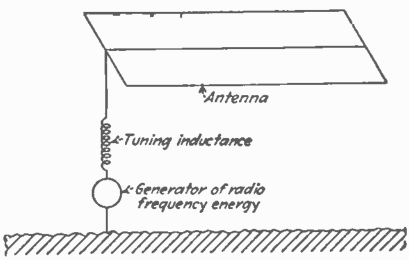


FIG. 2.—A simple system for producing radio waves, consisting of flat-top antenna, tuning inductance to bring antenna circuit into resonance at the frequency of the waves to be radiated, and a generator of radio-frequency energy.

It is apparent from these considerations that the size of radiator required is inversely proportional to the frequency. High-frequency waves can therefore be produced by a small radiator, while low-frequency waves require a high antenna system for effective radiation. The practical result of this fact is that the antennas of low-frequency transmitting stations are sometimes suspended from towers over 500 ft. high and yet are less efficient radiators than an antenna of one-tenth the height operating on a very high radio frequency.

Every radiator has directional characteristics as a result of which it sends out stronger waves in certain directions than in others. Thus, while a vertical wire radiates the same amount of energy in directions that are perpendicular to the wire, the radiation in a vertical plane varies from a maximum in a horizontal direction to zero in a vertical direction as shown in Fig. 3. Directional characteristics of antennas are taken advantage of to concentrate the radiation toward the point to which it is desired to transmit.

The amount of energy sent out from any radiating system is proportional to the square of the radio-frequency current that flows in the radiator. Since all of the common sources of radio-frequency energy are relatively low-voltage, high-current sources it is necessary that the radiating system offer a relatively low impedance to the radio-frequency energy that is to be transmitted. This is accomplished by tuning the antenna circuit to resonance with the frequency to be radiated, which makes the impedance of the antenna circuit low and enables a relatively small applied voltage to produce a very large antenna current and hence a high radiated energy. This is the only reason for tuning the transmitting antenna, as the mere tuning of the radiating systems to the frequency being transmitted does not increase the radiated power per ampere of current. The tuning is accomplished by inserting an inductance or a condenser in series with the antenna, as circumstances require. Thus in the flat-top antenna of Fig. 2 the antenna has a capacity reactance and so is tuned by the use of the inductance coil shown in the figure.

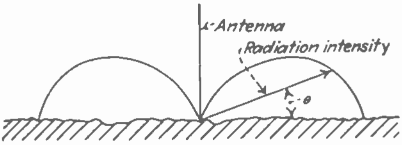


FIG. 3.—Directional characteristics in a vertical plane of radiation from an antenna consisting of a vertical wire. The length of the radius vector from the base of the antenna represents the relative intensity of wave radiated in the direction of the vector.

3. Generation and Control of Radio-frequency Power.—The radio-frequency power required by the radio transmitter is practically always obtained from a vacuum-tube oscillator. Vacuum-tube oscillators are capable of converting direct-current power into alternating-current energy of any desired frequency up to 300,000,000 cycles or higher. Over the range of frequencies used in commercial radio communication, *i.e.*, 12 to 30,000 kc, the power that can be obtained from vacuum-tube oscillators is in the order of tens to hundreds of kilowatts, and the efficiency with which the direct-current power is transformed into alternating-current energy is in the neighborhood of 50 per cent or higher.

A number of other methods of obtaining radio-frequency energy have been used at one time or another during the history of radio. Among these are the high-frequency alternator, the Poulsen arc, the frequency multiplier, and the oscillatory spark discharge. The high-frequency alternator is a special high-speed inductor-type alternator with many poles. Such alternators can generate several hundreds of kilowatts with reasonable efficiency when operating at frequencies of 50,000 cycles or less. A number of high-frequency alternators are now in commercial use, although it is improbable that any more will ever be built.[1] The frequency multiplier utilizes a moderately high-frequency

[1] A description of the widely used Alexanderson alternator is to be found in an article by Ernst F. W. Alexanderson, Trans-oceanic Radio Communication, *Proc. I.R.E.*, vol. 8, p. 263, August, 1920. For information on this as well as other types of

alternator from which the desired radio frequency is obtained by the use of magnetic-type harmonic generators. In this way it is possible with an alternator giving a frequency of 5000 cycles, to produce considerable quantities of power at frequencies of from 20,000 to 40,000 cycles. This type of arrangement at one time had a very prominent place in radio but has now practically disappeared.[1]

The Poulsen arc takes advantage of the negative resistance characteristic of an electric arc to convert direct-current power into radio-frequency energy. The arc as ordinarily employed takes place between carbon and copper electrodes in an atmosphere of hydrocarbon vapor and with a magnetic field at right angles to the axis of the arc. Such an arrangement when properly designed will generate large quantities of radio-frequency energy at a fair efficiency. The Poulsen arc operates most efficiently at frequencies below several hundred thousand cycles, but it will function after a fashion up to frequencies approaching 2000 kc. For the frequencies to which it is adapted the arc is a very simple and rugged generator of radio-frequency energy, and hundreds are now in regular use. It is, however, being slowly replaced by the vacuum-tube oscillator because of the latter's flexibility, frequency stability, and freedom from harmonics.[2]

The oscillatory spark discharge was the earliest and for many years the only method known for the generation of radio-frequency power. In this type of transmitter a condenser is charged to a high potential, which then breaks down a spark gap, permitting an oscillatory discharge through an inductance. This process is repeated about one thousand times each second. The spark transmitter thus radiates a series of wave trains, each of which is a damped sinusoidal oscillation. This method is capable of generating large quantities of radio-frequency energy with good efficiency but is in disfavor because the radiated waves are not simple sine waves but rather waves of a number of frequencies superimposed on each other. The result is excessive interference with radio signals being transmitted on slightly different frequencies.[3]

---

alternators see G. G. Blake, "History of Radio Telegraphy and Telephony," pp. 230–232.

[1] Extensive discussions of frequency multipliers are to be found in practically every book on radio communication written prior to about 1920.

[2] For further information regarding the Poulsen arc the reader is referred to the following articles: Leonard F. Fuller, The Design of Poulsen Arc Converters for Radio Telegraphy, *Proc. I.R.E.*, vol. 7, p. 449, October, 1919; P. O. Pederson, On the Poulsen Arc and Its Theory, *Proc. I.R.E.*, vol. 5, p. 255, August, 1917; Some Improvements in the Poulsen Arc, *Proc. I.R.E.*, vol. 9, p. 434, October, 1921, and vol. 11, p. 155, April, 1923.

[3] An extensive art has been developed in connection with the spark generator of radio-frequency energy, which though obsolete as far as radio communication is concerned, is still of value in other applications. Interested readers will find an excellent treatment of the spark generator in J. H. Morecroft, "Principles of Radio Communication," 2d ed., Chap. V.

Modulation.—The transmission of information by radio waves requires that some means be employed to control the radio waves by the desired intelligence. In radio telegraphy this control is obtained by turning the transmitter on and off in accordance with the dots and dashes of the telegraph code, as illustrated in Fig. 4. In radio telephony the transmission is accomplished by varying the amplitude of the radio-frequency wave in accordance with the pressure of the sound wave being transmitted. Thus the sound wave shown at *d* in Fig. 4 would be transmitted from a radio-telephone station by causing the amplitude of the radiated

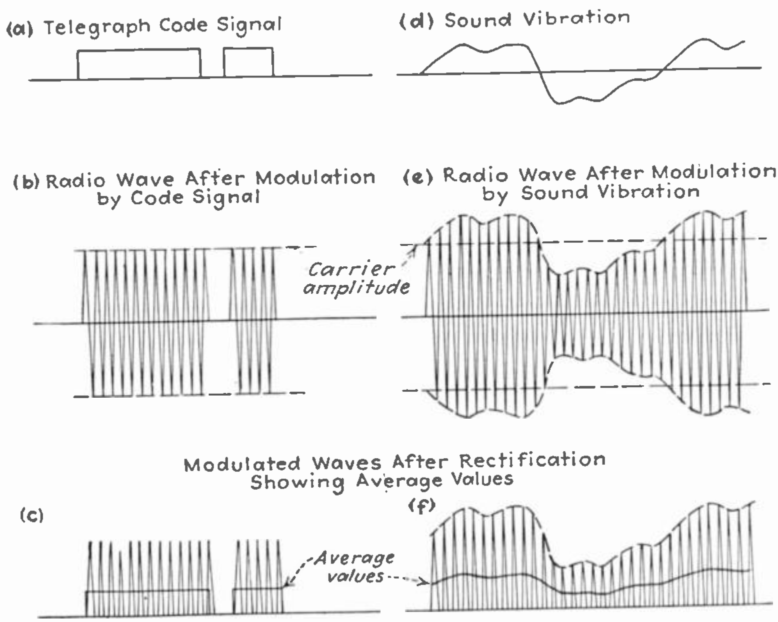


FIG. 4.—Diagram showing how a signal may be transmitted by modulating the amplitude of a radio wave, and how the original signal may be recovered from the modulated wave by rectification.

wave to vary as shown at *e*. In the transmission of pictures by radio a similar method is employed, in which the amplitude of the wave radiated at any time is made proportional to the light intensity of the part of the picture that is being transmitted at that instant.

When the amplitude of the alternating-current wave is varied from time to time, the wave is said to be modulated. Thus the wave radiated from a radio-telephone station is modulated by the voice or sound wave, while during the transmission of a picture the modulation is in accordance with the light intensities of different portions of the picture, and in the case of radio telegraphy the modulation is by the telegraph code. Except in the case of telegraphy the modulation of the radio-frequency wave is usually accomplished by means of vacuum tubes that control the amplitude of the generated or radiated high-frequency energy in accordance with the intelligence that is to be transmitted.

4. Reception of Radio Signals.—In the reception of radio signals it is first necessary to abstract energy from the radio waves passing the receiving point. After this has been done, the radio receiver must first separate the desired signal from other signals that may be present, and then reproduce the original signal from the radio waves. In addition, arrangements are ordinarily provided for amplification of the received energy so that the output of the radio receiver can be greater than the energy abstracted from the wave.

Any antenna system capable of radiating electrical energy is also able to abstract energy from a passing radio wave because the electromagnetic flux of the wave in cutting across the antenna conductors induces a voltage that varies in exactly the same way as the current flowing in the antenna radiating the wave. The amount of voltage induced in an antenna is equal to the product of the effective antenna height and the strength of the wave, and the resulting current flowing in the antenna is the current that is produced by this induced voltage acting against the impedance of the circuit. The energy represented by the induced current flowing in the antenna system is abstracted from the passing wave and will be greatest when the impedance of the antenna system has been reduced to a minimum by making the antenna circuit resonant to the frequency of the wave to be received.

The characteristics of an antenna when used for receiving radio signals are the same as in sending. Thus an efficient transmitting antenna is equally efficient when used for the reception of radio signals, and in general the size of receiving antenna will be proportional to the wave length being received, just as is the case in transmitting. The directional characteristics of an antenna are the same for reception as transmission; that is, if an antenna radiates most of the energy delivered to it in one direction, then that antenna will have a larger voltage induced in it by waves coming from the direction in which radiation is large than by those coming from other directions. Similar types of aerial systems are used for both radiating and receiving electromagnetic waves, the only difference being that receiving antennas can be less expensive and smaller than transmitting antennas because amplification at the receiver can readily make up for a less efficient antenna.

Since every wave passing the receiving antenna induces its own voltage in the antenna conductors it is necessary that the receiving equipment be capable of separating the desired signal from the unwanted signals that are also inducing voltages in the antenna. This separation is made on the basis of the difference in frequency between transmitting stations and is carried out by the use of resonant circuits which can be made to discriminate very strongly in favor of a particular frequency. It has already been pointed out that by making the antenna circuit resonant to a particular frequency the energy abstracted from radio waves of that frequency will be much greater than the energy from waves of other frequencies and this alone gives a certain amount of separation between signals. Still greater selective action can be obtained by the use of additional suitably adjusted resonant circuits located somewhere in the receiver in such a way as to reject all but the desired signal. The ability to discriminate between radio waves of different frequencies is called selectivity and the process of adjusting circuits to resonance with the frequency of a desired signal is spoken of as tuning.

Detection.—The process by which the signal being transmitted is reproduced from the radio-frequency currents present at the receiver is called detection. Where the intelligence is transmitted by varying the amplitude of the radiated wave, detection is accomplished by rectifying the radio-frequency currents. The rectified current thus produced varies in accordance with the signal originally modulated on the wave radiated at the transmitter and so reproduces the desired signal. Thus when the modulated wave shown at *e* of Fig. 4 is rectified, the resulting current is shown at *f* and is seen to have an average value that varies in accordance with the amplitude of the original signal. In the transmission of code signals by radio, the rectified current reproduces the dots and dashes of the telegraph code as shown at Fig. 4*c* and could be used to operate a telegraph sounder. When it is desired to receive the telegraph signals directly on a telephone receiver it is necessary to break up the dots and dashes at an audible rate in order to give a note that can be heard, since otherwise the telephone receiver would give forth a succession of unintelligible clicks.

Although intelligible radio signals have been received from stations thousands of miles distant using only the energy abstracted from the radio wave by the receiving antenna, much more satisfactory reception can be obtained if the received energy is amplified. This amplification may be applied to the radio-frequency currents before detection, in which case it is called radio-frequency amplification, or it may be applied to the rectified currents after detection, which is called audio-frequency amplification. The use of amplification makes possible the satisfactory reception of signals from waves that would otherwise be too weak to give an audible response. It also permits the strength of the signal as heard in the telephone receiver or any other indicating device to be raised to any desired volume, permitting radio reception in noisy locations, such as on airplanes, and making possible the use of loud-speakers.

The only satisfactory method of amplifying radio signals that has been discovered is by the use of vacuum tubes, and before such tubes were discovered radio reception had available only the energy abstracted from the radio wave by the receiving antenna. As a result of the small amplitude of this energy the signals were always weak, and radio reception from other than local stations was possible only in very quiet places.

5. Nature of a Modulated Wave.—The modulated wave that is sent out by a radio station represents an oscillation of varying amplitude and so consists of a number of waves of different frequencies superimposed upon each other. The actual nature of a modulated wave can be deduced by writing down the equation of the wave and making a mathematical analysis of the result. Thus, in the case of the simple sine-wave modula-

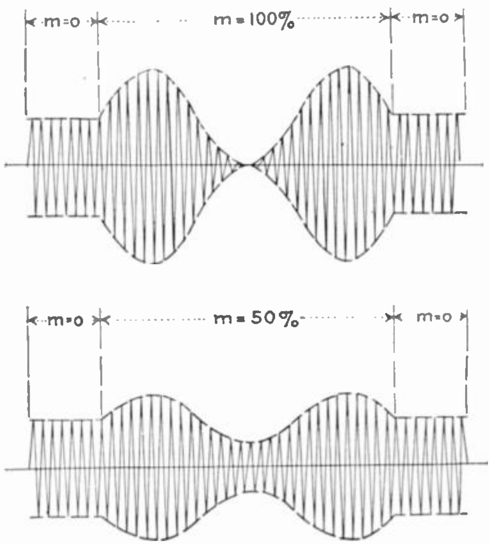


FIG. 5.—Waves having different degrees of simple sine-wave modulation.

tions shown in Fig. 5, the amplitude of the radio-frequency oscillation is given by $E = E_0 + mE_0 \sin 2\pi f_s t$ in which E_0 represents the average amplitude, f_s the frequency at which the amplitude is varied, and m is the ratio of amplitude variation from the average to the average amplitude, which is called the degree of modulation. Waves with several degrees of modulation are shown in Fig. 5. The equation of these modulated waves can be written as

$$e = E_0(1 + m \sin 2\pi f_s t) \sin 2\pi ft \tag{2}$$

in which f is the frequency of the radio oscillation. Multiplying out the right-hand side of Eq. (2) gives

$$e = E_0 \sin 2\pi ft + mE_0 \sin 2\pi f_s t \sin 2\pi ft$$

By expanding the last term into functions of the sum and difference angles by the usual trigonometric formula, the equation of a wave with simple sine-wave modulation is

$$e = E_0 \sin 2\pi ft + \frac{mE_0}{2} \cos 2\pi(f - f_s)t - \frac{mE_0}{2} \cos 2\pi(f + f_s)t \tag{3}$$

Equation (3) shows that the wave with sine-wave modulation consists of three separate waves. The first of these is represented by the term $E_0 \sin 2\pi ft$ and is called the carrier. Its amplitude is independent of the presence or absence of modulation and is equal to the average amplitude of the wave, which is independent of the degree of modulation. The two other components are alike as far as magnitude is concerned, but the frequency of one of them is less than that of the carrier frequency by an amount equal to the modulation frequency, while the frequency of the second is more than that of the carrier by the same amount. These two components are called the side-band frequencies and carry the intelligence that is being transmitted by the modulated wave. The frequency by which the side bands differ from the carrier frequency represents the modulation frequency, while the amplitude of the side-band components compared with the amplitude of the carrier determines the degree of modulation, *i.e.*, the size of the amplitude variations that are impressed upon the radiated wave.

When the modulation is more complex than the simple sine-wave-amplitude variation of Fig. 5 the effect is to introduce additional side-band components. The carrier wave is always the same, irrespective of the character of the modulation, and represents the average amplitude of the wave, but there is a pair of side-band frequencies for each frequency component in the modulation. Thus if the wave of a radio-telephone transmitter is modulated by a complex sound wave containing pitches of 1000 and 1500 cycles, the modulated wave will contain one pair of 1000-cycle side-band components and one pair of 1500-cycle side-band components. The amplitude of any side-band component is always one-half of the amplitude of that particular frequency component that is contained in the modulation envelope.

Significance of the Side Bands.—The carrier and side-band frequencies are not a mathematical fiction, but have a real existence, as is evidenced by the fact that the various frequency components of a modulated wave can be separated from each other by suitable filter circuits. The side-band frequencies can be considered as being generated as a result of varying the amplitude of the wave. They are only present when the amplitude is being varied, and their magnitude and frequency are determined by the character of the modulation. The carrier frequency on the other hand is independent of the modulation, being the same even when no modulation is present.

The intelligence transmitted by the modulated wave is carried by the side-band components and not in the carrier, *i.e.*, the intelligence is conveyed by the variations in the amplitude of the wave and not by the average amplitude. It is therefore desirable to put as much power into the side-band frequencies as is possible, which is equivalent to saying that the wave amplitude should be varied through the widest possible range. When the amplitude is carried clear to zero during the modulation cycle, the modulation is at a maximum, or 100 per cent, and the side bands contain the maximum amount of power possible. With sine-wave modulation such as shown in Fig. 5 this maximum side-band power is one-half of the carrier power. With degrees of modulation less than 100 per cent the side bands will contain correspondingly less power.

It is apparent that the transmission of intelligence requires the use of a band of frequencies rather than a single frequency. In speech and music there are important frequency components as high as 5000 cycles, so that speech and music modulated upon a wave will produce side-band components extending as far as 5000 cycles on each side of the carrier frequency. A radio-telephone station therefore utilizes a frequency band about 10,000 cycles wide in transmitting high-quality signals. If this entire band is not transmitted equally well through space and by the circuits through which the modulated wave currents must pass, then the side-band frequency components that are discriminated against will not be reproduced in the receiving equipment with proper amplitude. With telegraph signals the required side band is relatively narrow because the amplitude of the signals is varied only a few times a second, but a definite frequency band is still required. If some of the side-band components of the code signal are not transmitted, the received dots and dashes run together and may become indistinguishable.

CHAPTER II

CIRCUIT CONSTANTS

6. Inductance.—Whenever a current flows in an electrical circuit there is produced magnetic flux that links with (*i.e.*, encircles) the current. The amount of such flux actually present with a given current is measured in terms of a property of the circuit called the inductance and depends upon the arrangement of the circuit and the presence or absence of magnetic substances.

Inductance can be defined as the flux linkages per ampere of current producing the flux; that is

$$\text{Inductance } L \text{ in henries} = \frac{\text{flux linkages}}{\text{current producing flux}} 10^{-8} \qquad (4)$$

A flux linkage represents one flux line encircling the circuit current once. Thus in Fig. 6 flux line *aa* contributes eight flux linkages toward the coil inductance because it circles the current flowing in the coil eight times. On the other hand flux line *b* of the same coil contributes only one-half of a flux linkage toward the coil inductance because this particular line encircles only one-half of the coil current.

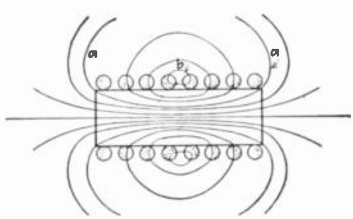


FIG. 6.—Flux and current distribution in typical single-layer air-cored inductance coil. The current density is indicated by the depth of shading.

The inductance of an electrical circuit is computed by assuming a convenient current flowing in the circuit, determining the magnetic flux produced by the current, and then counting up the total number of flux linkages present in the circuit. The inductance in henries is this total number of flux linkages multiplied by 10^{-8} and divided by the circuit current.

The inductance of a circuit is the measure of a number of electrical properties. When the current varies, as, for example, in the case of an alternating current, the amount of flux linking with the current also varies and in doing so induces a voltage in the circuit just as magnetic flux always does when cutting a conductor. The magnitude of this induced voltage can be expressed in terms of the circuit inductance and rate of change of current.

$$\text{Induced voltage in circuit caused by change in current} = -L\frac{di}{dt} \qquad (5)$$

A positive voltage in this equation is a voltage acting in the direction of the current. The voltage that must be applied to the circuit to overcome the induced potential is equal and opposite to that given by Eq. (5). It will be observed that the voltage induced in the circuit by an increase of current is always in a direction that tends to reduce the current, while a decreasing current gives a negative di/dt and produces a positive induced voltage that tends to keep the current flowing. For the particular case where the current is alternating, $di/dt = j\omega I$, and substitution in Eq. (5) shows that the applied voltage required to overcome the induced voltage is the familiar quantity $j\omega LI$.

Flux linkages can be created only by the expenditure of energy, and these linkages represent electrical energy stored in the form of magnetic flux. The amount of such storage depends upon the circuit inductance and current according to the relation[1]

$$\text{Energy in joules stored in magnetic field} = \tfrac{1}{2}LI^2 \qquad (6)$$

Inductance Coils with Non-magnetic Cores.—Inductance coils intended for use at radio frequencies almost always have non-magnetic cores. This is because the energy losses in magnetic materials at radio frequencies are so great as to make magnetic cores out of the question except at the very lowest radio frequencies. Inductance coils with non-magnetic cores are also used at audio frequencies when it is necessary to avoid effects introduced by the variation of the permeability of magnetic materials with changing flux density.

Skilled mathematicians have derived formulas that give the inductance of all the commonly used types of coils with non-magnetic cores in terms of the coil dimensions. These formulas are usually both complicated and hard to derive because of the difficulty encountered in calculating the magnetic flux produced by a current flowing in the coil. In order to make such formulas of practical value they are always simplified by the use of coefficients. Thus the inductance of a single-layer solenoid, such as shown in Fig. 6, is given by the formula

$$\text{Inductance in microhenries} = N^2dF \qquad (7)$$

where

N = number of turns
d = diameter of coil measured to center of wire
F = constant that depends only upon the ratio of length to diameter and is given in Fig. 7.

The quantity F depends in a complicated way upon the ratio of coil length to diameter, but once this relationship has been determined, the

[1] Equations (5) and (6) are direct consequences of Eq. (4) and are derived in most textbooks on alternating currents. For example, see V. Karapetoff, "The Magnetic Circuit," Chap. X.

value of F for different ratios can be computed once for all and presented in a curve such as Fig. 7 or by a table.

Formulas giving the inductance of all the commonly used types of coils having non-magnetic cores are given in Appendix A. By the aid of these formulas it is possible to compute the inductance of practically any type of coil with as great accuracy as the geometrical dimensions are ordinarily known. Guessing at the number of turns and the coil dimensions to obtain a desired inductance is therefore never necessary.

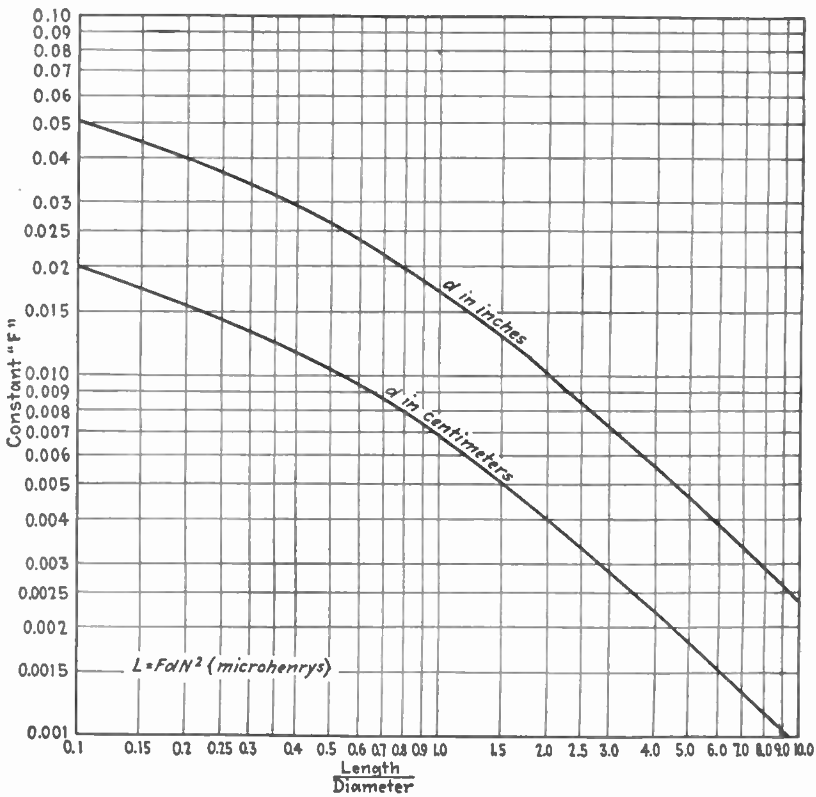


FIG. 7.—Values of constant F for use in Eq. (7), to obtain the inductance of single-layer solenoid.

The inductance of all coils with non-magnetic cores is proportional to the square of the number of turns if the dimensions such as length, diameter, depth of winding, etc., are kept constant as the number of turns is altered. The reason for this behavior lies in the fact that if the coil dimensions are kept constant, the amount of magnetic flux produced by a given coil current and the number of times each flux line links with the coil current are both proportional to the number of turns.

The inductance of all coils having the same number of turns and the same shape is always proportional to the size (*i.e.*, to a linear dimension, such as length or radius) of the coil. Thus if two coils have the same number of turns, but one is twice as big as the other in every dimension (such as diameter, length, width, depth of winding, etc.), then the larger coil will have twice the inductance of the smaller one. This rule can be verified by examining the inductance formulas in Appendix A and results from the fact that the cross section of the flux paths is proportional to the square of the linear dimension of the coil, while the length of these paths varies directly as the linear dimension.

Inductance Coils with Magnetic Cores.—Magnetic cores greatly increase the flux produced by a given magnetomotive force (*i.e.*, a given number of ampere turns), and so make it possible to obtain a large inductance in a small volume and with the use of only a small amount of material. Because of this, inductance coils with magnetic cores are used in preference to coils with non-magnetic cores wherever this is possible. There are, however, two important factors limiting the usefulness of magnetic materials in coils intended for communication work. In the first place the permeability of all magnetic materials varies with the flux density and the previous magnetic history so that the inductance of a coil with a magnetic core depends somewhat upon the current flowing through the winding, and, furthermore, when this current is alternating, the inductance will be different at different parts of the cycle, which causes the production of harmonics. In the second place the eddy-current losses increase so rapidly with frequency as to make magnetically cored coils useless even at the lower radio frequencies unless special methods of subdividing the core are employed. Coils with magnetic cores find their principal usefulness where it is not essential that the inductance be absolutely independent of current, and where the frequency is at the same time below 15,000 cycles.

It is relatively easy to compute the inductance of coils with magnetic cores with at least fair accuracy because the high permeability of magnetic materials restricts practically the entire flux to the magnetic core. This results in simple flux paths that can be readily handled in computations when the magnetic characteristics of the core material are known. The inductance of ordinary types of coils with magnetic cores is given to a good approximation by the formula[1]

$$L = 1.257N^2P \cdot 10^{-8} \tag{8}$$

where

L = inductance in henries
N = number of turns
P = permeance of the magnetic circuit of coil, assuming permeability of air as unity.

[1] Equation (8) is derived as follows: The flux produced by a current I flowing in a coil of N turns with a core of permeance P is $1.257INP$ (*i.e.*, magnetomotive force

This equation can be used in the calculation of inductance coils having a short air gap in the magnetic circuit provided the permeance P in the equation is interpreted to mean the permeance of the complete magnetic circuit, which in this case consists of the air gap in series with magnetic material.

The inductance of coils with magnetic cores is proportional to the square of the number of turns if the dimensions of the coil are kept constant, and to the linear dimensions if the number of turns and proportions are the same. These two relations are identical with those that exist in coils with non-magnetic cores and arise from the same reasons.

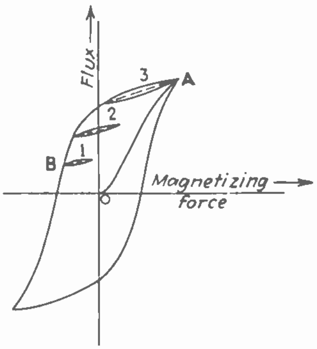


FIG. 8.—Typical hysteresis loop showing displaced minor hysteresis loops 1, 2, and 3, obtained when a small alternating current is superimposed upon a direct-current magnetization. The effective permeability to the superimposed alternating current is proportional to the slope of the line joining the tips of the displaced hysteresis loop.

Incremental Permeability.—The behavior of coils having magnetic cores may become very complicated when the current through the winding contains several components of different frequencies. The most important example of this character is when an alternating current is superimposed upon a direct current. In this case the effective inductance offered to the alternating-current component depends upon the magnitudes of both currents and upon the previous magnetic history of the core. When a core that has been thoroughly demagnetized is first magnetized the relation between current in the winding and core flux is the usual B-H curve shown at OA in Fig. 8. If the current is then successively reduced to zero, reversed, brought back to zero, reversed to the original direction, etc., the flux goes through the familiar hysteresis loop shown in Fig. 8. A direct current flowing through the coil winding brings the magnetic state of the core to some point on the hysteresis loop, such as B in Fig. 8, and it is to be observed that the flux with a given value of direct current depends not only upon the magnitude of current but also upon which side of the hysteresis loop one is on, and upon the width

times permeance of the magnetic circuit times $4\pi/10$), and as every flux line links with every turn, the number of flux linkages is $1.257IN^2P$, which substituted in Eq. 4 gives Eq. (8). This derivation is approximate in that it neglects the small amount of magnetic flux that does not stay confined to the magnetic material and so gives results that are slightly low. The additional inductance contributed by these flux lines can be calculated with fair accuracy by estimating the flux linkages contributed by this extra flux. Calculations of this type are described at length in many books dealing with magnetic-circuit calculations, as for example, V. Karapetoff, "The Magnetic Circuit."

of the loop, both of which are determined by the previous magnetic history.

When an alternating current is now superimposed on this direct current the result is to cause the flux in the iron core to go through a displaced minor hysteresis loop that is superimposed upon the usual hysteresis curve. Examples of such displaced hysteresis loops are shown at 1, 2, and 3 in Fig. 8. One tip of every displaced loop is always on the normal loop (or upon the *B-H* curve), and the two tips of a minor loop always correspond to the maximum and minimum values of current in the coil (*i.e.*, the direct current respectively plus and minus the crest a-c current). The permeance of the core to the alternating current that is superimposed upon the direct current, and hence the effective coil inductance offered to the alternating current, is proportional to the slope of the line (shown dotted in Fig. 8) joining the two tips of the displaced hysteresis loop. The permeability shown by a magnetic material to alternating currents superimposed upon direct currents is called the incremental permeability (*i.e.*, the permeability to a small increment of alternating magnetomotive force) and has been the subject of considerable study.[1] The most important characteristics of incremental permeability (and hence of the inductance to superimposed alternating currents) are: (1) For a given alternating current the incremental permeability (and hence the inductance) to the superimposed alternating current will be less the greater the direct current upon which the alternating current is superimposed; and (2) with a given direct current the incremental permeability, and hence the inductance to the alternating current, will increase as the superimposed alternating current becomes larger. These characteristics are clearly brought out by the results shown in Fig. 9.

When an alternating current of one frequency is superimposed upon a current of another frequency instead of upon direct current the situation becomes very complicated. The general result is that the effective inductance to each frequency is altered by the presence of the other current, and furthermore the inductance offered to each frequency component varies cyclically at the frequency of the other component.

When it is necessary to have an inductance that does not change appreciably with the current flowing through the windings, or when it is desired to minimize the various effects that result when the magnetic flux is not proportional to the coil current, an air gap is left in the magnetic circuit. The total reluctance of the magnetic circuit of the coil is then the sum of the reluctances of the air gap and of the magnetic part of the magnetic circuit, and by making the former much the largest (from five to twenty-five times that of the latter), variations in the reluctance (or permeability) of the core material have only a small effect on the total

[1] See Spooner, Thomas, Permeability, *Trans. A.I.E.E.*, vol. 42, p. 340, 1923.

reluctance of the magnetic circuit. The presence of an air gap reduces the detrimental effects of the magnetic material but does not eliminate them, and the length of gap to use therefore depends upon the extent to which the non-linear properties of the magnetic material must be minimized.

The length of air gap usually required is only a small fraction of the total length of the magnetic circuit because the permeability of the magnetic material is hundreds of times that of air. Thus, although an air gap in the magnetic material reduces the coil inductance to a much

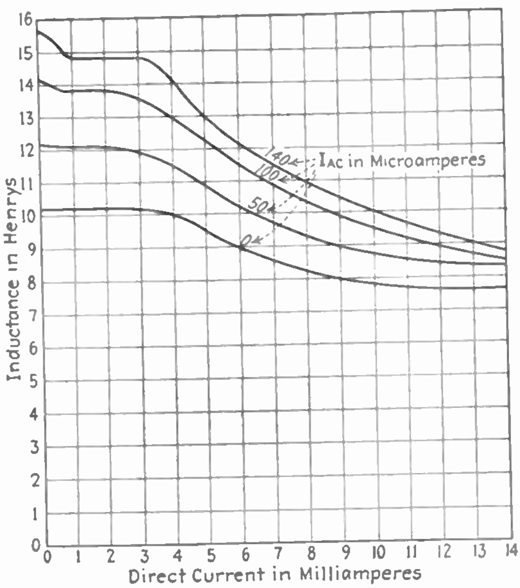


FIG. 9.—Curve giving measured inductance of a coil as a function of direct-current magnetization with several values of superimposed alternating current, showing how the inductance decreases with increased direct current and reduced alternating current.

smaller value than would be obtained with no gap, enough magnetic material is present in the coil greatly to increase the permeance of the magnetic circuit and hence the inductance over the value obtainable with no magnetic material whatsoever in the core. The necessary air gap can be supplied in a number of ways. Under some conditions the air gaps at the joints in the laminations are sufficient. In other cases a non-magnetic spacer is placed at a butt joint to increase the gap, while sometimes the cores are made up of insulated particles of powdered magnetic material pressed together to form a core that has minute air gaps distributed throughout its structure.

Magnetic Materials.—The magnetic materials most widely used in communication work are ordinary transformer iron and silicon steel.

There are in addition two remarkable magnetic alloys that are finding increasing use in communication work. These are permalloy, which is a nickel-iron alloy characterized by a remarkably high permeability at low flux densities, and perminvar, a nickel-iron-cobalt alloy, that has a moderately high permeability at low flux densities but is chiefly remarkable for a permeability that is practically constant for magnetomotive forces up to about two ampere turns per centimeter, coupled with an extremely small hysteresis loss.[1]

7. Mutual Inductance and Coefficient of Coupling.—When two inductance coils are so placed in relation to each other that flux lines produced by current in one of the coils link with the turns of the other coil as shown in Fig. 10*a* the two inductances are said to be inductively coupled. The effects which this coupling produces can be expressed in terms of a property called the mutual inductance, which is defined by the relation:

$$\text{Mutual inductance } M \text{ in henries} = \frac{\begin{pmatrix}\text{flux linkages in second coil} \\ \text{produced by current in first coil}\end{pmatrix}}{\text{current in first coil}} 10^{-8} \qquad (9)$$

$$= \frac{\begin{pmatrix}\text{flux linkages in first coil} \\ \text{produced by current in second coil}\end{pmatrix}}{\text{current in second coil}} 10^{-8} \qquad (9a)$$

Formulas (9) and (9*a*) are equivalent and give the same value of mutual inductance. The flux linkages produced in the coil that has no current in it are counted just as though there was a current in this coil, so that the number of times a flux line would encircle an imaginary coil current is the number of linkages contributed by this particular line. In adding up the flux linkages it is important to note that different flux lines may link with the same coil in opposite directions, in which case the total number of linkages is the difference between the sums of positive and negative linkages. The mutual inductance may therefore be positive or negative depending upon the direction of the linkages.

The problem of calculating mutual inductance is similar in all respects to the problem of computing inductance, and formulas have been worked out by which the mutual inductance can be calculated with good accuracy in all the ordinary types of configurations. Some of the more important cases are treated in Appendix A.

Coefficient of Coupling.—The maximum value of mutual inductance that can be obtained between two coils having inductances L_1 and L_2 is $\sqrt{L_1L_2}$. The ratio of mutual inductance actually present to this

[1] For more detailed descriptions of these alloys see: H. D. Arnold and G. W. Elman, Permalloy, a New Magnetic Material of Very High Permeability, *Bell System Tech. Jour.*, vol. 2, p. 101, July, 1923; and G. W. Elman, Magnetic Properties of Perminvar, *Bell System Tech. Jour.*, vol. 8, p. 21, January, 1929.

maximum possible value is called the coefficient of coupling, which can therefore be expressed by the relation:

$$\text{Coefficient of coupling} = k = \frac{M}{\sqrt{L_1 L_2}} \tag{10}$$

The coefficient of coupling is a convenient constant because it expresses the extent to which the two inductances are coupled, independently of the size of the inductances concerned. In air-cored coils, such as used in radio, a coupling coefficient of 0.5 is considered high and is said to represent "close" coupling, while coefficients of only a few hundredths represent "loose" coupling.

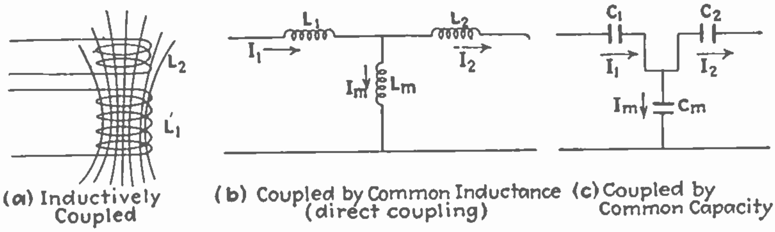


FIG. 10.—Several simple methods of coupling two circuits.

When two coils of inductance L_1 and L_2 between which a mutual inductance M exists are connected in series, the equivalent inductance of the combination is $L_1 + L_2 \pm 2M$. The term $2M$ takes into account the flux linkages in each coil due to the current in the other coil. These mutual linkages may add to or subtract from the self-linkages, depending upon the relative direction in which the current passes through the two coils. When all linkages are in the same direction the total inductance of the series combination exceeds by $2M$ the sum of the individual inductances of the two coils; while if the mutual linkages are in opposition to the other linkages the inductance of the combination is less than the sum of the two coil inductances by $2M$. This property can be taken advantage of to measure the mutual inductance between two coils. The procedure is to connect the two coils in series and measure the equivalent inductance of the combination. The connections to *one* of the coils are then interchanged and the equivalent inductance is measured again. The difference between the two measured inductances is then $4M$.

Any two circuits so arranged that energy can be transferred from one to the other are said to be coupled even though this transfer of energy takes place through some means such as a condenser, resistance, or inductance, that is common to the two circuits rather than by the aid of a mutual inductance. Examples of various methods of coupling are shown in Fig. 10. Any two circuits that are coupled by a common impedance have a coefficient of coupling that is equal to the ratio of the common impedance to the square root of the product of the total impedances of the same kind as the coupling impedance that are present in the two circuits. Thus with case b in Fig. 10 where the coupling is furnished by the common inductance L_m the total inductances of the two circuits are $L_1 + L_m$ and $L_2 + L_m$, and the coefficient of coupling is given by the equation

$$\text{Coefficient of coupling } k \text{ for Fig. } 10b = \frac{L_m}{\sqrt{(L_1 + L_m)(L_2 + L_m)}} \qquad (11)$$

In Fig. 10c the coupling element is a common capacity C_m, and the coefficient of coupling is[1]

$$\text{Coefficient of coupling for Fig. } 10c = \frac{\sqrt{C_1C_2}}{\sqrt{(C_m + C_1)(C_m + C_2)}} \qquad (12)$$

8. Condensers and Dielectrics.—Electrostatic capacity exists wherever an insulator (*i.e.*, dielectric) separates two conductors between which a difference of potential can exist. Applying a voltage between these conductors causes an electric charge to flow into them with the resulting production in the dielectric of an electrostatic stress that represents stored electrical energy.[2] The combination of the conducting plates separated by the insulating dielectric is called a condenser. The amount of energy stored in this way in the electrostatic field of the condenser depends upon the voltage that is applied to the electrodes, and upon the electrostatic capacity of the combination. The capacity increases directly with the area and inversely with the thickness of the dielectric and depends to a considerable extent upon the kind of dielectric being used. Electrostatic capacity is measured in farads, which is a capacity of such a size that when 1 volt is applied across a capacity of 1 farad a charge of 1 coulomb (equivalent to 1 amp. of current flowing for 1 sec.) will be stored. The farad is a very large unit and is frequently subdivided into the microfarad (abbreviated μf).

[1] Equation (12) is readily derived by following the rule that the coefficient of coupling is equal to the ratio of common impedance to the square root of the product of primary and secondary impedance of the same kind as the coupling element. Thus, in Fig. 10c, the capacity of the primary is $\frac{C_1C_m}{C_1 + C_m}$, and that of the secondary is $\frac{C_2C_m}{C_2 + C_m}$. The coupling reactance is $1/\omega C_m$, while the primary and secondary reactances are $\frac{C_1 + C_m}{\omega C_1C_m}$ and $\frac{C_2 + C_m}{\omega C_2C_m}$, respectively. The coefficient of coupling is then

$$k = \frac{1/(\omega C_m)}{\sqrt{\frac{C_1 + C_m}{\omega C_1C_m} \cdot \frac{C_2 + C_m}{\omega C_2C_m}}}$$

which reduces to Eq. (12).

[2] For many years it was thought that the charge resided in the dielectric instead of the conductors, but it has been recently discovered that in circumstances where the charge appeared to exist in the dielectric it was really on the surface of the dielectric in a conducting moisture film that acted as an extension of the electrodes.

The principal properties of condensers are summarized in the following equations, in which C is in farads, and E in volts.[1] The energy stored in a capacity C when charged to a voltage E is

$$\text{Stored energy in joules} = \tfrac{1}{2}CE^2 \tag{13}$$

The charge stored in a capacity C when charged to a voltage E is

$$\text{Stored charge in coulombs} = CE \tag{14}$$

The current that flows into or out of the condenser at any instant is the rate of change of charge, and so is given by the relation

$$\text{Current flowing into condenser} = \frac{dQ}{dt} = C\frac{dE}{dt} \tag{15}$$

When the voltage is sinusoidal, $C\frac{dE}{dt}$ reduces to $j\omega CE$. In the case where the condenser electrodes are large plates having a constant spacing, the capacity is given by the equation

$$\text{Capacity in } \mu\mu\text{f} = 0.08842K\frac{A}{d} \tag{16}$$

where A is the area of active dielectric in square centimeters, d the spacing between plates in centimeters, and K a constant, called the dielectric constant, that depends upon the dielectric material and is substantially independent of frequency. Values of K for common dielectrics are given in Table II.[2]

TABLE II.—CHARACTERISTICS OF DIELECTRICS AT RADIO FREQUENCIES WITH NORMAL ROOM TEMPERATURE

Material	Dielectric constant	Power factor
Air	1.00	0.000
Mica	2.94	0.0004
Hard rubber	2.50 to 3.00	0.007 to 0.014
Glass	4.90 to 7.00	0.004 to 0.016
Bakelite derivatives	3.50 to 6.00	0.037 to 0.073
Celluloid	4.10	0.042
Fiber	4 to 6	0.04 to 0.06
Wood (without special preparation)		
Oak	3.3	0.039
Maple	4.4	0.033
Birch	5.2	0.065
Transformer oil	2.5	
Castor oil	5.0	
Porcelain	4.4	

[1] The derivation of these equations is to be found in most textbooks of physics and electrical engineering and will therefore not be given here.

[2] The data for Tables II and III were in large measure obtained from the following sources: E. T. Hoch, Power Losses in Insulating Materials, *Bell System Tech. Jour.*,

TABLE III.—EFFECT OF FREQUENCY AND TEMPERATURE ON DIELECTRIC CHARACTERISTICS

Material	Temperature, degrees centigrade	Frequency, kilocycles	Dielectric constant	Power factor
Bakelite derivative	21	500	5.6	0.054
	71	500	6.9	0.11
	120	500	10.4	0.38
Hard rubber	21	500	3.0	0.009
	71	500	3.1	0.021
	120	500	3.2	0.065
Bakelite derivative	21	295	5.9	0.051
	21	500	5.8	0.051
	21	670	5.7	0.051
	21	1040	5.6	0.058
Hard rubber	21	210	3.0	0.009
	21	440	3.0	0.009
	21	710	3.0	0.009
	21	1126	3.0	0.010

Losses in Condensers.—A perfect condenser when discharged gives up all of the electrical energy stored in it, but this ideal is never completely realized since all actual condensers return less energy than was expended in charging them. The energy not delivered by the discharge represents losses that cause the condenser to have a power factor that is not zero. Practically all of the electrical losses of the usual condenser are due to dielectric hysteresis, which is the result of molecular friction in the dielectric and is similar in character to magnetic hysteresis. The energy lost in the condenser as a result of dielectric hysteresis appears in the form of heat generated within the dielectric.

The power factor of a condenser is determined by the type of dielectric used and is practically independent of the condenser capacity, the applied voltage, the voltage rating, or the frequency. This constancy of the power factor under such widely different conditions makes the power factor of a dielectric of fundamental importance where the losses in the condenser are to be considered. While the power factor of a condenser is determined by the dielectric used it is also affected by the conditions under which the dielectric operates. Thus the power factor always becomes higher as the temperature is raised. Moisture absorbed in the dielectric also has an extremely unfavorable effect on the power factor, and with dielectrics, such as paper and fiber which readily absorb moisture, the power factor may become extremely poor if the humidity is appreciable.

vol. 1, p. 110, October, 1922; R. V. Guthrie, Jr., Electrical Constants of Dielectrics for Radio Frequencies, *Proc. I.R.E.*, vol. 12, p. 841, December, 1924.

Table II gives information relative to the power factor of various dielectrics commonly used in radio work. In Table III the effect which changes in frequency and temperature have upon the power factor and dielectric constant is given for several dielectrics. The reason for the constancy of the power factor with changes in frequency is that the dielectric loss per cycle is almost independent of the number of cycles per second, so that irrespective of the frequency a constant proportion of the energy that is supplied to the condenser disappears as dielectric hysteresis.

The losses of a dielectric are sometimes expressed in terms of the angle by which the current flowing into the condenser fails to be 90° out of phase with the voltage. This angle is called the phase angle of the dielectric, and its value in radians is equal to the power factor. Thus a power factor of 0.01 represents a phase angle of 0.01 radian or 0.573°.

Equivalent Series and Shunt Resistance.—While a comparison of the losses of different dielectrics can be most satisfactorily expressed in terms

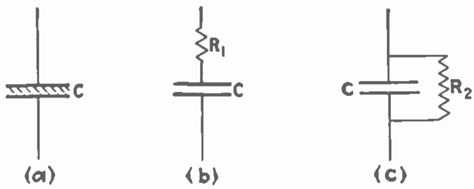


FIG. 11.—Representation of imperfect condenser by a perfect condenser of same capacity with series resistance and by a perfect condenser with shunt resistance.

of power factor or phase angle, the effect which the losses of a particular condenser have on the associated electrical circuits can be taken into account by replacing the actual condenser with a perfect condenser of the same capacity and having a resistance in series as shown in Fig. 11*b* or a resistance in parallel as in Fig. 11*c*. The value of the series or shunt resistance is so selected that the power factor of the perfect condenser associated with the resistance is the same as the power factor of the actual condenser. The value of the series resistance can be computed in terms of the power factor, condenser capacity C, and frequency f in the usual way and is given to a high degree of accuracy by the equation

$$\text{Series resistance} = R_1 = \frac{\text{power factor}}{2\pi fC} \tag{17}$$

In the same way the shunt resistance that can be used to represent the actual losses of the condenser is related to the power factor, capacity, and frequency to a high degree of accuracy by the equation

$$\text{Shunt resistance} = R_2 = \frac{1}{(2\pi fC)(\text{power factor})} \tag{18}$$

The relationship between the shunt resistance and series resistance that can be used to represent the losses of a given condenser at a given frequency can be obtained by combining Eqs. (17) and (18) to eliminate the power factor, which yields the result

$$\left.\begin{aligned} R_1 &= \frac{1}{R_2(2\pi fC)^2} \\ R_2 &= \frac{1}{R_1(2\pi fC)^2} \end{aligned}\right\} \tag{19}$$

Both the equivalent series and shunt resistances of a condenser having a constant power factor are independent of the voltage applied to the condenser but vary inversely with the condenser capacity and the frequency.

In addition to dielectric hysteresis it is possible for other losses to exist in a condenser. Among these are the losses caused by resistance of the condenser plates, the loss from leakage currents passing through the dielectric, and losses arising from corona discharges. The ohmic loss is entirely negligible if the condenser is properly constructed. The effects of leakage are wholly negligible at ordinary frequencies and only become of importance when a condenser is used to block off a direct current, in which case even a small leakage may be of importance. The leakage of a condenser is greatly increased by moisture in the dielectric and by raising the temperature of the dielectric. The leakage is ordinarily expressed in terms of the equivalent resistance between the condenser terminals and is inversely proportional to the capacity of the condenser. Corona occurs only in high-voltage condensers and is to be avoided in all circumstances as it represents a large power loss and in addition starts chemical actions that cause rapid deterioration of all dielectrics in the vicinity.

Condensers with solid dielectrics are ordinarily freed of moisture and then impregnated with some type of insulating compound during manufacture. This has the effect of making the condenser moisture-proof and thus insures that the power factor and leakage will be maintained at a low value irrespective of weather conditions. In high-voltage condensers such impregnation when done by a vacuum process also has the effect of eliminating air spaces between the dielectric and the condenser plates, which removes the possibility of corona and therefore greatly increases the voltage that may be safely applied to the condenser.

Condensers for Radio Purposes.—Paper, mica, and air are the dielectrics most frequently employed in condensers used in radio work. Paper dielectric is inexpensive and gives a large capacity in a small volume but has the disadvantage of fairly high electrical losses. Paper condensers are usually constructed by insulating two strips of tin or copper foil with enough layers of paper strips to give the necessary voltage strength and then rolling these into a compact bundle that is made moisture-proof by sealing with insulating compound. The dielectric losses and the insulation strength of different grades of paper vary widely, and, unless paper especially made for condenser use is employed, the leakage and dielectric hysteresis will be high and the voltage rating low.

Mica is widely used as a dielectric in high-grade condensers because of its high voltage strength and its superiority over practically all other solid dielectrics in the matter of losses. This superiority is especially pronounced at radio frequencies. Mica is expensive, however, and is not used where paper is satisfactory (*i.e.*, when the frequency is low, or when the losses are unimportant), except when the condenser capacity is so small that the cost of the dielectric is insignificant. Mica condensers are constructed by piling alternate layers of metallic and mica sheets upon one another and then connecting alternate layers of conductors together to form the two electrodes of the condenser. This type of construction is necessary because mica is damaged when rolled. The quality of the mica used is very important, and only the best grades are satisfactory for electrical purposes.

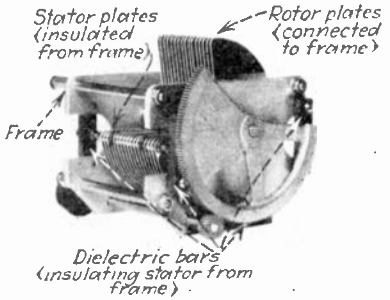


FIG. 12.—Air-dielectric variable condenser showing details of construction. In this condenser the rotor plates are connected to the end plates, which is known as the grounded-rotor type of construction.

Air is a perfect dielectric, having no dielectric hysteresis loss whatsoever. The principal losses that condensers with air dielectric have when operating below the corona voltage are those introduced by the dielectric material used in mounting the plates. Air condensers are extremely bulky in proportion to capacity because of the low dielectric constant of air and because a relatively large clearance between plates is necessary to insure separation in view of mechanical irregularities.

Variable Condensers with Air Dielectric.—The chief use of air as a dielectric is in condensers used to adjust the resonant frequency of tuned circuits, where a small condenser having low losses at radio frequencies and a capacity that is continuously variable is required. Air-dielectric condensers for this purpose are usually constructed as shown in Fig. 12, with a series of fixed plates, spaced by washers, as one electrode, and a series of similarly spaced rotating plates as the other electrode. The rotating plates mesh with the fixed plates without touching them, thus causing the condenser to have a capacity determined by the angle of rotation. The way in which the capacity of the variable condenser varies with the angle of rotation can be controlled either by cutting the stationary or the rotating plates to a special shape, or by use of an air gap between the two sets of plates that varies with the angle of rotation. In order to insure that the condenser will have low losses the dielectric used to insulate the fixed and moving plates from each other should be arranged in a way that will keep the voltage gradient in it low and should have the lowest possible power factor. Hard rubber is widely used for this purpose and is superior to bakelite, which has a much greater power factor. The details of construction of a typical condenser can be seen in Fig. 12. In this condenser the metal end plates are connected to the moving plates by a hair-spring and act as a shield for the fixed plates from which they are insulated by the strips of dielectric. The usual variable air-dielectric condenser has a capacity of $0.0005 \mu f$ or less. Larger sizes are rather bulky, and air condensers with capacities in excess of $0.002 \mu f$ are rarely made.

Condensers in which the fixed and rotating plates are both semicircular have a capacity that is roughly proportional to the angle of rotation and are therefore called straight-line-capacity condensers. For some purposes it is desirable to have a condenser in which the capacity is proportional to the square of the angle of rotation. Such a condenser, when used to resonate with a fixed inductance, causes the wave length at resonance to be proportional to the angle of rotation, and so is called a straight-line wave-length condenser. Still other types of condensers are so constructed that the resonant frequency is proportional to the angle of rotation, thus giving a straight-line-frequency type of condenser. The shape of the plates required in the straight-line wave-length and straight-line-frequency condensers is affected by the distributed capacity of the coil that is being tuned by the condenser, so that in order to get absolute straight-line wave-length or straight-line-frequency characteristics it is necessary to design the condenser to match a particular coil. The approximate shapes of rotor plates that will give these effects are given in Fig. 13, which also gives the variation of capacity with angle of rotation for these types of condensers.[1]

Miscellaneous Types of Condensers.—While paper, mica, and air are the principal dielectrics used in radio condensers, other materials have a limited field of usefulness. Where a large capacity or a high-voltage strength in a variable condenser of reasonable proportions is desired, it is possible to immerse an air-dielectric condenser in either transformer oil or castor oil. Transformer oil will slightly more than double the capacity obtained with air, and castor oil will make the capacity five times as great. Both of these oils also increase the possible voltage that may be applied to the condenser. The disadvantage of oils is that they

[1] Formulas for calculating the shape of plates in straight-line-frequency and wave-length condensers are to be found in: O. C. Roos, Simplified S.L.F. and S.L.W. Design, *Proc. I.R.E.*, vol. 14, p. 773, December, 1926.

have high dielectric losses except when entirely free from moisture, and a moisture-free condition is very difficult to maintain. Condensers for high voltage are sometimes made by using glass plates covered with tinfoil or other conducting material and immersed in transformer oil to prevent corona discharges. Another type of condenser is the electrolytic condenser, which has the advantage of giving an extremely large capacity in a moderate space.[1]

Electrolytic condensers consist of a pair of aluminum electrodes placed in a chemical solution. Their dielectric is a very thin film of

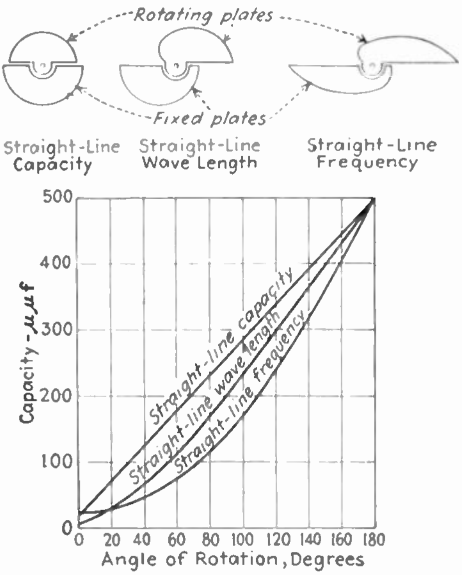


FIG. 13.—Characteristics of straight-line capacity, straight-line wave length, and straight-line frequency condensers, showing capacity as a function of angle of rotation and the approximate shape of plates for each type.

aluminum oxide that forms on the surface of the electrodes and gives a very high capacity because of its extreme thinness while withstanding a potential of several hundred volts before breaking down. The thickness of the film, and hence the capacity of the condenser, depends upon the voltage at which the condenser has been used, but a capacity in the order of 1 μf per square inch of film is typical. Thus a few square feet of electrodes will give a capacity of thousands of microfarads. Electrolytic condensers deteriorate with time as a result of destructive chemical actions but when properly built have a number of years of useful life. The losses in electrolytic condensers are high compared with those

[1] A good discussion of electrolytic condensers is to be found in H. O. Siegmund, The Aluminum Electrolytic Condenser, *Bell System Tech. Jour.*, vol. 8, p. 41, January, 1929.

of other condensers, but for many purposes, such as by-passing alternating currents, this is of secondary importance.

Although paper, mica, and air are the dielectrics generally used in condensers, the dielectric losses in other materials such as bakelite, hard rubber, fiber, porcelain, and various molded insulating compounds are important because these materials are often used in one form or another as insulators. The solid dielectric used in the air condenser to separate the two sets of plates is an excellent example of such use. Many other cases may be cited, such as tube bases, tube sockets, coil forms, panels, and so on. Although the dielectric losses which are present in such cases are generally small because of the small quantity of dielectric involved, they are sometimes very important, as is the case with the low-loss variable condenser used in tuning circuits to resonance, the requirements of which have already been discussed.

Condensers also rated at current and frequency.

Voltage Rating of Condensers.—The voltage that can be safely applied to a condenser depends upon the insulating strength of the dielectric used and upon the electrical losses in the dielectric. If the applied voltages exceed the dielectric strength the dielectric will spark through and the condenser will be destroyed. However, if the losses in the condenser are sufficient to cause a moderate amount of heating, the allowable voltage will be something less than the dielectric strength of the material because of the fact that all dielectrics deteriorate rapidly when heated. As the condenser losses are proportional to frequency the voltage rating of a particular condenser will be highest on direct-current potentials, somewhat less at low frequencies, and increasingly lower as the frequency is increased, until at high radio frequencies the allowable voltage is inversely proportional to the square root of the frequency and is only a small fraction of the insulation strength to direct current. It is therefore very important that condensers which must withstand high radio-frequency voltages have low losses. Otherwise the heat that is generated in the dielectric will raise the temperature to a point where deterioration and eventual destruction will take place.

Mica, because of its very low losses and high insulation strength, is the dielectric universally used for high-voltage service at radio frequencies. It is very important that high-voltage condensers with solid dielectrics be thoroughly impregnated to eliminate corona and the losses resulting from moisture. Transformer oil and some other types of oil can be used as a dielectric for high-voltage service but have the disadvantage of picking up moisture from the atmosphere and of possessing an insulating strength in bulk that is relatively low compared with the value in small quantities. Variable condensers with air dielectric are occasionally used in high-voltage radio-frequency service. Such condensers are extremely bulky in proportion to their capacity because the spacing between plates must be very large if the voltage rating is to be reasonably high. The result is that high-voltage air condensers are used only at high radio frequencies where the capacity is very small. The voltage that may be applied to an air-dielectric condenser is proportional to the air pressure; so enclosing the condenser in a container filled with air at a pressure of from 10 to 20 atmospheres greatly increases the potential that may be safely applied.

Inductance of Condensers.—All condensers have a certain amount of inductance in their circuit because of the fact that the current flowing through the condenser produces magnetic flux. The major part of this inductance occurs in the lead wires of the condenser, but a certain amount results from magnetic flux set up by the current in the condenser plates.

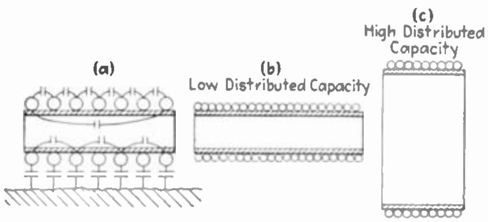


FIG. 14.—Some of the coil capacities that contribute to the distributed capacity of a single-layer coil, and coil shapes having high and low distributed capacity.

The inductance of condensers is very small even under the most unfavorable conditions and is therefore important only when the frequency is extremely high. Condenser inductance can be minimized by properly arranging the lead wires and the connections to the condenser plates.

Distributed Capacity of Coils.—Every electric circuit and every piece of electrical equipment has capacity associated with it because there are always dielectrics separating conductors between which voltage exists. These capacities are very often quite small, but at very high frequencies even a small condenser has a low reactance and so becomes important. The stray capacity of an inductance coil is a good and important example of this.

In an air-core inductance coil there are small capacities between adjacent turns, capacities between turns that are not adjacent, and capacities between terminal leads. In addition there can be capacities to ground from each turn. Some of the different capacities that may exist in a typical coil are shown in Fig. 14*a*. It is to be noted that every turn has a capacity to every other turn and also a capacity to ground. Each of these capacities stores a quantity of electrostatic energy that is determined by the capacity and the fraction of the total coil voltage that appears across the turns involved. The total effect which the numerous small coil capacities have can be represented to a high degree of accuracy by assuming that these many capacities can be replaced by a single condenser of appropriate size shunted across the coil terminals. This equivalent capacity is called either the distributed capacity or the self-capacity of the coil.

The self-capacity of a coil is determined by the spacing between turns, and particularly between those turns near the opposite ends of the winding. Thus in a single layer coil the coil capacity can be kept low by spacing the individual turns as much as is practicable and then using a coil shape that places the ends of the coil some distance apart. In particular a coil shaped as shown in Fig. 14*b* will have much less distributed capacity than when proportioned as in Fig. 14*c* because of the greater distance between terminals in the former case. The material on which the winding is placed has a dielectric constant greater than that of air and so increases the distributed capacity more or less in proportion to the amount of solid dielectric present, which should therefore be kept to the smallest feasible amount.

In multilayer coils the distributed capacity tends to be high because the layer arrangement of the winding causes turns from different parts of the winding to be located near each other. Thus in the two-layer winding shown at Fig. 15*a*, in which the turns are numbered in order, the first and last turns are adjacent, which causes the voltage between these turns to be large and hence greatly increases the effect of the shunting capacity. An improvement over the simple layer winding of Fig. 15*a* is obtained by using the type of winding shown in Fig. 15*b*, known as a bank winding, in which adjacent turns represent parts of the coil that are close together, while the ends of the windings are fairly far apart. Another method of keeping down the self-capacity of a multilayer coil is to use many layers with only a few turns per layer, as in Fig. 15*c*, which has the effect of separating the ends of the coil. Still another device for reducing self-capacity is shown in Fig. 15*d*, in which the layers are spaced to reduce the capacity from turns in one layer to turns in adjacent layers. Many ingenious methods have been devised for incorporating these principles into types of construction having desirable mechanical rigidity, and several types of multilayer coils with low capacity are produced commercially.

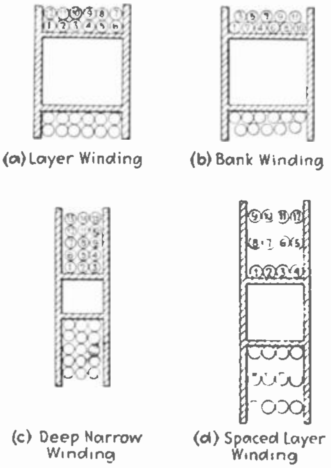


FIG. 15.—Several types of multilayer windings.

The various coil capacities that contribute toward the distributed capacity include as part of their dielectric the wire insulation and the form upon which the coil is wound. Since these materials have appreciable dielectric hysteresis there is a loss associated with the distributed capacity of the coil which is referred to as the dielectric loss of the coil and has the effect of increasing the effective coil resistance. In order to keep the dielectric loss small the coil form should be of material having low dielectric losses and should be no thicker than mechanical considerations make advisable. It is also important that the winding be covered with a moisture-proof binder such as collodion to prevent absorption of moisture by the cotton or silk insulation of the wire. The dielectric losses are low and often negligible in well-constructed coils but may reach an undesirable magnitude if no attention is paid to them.

9. The Effective Resistance of Coils and Conductors at Radio Frequencies.—The effective resistance offered by conductors to radio frequencies is considerably more than the ohmic resistance measured with direct currents. This is because of an action known as skin effect, which causes the current to be concentrated in certain parts of the conductor and leaves the remainder to contribute little or nothing toward carrying the current. As a result of this effect *it is necessary to generalize the concept of resistance when dealing with radio frequencies by considering the resistance to be that quantity which when multiplied by the square of the current will give the energy dissipated in the circuit.*

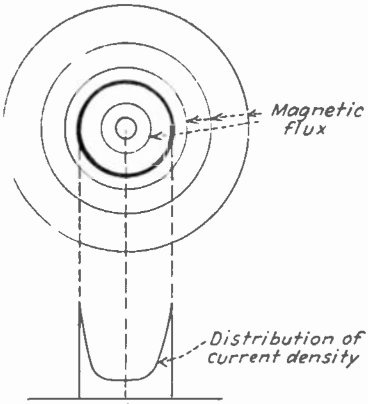


FIG. 16.—Isolated round conductor, showing flux paths and current distribution at radio frequencies. Note that the current density is highest for parts of the conductor encircled by the fewest flux lines.

Skin Effect in an Isolated Conductor.—A simple example of skin effect, and one which makes its nature clear, is furnished by an isolated round wire. When a current is flowing through such a conductor the magnetic flux that results is in the form of concentric circles as shown in Fig. 16. It is to be noted that some of this flux exists within the conductor and therefore links with, *i.e.*, encircles, current near the center of the conductor while not linking with current flowing near the surface. The result is that the inductance of the central part of the conductor is greater than the inductance of the part of the conductor near the surface because of the greater number of flux linkages existing in the central region. At radio frequencies the reactance of this extra inductance is sufficiently great seriously to affect the flow of current, most of which flows along the surface of the conductor where the impedance is low rather than near the center where the impedance is high. The center part of the conductor therefore does not carry its share of the current and the effective resistance is increased, since in effect the useful cross section of the wire is very greatly reduced. The actual type of current distribution obtained in the case of a round wire is as shown in Fig. 16.

The ratio which the effective alternating-current resistance bears to the direct-current resistance of a conductor increases with frequency, with conductivity of the conductor material, and with the size of conductor. This results from the fact that a higher frequency causes the extra inductance at the center of the conductor to have a higher reactance, while a greater conductivity makes the reactance of the extra inductance of more importance in determining the distribution of current, and a greater cross section provides a larger central region. It is to be noted, however, that a larger conductor always has less radio-frequency resistance than a smaller one because although the ratio of alternating-current to direct-current resistance is less favorable this is more than made up by the greater amount of conductor cross section present. The actual value of the ratio of alternating-current to direct-current resistance for a solid round conductor can be calculated with the aid of tables and formulas found in every electrical handbook.

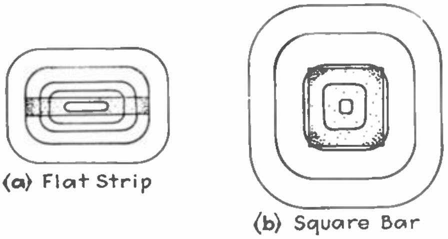


FIG. 17.—Isolated strip and bar conductors, showing approximate flux paths and current distribution. The current density is indicated by the density of shading and is seen to be greatest for those parts of the conductor encircled by the fewest flux lines.

When skin effect is present the current is redistributed over the conductor cross section in such a way as to make most of the current flow where it is encircled by the smallest number of flux lines. This general principle controls the distribution of current irrespective of the shape of the conductor involved. Thus with a flat-strip conductor, such as shown in Fig. 17, the current flows primarily along the edges, where it is surrounded by the smallest amount of flux, and the effective resistance will be high because most of the strip carries very little current. This illustration makes clear that it is not the amount of conductor surface that determines the resistance to alternating current but rather the way in which the conductor material is arranged. Another example is formed by the square-bar conductor of Fig. 17, in which the current is concentrated along the four corners as indicated in the figure because the flux linkages are least for this portion of the cross section.

Where it is important that an isolated conductor have very low resistance to radio-frequency currents it is preferable to use thin tubular conductors instead of round wire of the same cross section, since the central part of the wire carries very little current while all of the material of a tube is effective. Another way of making more effective use of the conducting material is to form the conductor from a large number of small enameled wires connected in parallel at their ends but insulated from each other throughout the rest of their length and thoroughly interwoven. If the stranding is properly done each conductor will, on the average, link with the same number of flux lines as every other conductor, and the current will divide evenly between the strands. If each strand is small it will have relatively little skin effect over its cross section with the result that all of the material is effective in carrying the current, and a radio-frequency resistance approximating the direct-current resistance results. A stranded cable of this type is called a litz (or *litzendraht*) conductor.

Skin Effect in Coils.—The same principle that governs the current distribution in an isolated conductor also determines the distribution of current in the conductors of a coil; that is, the current density is greatest in those parts of each coil conductor encircled by the smallest number of flux lines. The skin effect in coils is however much more complicated and much greater in magnitude than in isolated conductors because each turn of the coil produces flux that causes skin effect in adjacent turns. As a result the radio-frequency resistance of coils may be as much as several hundred times the resistance to direct currents. The approximate current distribution in the conductors of a typical radio coil is indicated by the shading in Fig. 6, which also shows the flux paths and so brings out the relation between current density and flux linkages.

The losses in a coil are most conveniently expressed in terms of the ratio of the coil reactance ωL to the effective coil resistance R. This ratio approximates the reciprocal of the coil power factor and is so important in the theory of resonant circuits that it is considered as a fundamental coil property and is usually referred to by the symbol Q; that is

$$Q = \frac{\text{coil reactance}}{\text{coil resistance}} = \frac{\omega L}{R} \tag{20}$$

The effective coil resistance R includes any dielectric loss which the coil may have, but in a well-constructed coil the effective coil resistance R is caused almost solely by skin effect.

It is convenient to express the characteristics of a coil in terms of the ratio of coil reactance to effective resistance because this ratio Q is approximately constant for the same coil over a wide range of frequencies as a result of the fact that the effective radio-frequency resistance of a coil is roughly proportional to frequency. Furthermore the value of Q for equally well designed coils intended for use at different frequencies is approximately the same; that is, a value of Q which is considered as denoting an efficient coil of a size suitable for use at 100 kc also represents the Q of an efficient coil for 1000-kc service. The values of Q actually obtained with coils used in receiving equipment range from fifty to several hundred, with values somewhat greater than these being frequently encountered in transmitter inductances. In general the value of Q that can be obtained with reasonable effort tends to become somewhat less at higher frequencies, so that while a Q of 100 represents only a moderately good coil at 100 kc, it denotes an especially well designed coil at a frequency of 10,000 kc.

Factors Influencing the Q of a Coil.—The actual value of $\omega L/R$ depends primarily on the coil construction and is determined by complicated factors for which a complete mathematical solution has never been made.

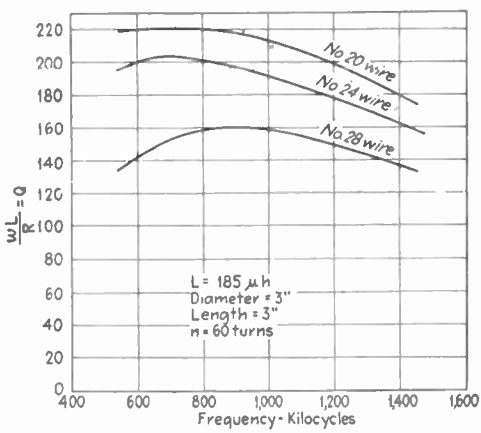


FIG. 18.—Curves showing values of Q as a function of frequency for three coils differing only in wire size. The largest size wire has over six times the cross section of the smallest, and yet the latter has a value of Q only about 35 per cent less.

In spite of this there are a number of general principles that can be used as a guide in the design of coils. To begin with, if coils differing only in wire diameter are compared, it will be found that for each frequency there will be a particular size of wire that will give the lowest radio-frequency resistance and hence the highest Q. This best size of wire will be different for different frequencies and is often, although not necessarily, the largest wire that can be wound in the space available. As long as the wire is not too small, changes in wire size have surprisingly little effect on coil alternating-current resistance because the increased skin effect of the larger conductors offsets in large measure the greater cross section. It is only when the frequency is so low or the wire so small as to give little skin effect that the radio-frequency resistance of the coil becomes markedly influenced by the direct-current resistance. These effects of wire size are graphically shown by the curves of Fig. 18, which give Q as a function of frequency for several coils differing only in wire size, and in which increasing the direct-current resistance over six times reduces the coil Q by about 35 per cent.

Increasing the size of a coil while maintaining the inductance, proportions, and direct-current resistance constant tends to reduce the effective radio-frequency resistance and hence give a larger Q. Furthermore the larger coil provides a winding space for larger wire than the smaller one, which if utilized will still further increase the Q as the dimensions are increased. A large coil can utilize a large conductor to better advantage than a small coil because the increased space over which the flux is distributed reduces the flux density in the vicinity of the conductors

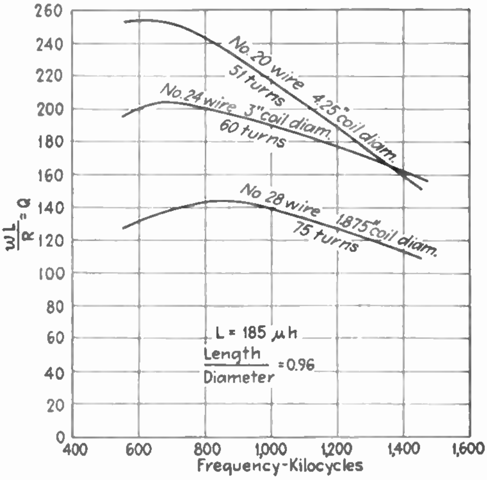


FIG. 19.—Curves showing values of Q as a function of frequency for three coils having the same inductance and same ratio of length to diameter, but differing in size. Increasing the size of the coil increases the Q at the frequency where the Q is highest and tends to lower the frequency at which the coil is most efficient.

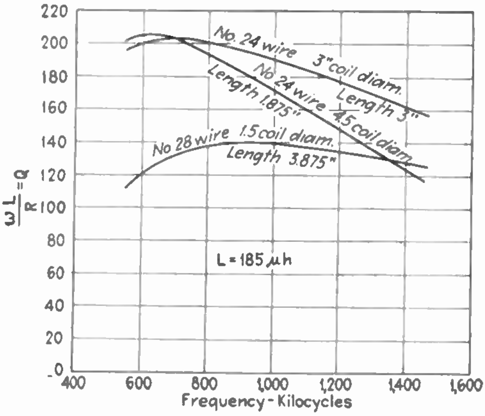


FIG. 20.—Curves giving values of Q as a function of frequency for three coils having the same inductance but differing in ratio of length to diameter. These show how there is a best shape.

and hence reduces the skin effect to reasonable values even with large wire. The effect which changes in size can have on the Q of a coil is demonstrated by the curves of Fig. 19.

The best shape for a coil having a given inductance is neither a very long coil with a small diameter nor a short coil with a large diameter but rather one of intermediate proportions. The curves of Fig. 20 are typical examples of the differences that can be expected from different shapes.

The types of conductors most frequently used in winding coils are round wire, litz, tubing, and flat- or edgewise-wound strip. Ordinary wire is one of the best shapes and is the standard with coils intended for use in radio receivers. Coils wound with litz usually have somewhat less radio-frequency resistance than when a solid wire of the same cross section is used, but in receiving coils the gain obtained from litz is generally not considered as being worth the additional expense. The advantages to be gained by using litz in coils are greatest at frequencies below 2000 kc, since at very high frequencies irregularities in stranding introduce effects which tend to compensate for the otherwise favorable properties of litz. A comparison of the Q of two coils, one of which is wound with litz and the other with solid wire of the same cross section, is shown in Fig. 21.

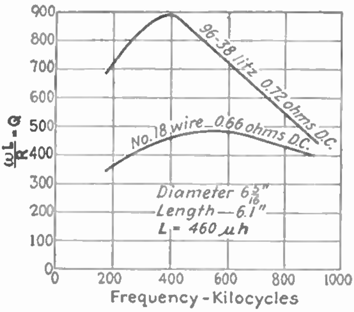


Fig. 21.—Curves giving values of Q as a function of frequency for two coils identical in every way except that one is wound with solid wire and the other with litz of approximately the same cross section. The litz coil has an exceptionally high Q.

The principal use of litz is in the large coils of low-frequency radio transmitters where very heavy currents must be handled and where the inductance is large. For such purposes litz wire has no competitor. Transmitter coils for high-frequency transmitters, or for low power service at low frequencies, are usually made of copper tubing or strip. Tubing is becoming more and more the standard for such service because it has desirable mechanical rigidity combined with a better current distribution over its cross section than does either flat- or edgewise-wound strip and hence has a better Q in proportion to the amount of conductor material employed. The reasons for the inferiority of strip are the same with coils as in connection with isolated conductors and are evident when the flux paths of coils are considered in relation to the conductor cross section.

Large quantities of data giving coil resistance in countless specific cases are to be found scattered throughout the literature, but no attempt has been made here to summarize the published results because they represent a huge mass of figures which have as yet defied organization in any systematic manner.[1] The important things to keep in mind are the general principles discussed above and the fact that the effective resistance of a coil to alternating currents is determined primarily by the flux distribution in the conductors and has very little relation to the direct-current resistance of the wire.

10. Types of Coils Used in Radio Work. *Coils for Radio Receivers.*—The vast majority of coils for radio-frequency service is used in resonant

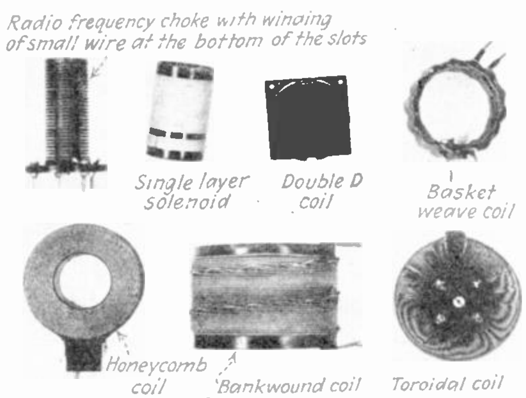


FIG. 22.—Representative coils of different types designed for use in radio receivers.

circuits of radio receivers. Coils for this purpose must have low losses (*i.e.*, a high Q), moderate size, and reasonably small distributed capacity. The type of coil most frequently employed to meet these requirements is the single-layer solenoid, which when properly constructed has very low losses in proportion to its inductance, and also small distributed capacity. The principal limitation to the use of the single-layer solenoid is in the space required. Where a large inductance is necessary the single-layer solenoid must be either so large as to reach unmanageable proportions for a radio receiver or must be wound with wire that is so small as to give a low Q.

In addition to the solenoid a number of other types of coils are sometimes used. Most common of these is the basket-weave type of coil shown in Fig. 22, which is used extensively in coils intended for service at very high frequencies. This type of coil is wound over a series of pins mounted in holes located in the circumference of a circle and, after wind-

[1] Useful information of this sort is to be found in the following publications: August Hund and H. B. De Groot, Radio-frequency Resistance and Inductance of Coils Used in Broadcast Reception, *Bur. Standards Tech. Paper*, 298; E. L. Hall, Resistance of Conductors of Various Types and Sizes as Windings of Single-layer Coils at 150 to 6000 Kilocycles, *Bur. Standards Tech. Paper* 330; J. H. Morecroft, "Principles of Radio Communication."

ing, is sewed together with thread so that when reasonably stiff wire is used the coil maintains its shape without the use of a dielectric form and thus has negligible dielectric losses. The size of basket-weave coils that are practicable is limited by mechanical considerations to relatively small inductances.

Where more inductance is required than can be obtained in a single-layer solenoid of reasonable size with wire that is not excessively small it is necessary to resort to some type of multilayer coil. Such coils are characterized by a relatively high inductance in proportion to the dimensions of the coil and usually have losses and distributed capacity somewhat in excess of the values that would be obtained for a single-layer coil of the same inductance. Various types of multilayer coils are used, such as the bank winding and the spaced-layer arrangements discussed in Sec. 8 and illustrated in Figs. 15 and 22. One of the most ingenious of the multilayer coils is the honeycomb type shown in Fig. 22, which has a few relatively widely spaced turns per layer wound diagonally in such a way that there are several layers separating the nearest parallel strands. The coil is wound on suitably arranged pins, which are later removed, and is held together and made self-supporting by impregnation with a binder.

It is frequently important that the major portion of the magnetic field produced by an inductance be confined to a small space in order to prevent mutual inductance with neighboring coils and circuits. With the single-layer solenoid this restriction limits the size of coil that may be used because the volume covered by the magnetic flux is proportional to the size of the solenoid. Considerations of this sort frequently make very small coils preferable to larger coils that have a lower loss but which develop greater mutual inductance with adjacent circuits. Various efforts have been made to devise types of coils that have small or negligible external magnetic fields. Toroidal coils, illustrated in Fig. 22, have substantially no external magnetic field but have not met with much favor for radio use because of the fact that the amount of wire that must be used in a toroid is so much greater than that required to give the same inductance with the single-layer coil that the radio-frequency resistance of the toroidal coil is high unless the coil dimensions are very great. Various types of astatic coil windings have also been used to some extent. One of these is shown in Fig. 22 and consists of two D-shaped coils side by side and so connected that the magnetic flux tends to pass down through one cylinder and up through the other one. The field at some distance away is relatively small because at such locations the effects of the two coils are in substantial opposition. Such a coil might be considered as a modified toroidal coil and has characteristics in regard to losses and distributed capacity intermediate between those of the toroid and the single-layer solenoid.

Coils for Use in Transmitters.—Coils used in radio transmitters must meet requirements that differ somewhat from those of receivers because of the heavy currents and high voltages that are present in high-power transmitters and which heat the conductors and insulation as a result of conductor resistance and dielectric hysteresis. The conductors used in such coils are most frequently made of heavy round rod or of tubing,

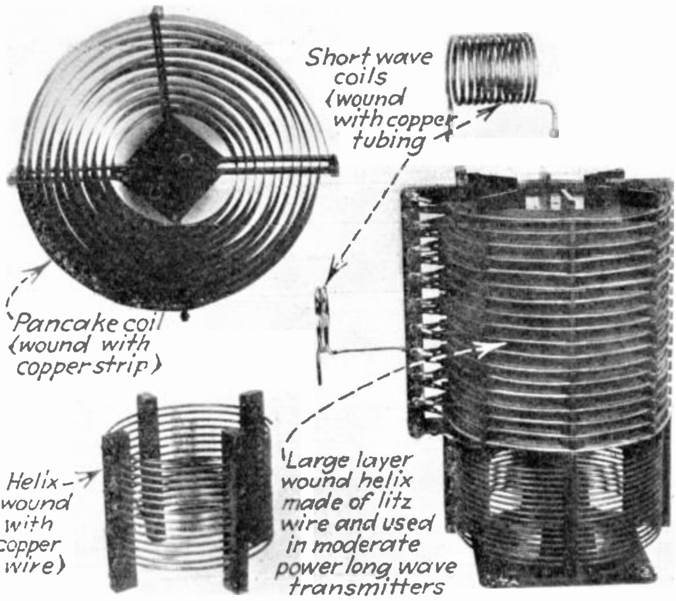
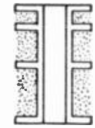


FIG. 23.—Representative coils of different types designed for use in radio transmitters.

which is bent into position and held there by strips of dielectric. Litz conductors are generally employed in coils that are required to carry very large currents of low radio frequencies while copper strip is also used at times. A number of types of coils built for use in radio transmitters of moderate power are shown in Fig. 23, which gives an idea of the methods of construction commonly employed. There are numerous design problems encountered in coils that are to handle large radio currents. In particular the high voltages developed across such coils have a tendency to produce corona at the points where the electrostatic stress is highest. Since even the slightest trace of corona at radio frequencies results in high energy loss, great care must be used in arranging the details of the construction.[1]

[1] An excellent discussion of the problems involved in the design of large inductances for use in high-power transmitting stations is to be found in W. W. Brown and J. E. Love, Design and Efficiencies of Large Air Core Inductances, *Proc. I.R.E.*, vol. 13, p. 755, December, 1925.

Radio-frequency Choke Coils.—Certain types of radio circuits call for an inductance coil having a very high impedance to radio-frequency currents lying within a certain range of frequencies. The losses of such coils are quite unimportant, but the impedance when taking into account the distributed capacity as well as the inductance is required to be very high. Coils of this type are called radio-frequency choke coils and must have extremely low distributed capacity. The two methods commonly used in constructing such coils are shown at *a* and *b* of Fig. 24. In the first method the coil is very long in proportion to its diameter and has a large number of turns of small wire. The great length insures a small distributed capacity, and the many turns of small wire give a large inductance. In the second method the coil is wound in a series of spaced sections as shown in the figure. The over-all distributed capacity of the coil is kept small by making the multilayer winding in each section deep and narrow and by employing a number of sections in series. The resistance of radio-frequency coils is unimportant, and the size of wire is selected from the point of view of current-carrying capacity.


(b) Sectionalized Coil

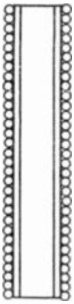

(a) Long Narrow Solenoid

FIG. 24.—Types of radio-frequency choke coils wound to give a high inductance with small distributed capacity.

Variable Inductances.—Inductances that are continuously variable can be constructed by connecting two coils in series and then varying the total inductance $L_1 + L_2 + 2M$ of the combination by changing the mutual inductance. Such variable inductances are called variometers and have many uses. When the two coils both have an inductance of L, and the maximum coefficient of coupling that can be obtained is k, the inductance can be varied from $2L(1 - k)$ to $2L(1 + k)$. The construction of a typical variometer is shown in Fig. 25, where the two coils can be adjusted from series aiding to series opposing by rotating the inner coil, and the maximum coefficient of coupling is made high by arranging the two inductances so that they coincide as nearly as is physically possible.

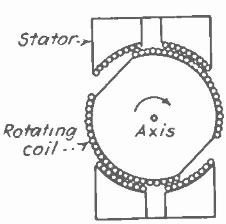


FIG. 25.—Constructional details of a typical variable inductance (variometer).

11. Electrostatic and Electromagnetic Shielding of Coils.—Under many conditions it is necessary to confine substantially all electrostatic and electromagnetic flux to a limited space in the neighborhood of the coil. This result can be accomplished by completely enclosing the coil in a container made of material having low electrical resistivity, such as copper, aluminum, or brass, or by a conductor made of magnetic material. Such an arrangement acts as an electrostatic shield because it is a conductor and so constitutes a Faraday cage that screens the space external to the cage from electrostatic effects within. As far as electrostatic shielding is concerned the exact nature of the shielding material is not highly important, and the shielding is substantially perfect if the container in which the coil is located is water-tight or if its joints are lapped.

When the magnetic flux that is to be shielded is of audio or low radio frequency it is preferable to make the coil shield of magnetic material of high permeability. Such material acts as a magnetic short circuit that prevents the magnetic flux lines from extending to the space outside the container in which the coil is located and thus gives a shielding effect with magnetic flux that is analogous in all respects to the effect which a Faraday cage has on electrostatic flux. In order to obtain complete shielding it is necessary that the magnetic material of the shield be sufficiently thick to give a low reluctance to the flux entering the shield at one point and departing from it at another, so that there will be negligible magnetomotive force developed between different points on the shield. Otherwise some magnetic flux will not be intercepted by the shield.

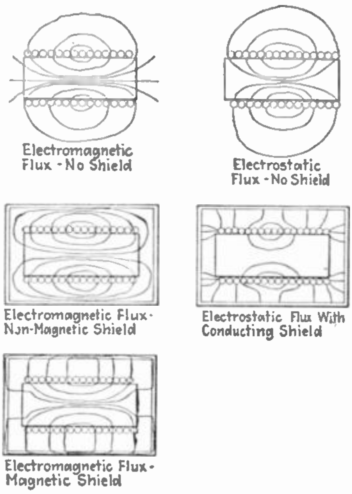


Fig. 26.—Paths of electrostatic and electromagnetic flux lines about the same coil with and without magnetic and non-magnetic shields.

Where the frequency of the magnetic field is high more satisfactory shielding can be obtained by the use of shields having high electrical conductivity. The magnetic flux in attempting to pass through such a shield induces voltages that set up eddy currents which in large measure prevent the magnetic flux from penetrating through the shield. The shielding effect produced by eddy currents increases with frequency and with the conductivity of the shielding material. This is because the voltage induced by a flux is proportional to the frequency and because the eddy currents produced by a given induced voltage are proportional to the conductivity of the shield. At the high radio frequencies the result is that non-magnetic shields of low electrical resistance, such as copper and brass, are better than high-resistance magnetic materials, such as iron. The effect which a non-magnetic shield has on the electrostatic and electromagnetic flux of a coil is shown in Fig. 26.

It is very difficult to shield a magnetic field completely, and if it is essential that the shielding be nearly perfect several concentric shields separated by appreciable air gaps must be used. The shields must also be electrically continuous, without joints to disturb the eddy current and flux paths. Thus when the shield is composed of aluminum or copper, a high degree of shielding can be obtained only when as many of the joints as possible are soldered, and when the remaining ones are carefully fitted so as to give a continuous contact having low electrical resistance. Otherwise the eddy currents will be diminished by the joint resistance, and the shielding will be incomplete.

Effect of Shielding on Coil Properties.—Placing a coil within a shield increases the coil's distributed capacity and the effective resistance, while the coil inductance is reduced with non-magnetic shields and increased when the shield is magnetic. The distributed capacity is increased as a result of the capacity between various parts of the coil and the shield. Restricting the magnetic-flux lines to the space within the shield by the use of non-magnetic material of high electrical conductivity increases the reluctance of the magnetic circuit (*i.e.*, decreases the amount of flux produced by a given coil current) and hence reduces the effective inductance of the coil by an amount that depends on the extent to which the shield interferes with the normal flux paths. When the shield is so large in comparison with the coil size as to interfere with only a small portion of the flux paths, the effect on the inductance will be small, whereas a close-fitting shield interferes with many flux lines and has the effect of greatly lowering the effective coil inductance. If the shield is of magnetic material the effect on inductance is just the opposite, since the magnetic material supplies a low-reluctance path to the magnetic flux that increases the flux (and hence the inductance) the closer the shield is to the coil.

The energy consumed by the eddy currents flowing in the shield, and by the iron losses if a magnetic shield is used, must be supplied by the coil and hence have the effect of increasing the effective coil resistance by an amount that depends primarily upon the extent to which the shield interferes with the normal flux paths. If high-conductivity shielding material is used and the shield is not too close to the coil the added losses will not be excessive. It is therefore possible to construct a shielded coil that will have a reasonably high Q even though the losses will not be as low as those of the same coil with the shield removed. An analysis of the effect which a non-magnetic shield has on effective coil inductance and resistance can be made on the basis of coupled circuits and is taken up in Sec. 18.

Electrostatic Shielding without Magnetic Shielding.—Electrostatic flux can be shielded without affecting magnetic fields by surrounding the space to be shielded with a conducting cage that is made in such a way as to provide no low-resistance path for the flow of eddy currents, at the same time offering a metallic terminal upon which electrostatic flux can end. One way of accomplishing this is to use a shield formed from a wire having one terminal grounded and the other free. Such a cage acts as a fairly good electrostatic shield but has very little influence on the magnetic field since there are no closed loops around which eddy currents can circulate. Another type of electrostatic shield that can be employed under some circumstances consists of metal foil so arranged that its surface is approximately parallel with the magnetic field; with an insulated gap provided where necessary to prevent the shield from becoming a short-circuited turn. Thus the electrostatic field of a coil can be shielded by wrapping conducting foil over the wire in the same direction that one would wrap the coil with tape, and if an insulated gap is left along a line that is parallel to the direction in which the flux lines go, the effect on the magnetic flux will be small.

12. Radio-frequency Coils with Magnetic Cores.—The use of magnetic cores in coils intended for radio-frequency service is limited by the fact that the eddy-current losses in magnetic material are proportional to the square of the frequency and so become excessively great when the frequency is high. The large eddy currents produced in the core also set up flux of their own which is in opposition to the main flux and therefore cause the effective inductance of the coil to become lower as the frequency is increased. This phenomena is known as magnetic skin effect, since it is equivalent to preventing the magnetic flux from penetrating very deeply into the magnetic material of the core and so is analogous to skin effect in conductors.

The eddy currents can be reduced by subdividing the magnetic core material and by the use of magnetic material of high electrical resistivity. This can be accomplished in a fairly satisfactory manner at the lowest radio frequencies by the use of extremely thin laminations of magnetic materials such as silicon steel. A still better method of dividing the core consists of reducing the core material to a fine dust, mixing this powder with an insulating compound that will surround the particles, and then forming the mixture under high pressure into solid cores, preferably of toroidal shape. This construction produces a core that is subdivided to a much greater extent than can be realized with laminations, and at the same time there are distributed throughout its length air gaps that prevent the core from saturating or changing its inductance with moderate flux densities.[1]

When properly constructed, coils with magnetic cores can be made to have a fairly high value of Q for radio frequencies up to about 100 kc. The losses and the magnetic skin effect become of increasing importance

[1] Dust cores are extensively used in telephone communication at frequencies below 50,000 cycles. A description of the method of preparation of cores made of permalloy dust and iron dust and their outstanding mechanical and electrical properties is to be found in the following articles: Buckner Speed, and G. W. Elmen, Magnetic Properties of Compressed Powdered Iron, *Jour. A.I.E.E.*, vol. 40, p. 596, July, 1921; W. J. Shackelton and I. G. Barber, Compressed Powdered Permalloy, *Trans. A.I.E.E.*, vol. 47, p. 429, April, 1928.

as the frequency is raised, with the result that magnetic cores are seldom used for frequencies above 100 kc and only occasionally at the lower radio frequencies where their characteristics are more favorable.

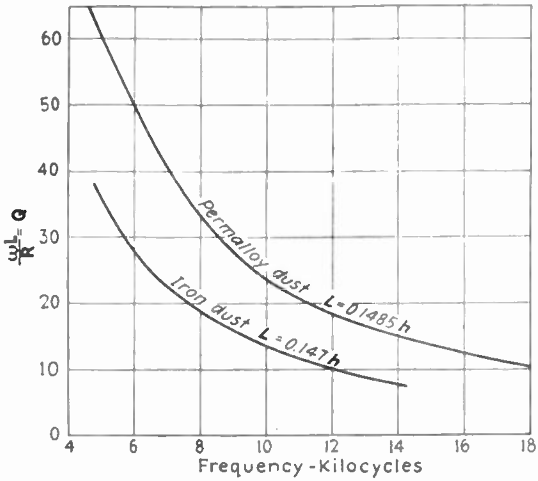


FIG. 27.—Curves showing values of Q as a function of frequency for a coil having a permalloy dust core and a coil having an iron dust core. These cores are intended for service at audio and low radio frequencies, and as the frequency is raised, their Q becomes less because of eddy-current losses in the core.

The Q of an iron dust and of a permalloy dust core are given in Fig. 27 as a function of frequency. These coils were both wound with relatively small wire because the copper loss is even then small compared with the iron loss.

CHAPTER III

PROPERTIES OF RESONANT CIRCUITS

13. Series Resonance.—When a constant voltage of varying frequency is applied to a circuit consisting of an inductance, capacity, and resistance, all in series, the current that flows depends upon frequency in the manner shown in Fig. 28. At low frequencies the capacitive reactance of the circuit is large and the inductive reactance is small, so that most of the voltage drop is across the condenser, while the current is small and leads the applied voltage by nearly 90°. At high frequencies the inductive reactance is large and the capacitive reactance low, resulting in a small current that lags nearly 90° behind the applied voltage, and most of the voltage drop is across the inductance. In between these two extremes there is a frequency, called the resonant frequency, at which the capacitive and inductive reactances are exactly equal and consequently neutralize each other, leaving only the resistance of the circuit to oppose the flow of current. The current at the resonant frequency is accordingly equal to the applied voltage divided by the circuit resistance, and is very large if the resistance is low.

The characteristics of a series resonant circuit depend primarily upon the ratio of inductive reactance ωL to circuit resistance R, that is, upon $\omega L/R$. This ratio is frequently denoted by the symbol Q and is called the circuit Q. In the usual resonant circuit the radio-frequency resistance of the circuit is made up almost solely of coil resistance because the losses in a properly constructed condenser are negligible in comparison with those of the coil. The result is that the circuit Q can ordinarily be taken as the Q of the coil alone, which was discussed in Sec. 9.

The general effect of different circuit resistances, *i.e.*, different values of Q, is shown in Fig. 28. It is seen that when the frequency differs from the resonant frequency by as much as several per cent the actual current is very nearly the current that would be obtained with no losses, and is practically independent of the circuit resistance. On the other hand the current at the resonant frequency is determined solely by the resistance. The effect of increasing the resistance of a series circuit is accordingly to flatten the resonance curve by reducing the current at resonance. This broadens the top of the curve, giving a more nearly uniform current over a band of frequencies near the resonant point, but does so by reducing the selectivity of the tuned circuit (*i.e.*, the ability to discriminate between voltages of different frequencies).

Analysis of Series Resonant Circuit.—In deriving the relations that exist in a series resonant circuit the following symbols will be used.

E = voltage applied to circuit
I = current flowing in circuit
f = frequency in cycles

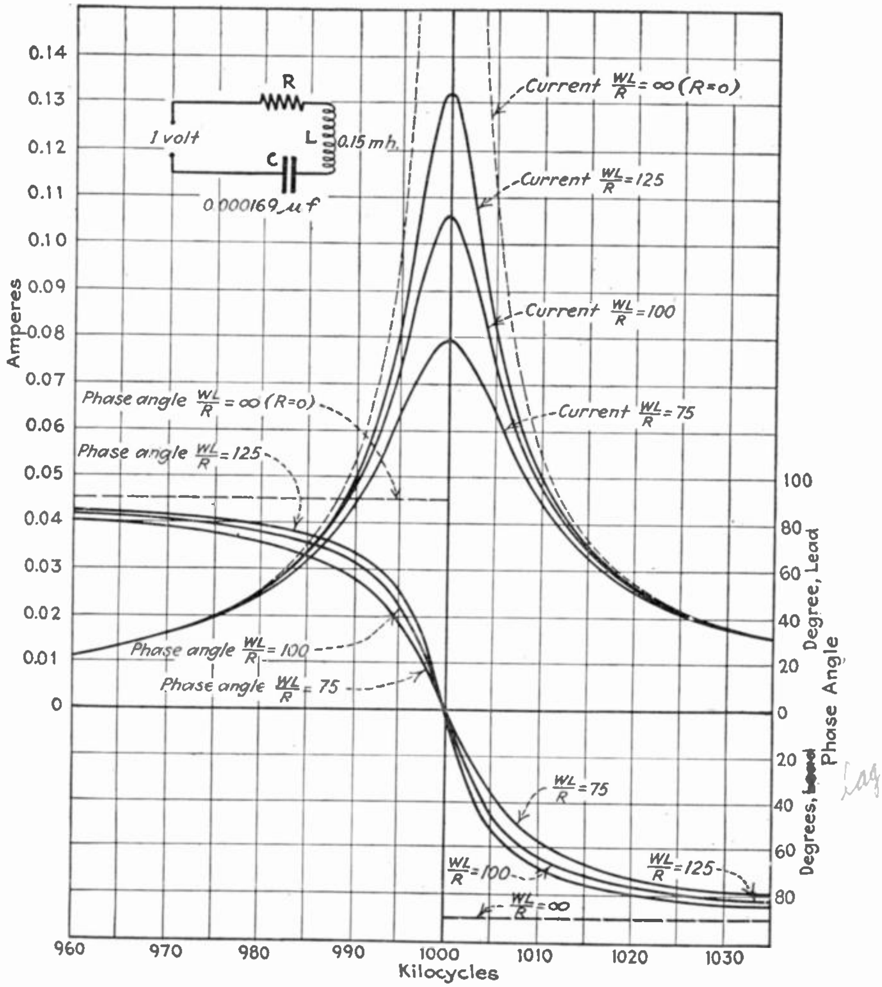


FIG. 28.—Magnitude and phase angle of current in series resonant circuit as a function of frequency for a constant applied voltage. Curves are given for three values of circuit resistance, and also for zero resistance.

$\omega = 2\pi f$
$Q = \omega L/R$
R = effective resistance of tuned circuit
L = inductance in henrys
C = capacity in farads

Z = impedance of series circuit
θ = phase angle of impedance
Subscript $_o$ denotes values at resonant frequency.

The resonant frequency is the frequency at which the inductive and capacitive reactances are equal, so is given by the expression

$$\omega L = \frac{1}{\omega C} \tag{21a}$$

or

$$\text{Resonant frequency} = f_o = \frac{1}{2\pi\sqrt{LC}} \tag{21b}$$

Equation (21*b*) shows that the resonant frequency depends only upon the product LC.

The impedance of the circuit is

$$Z = R + j\left(\omega L - \frac{1}{\omega C}\right) \tag{22}$$

and accordingly has a magnitude of

$$|Z| = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2} \tag{23a}$$

and a phase angle θ given by

$$\tan \theta = \frac{\left(\omega L - \frac{1}{\omega C}\right)}{R} \tag{23b}$$

The current flowing in the circuit is

$$I = \frac{E}{Z} = \frac{E}{R + j\left(\omega L - \frac{1}{\omega C}\right)} \tag{24}$$

At resonance $Z = R$, giving a current in phase with the applied voltage and having a magnitude

$$\text{Current at resonance} = I_o = \frac{E}{R} \tag{24a}$$

This is the maximum current that can flow in the circuit.

In the series resonant circuit the voltages across the condenser and the inductance will both be very much greater than the applied voltage when the frequency is in the vicinity of resonance. This is possible because the voltages across the condenser and inductance are nearly 180° out of phase with each other and so add up to a value that is much smaller than either voltage alone. Since the current at resonance is E/R, the voltage across the inductance at resonance is ωL times the current, or

$$\text{Voltage across } L \text{ at resonance} = E\frac{\omega L}{R} = EQ \tag{25}$$

The voltage across the condenser also has this same value since, at resonance, $\omega L = 1/\omega C$. Equation (25) shows that *at resonance the voltage across the inductance (or condenser) is Q times the applied voltage* (i.e., *there is a resonant rise of voltage in the circuit amounting to Q times*). Since Q can be expected to have a value in the neighborhood of 100, a series resonant circuit will develop a high voltage even with small applied potentials. Thus, if the applied voltage is 50, and $Q = 100$, the potential developed across the condenser is 5000 volts, and unless the condenser is built for high-voltage service the insulation will break down.

Universal Resonance Curve.—Equations (24) and (24*a*) can also be rearranged to express the ratio of current actually flowing to the current at resonance, in terms of the circuit Q and the fractional deviation of the frequency from resonance. When this is done the result is[1]

$$\frac{\text{Actual current}}{\text{Current at resonance}} = \frac{1}{1 + \delta + jQ\delta\left(\dfrac{2 + \delta}{1 + \delta}\right)} \tag{26}$$

where

$$Q = \omega_o L/R_o = \text{circuit } Q \text{ at resonant frequency}$$

$$\delta = \frac{(\text{actual frequency}) - (\text{resonant frequency})}{(\text{resonant frequency})}$$

The quantity δ represents the ratio of the number of cycles by which the actual frequency is off resonance to the resonant frequency and can be called the fractional deviation of the frequency from resonance. Thus a value $\delta = 0.01$ means that the actual frequency differs from the resonant frequency by 0.01 of the resonant frequency. The value of δ is positive or negative as the actual frequency is greater than or less than resonance

[1] The derivation of this equation follows:
The ratio of Eq. (24) to (24*a*) gives:

$$\frac{\text{Actual current}}{\text{Current at resonance}} = \frac{R_o}{R + j\left(\omega L - \dfrac{1}{\omega C}\right)}$$

$$= \frac{R_o}{R + j\left(\dfrac{\omega^2 LC - 1}{\omega C}\right)}$$

By the definition of δ,

$$\omega = \omega_o(1 + \delta)$$

Substituting this value of ω, and remembering that $\omega_o L = 1/\omega_o C$, gives

$$\frac{\text{Actual current}}{\text{Current at resonance}} = \frac{R_o}{R + j\left[\dfrac{(1 + \delta)^2 - 1}{1 + \delta}\right]\omega_o L}$$

$$= \frac{1}{\dfrac{R}{R_o} + jQ\delta\left(\dfrac{2 + \delta}{1 + \delta}\right)}$$

When Q is constant the radio-frequency resistance is proportional to frequency so that $R/R_o = \omega/\omega_o = (1 + \delta)$, which when substituted yields Eq. (26).

respectively. By expressing δ in terms of $1/Q$ it is possible to plot the *universal resonance curves of Fig. 29, from which the exact resonance curve of any series circuit may be obtained without calculation when Q is known.* These curves are extremely useful because they are independent of the

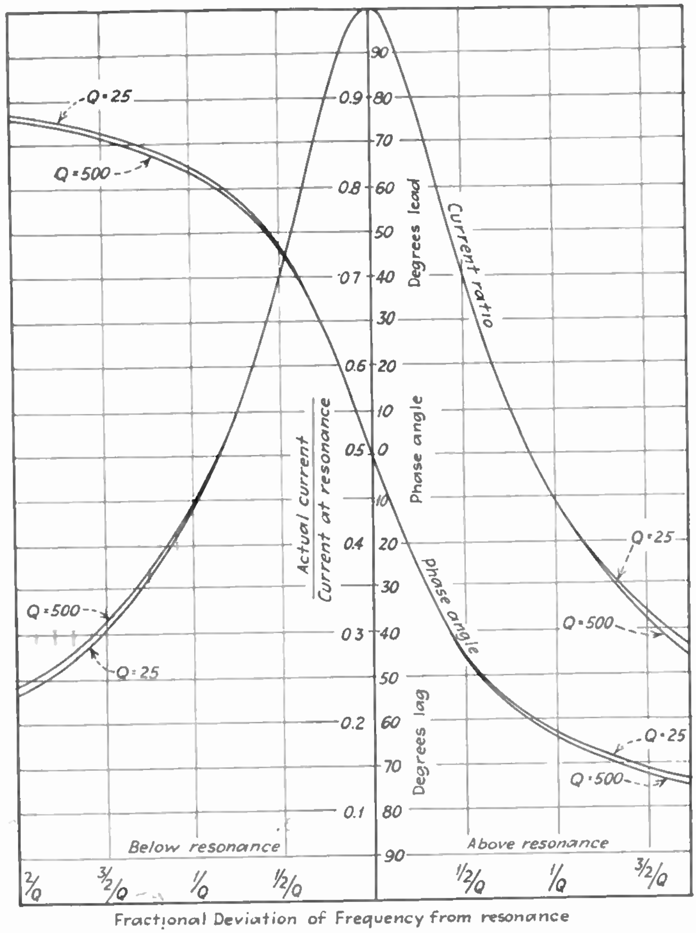


FIG. 29.—Universal resonance curve from which the exact ratio of actual current to current at resonance, as well as exact phase angle, can be determined for any series circuit in terms of the fractional deviation of the actual frequency from resonance. This curve can also be applied to the parallel resonant circuit by considering the vertical scale to represent the ratio of actual parallel impedance to the parallel impedance at resonance. When applied to parallel circuits angles shown in the figure as leading, are lagging, and *vice versa.*

resonant frequency of the circuit and of the ratio of inductance to capacity. It is to be noted that they are practically symmetrical about the resonant frequency and are substantially independent of Q.

The method of using Fig. 29 in practical calculations can best be seen from an example. Assume that a circuit with $Q = 125$ is resonant at 1000 kc, and it is desired to know how many cycles the frequency must be above resonance to reduce the current to one-half of the value at resonance. By referring to Fig. 29 the response above resonance is seen to be reduced to 50 per cent when the fractional deviation from resonance is $0.86/Q$, which represents 0.86/125 of 1000 kc or 6.88 kc. The phase angle of the current as obtained from the curve is 60° lagging.

The only assumption involved in Eq. (26) is that Q is the same at the frequency being considered as at the resonant frequency. *When this is true, Eq.* (26) *and the universal resonance curve involve no approximations whatsoever.* Over the limited range of frequencies near resonance, where the universal resonance curve finds its chief usefulness, the variations in Q are so small as to introduce negligible (*i.e.*, less than 1 per cent) errors from the use of the curve provided the value of Q existing at resonance is used in determining the fractional deviation from resonance.

Working Rules for Estimating Sharpness of Resonance.—Since the curves for different values of Q are almost identical in Fig. 29, particularly in the neighborhood of the resonant frequency, it is possible to state several easily remembered working rules that will enable one to estimate the sharpness of any resonance curve with an error of less than 1 per cent when only the Q of the circuit is known.[1] These rules follow:

Rule 1. *When the frequency of the applied voltage deviates from the resonant frequency by an amount that is* $1/2Q$ *of the resonant frequency the current that flows is reduced to* 70 *per cent of the resonant current, and the current is* 45° *out of phase with the applied voltage.*

Rule 2. *When the frequency of the applied voltage deviates from the resonant frequency by an amount that is* $1/Q$ *of the resonant frequency the current that flows is reduced to* 45 *per cent of the resonant current, and the current is* $63\frac{1}{2}°$ *out of phase with the applied voltage.*

Thus in the circuit considered in the above example the current would be reduced to 70 per cent of the value at resonance when the frequency is $\frac{1}{250}$ of 1000 kc, or 4000 cycles off resonance, and to 45 per cent of the resonant current for a frequency deviation of $\frac{1}{125}$ of 1000 kc, or 8000 cycles. Since the resonant rise of voltage in this circuit is 125 times, the rise of voltage is very nearly $0.7 \times 125 = 87.5$ times when the frequency is 4000 cycles off resonance, and is very close to $0.45 \times 125 = 56.25$ times at a frequency 8000 cycles from resonance.

[1] An error of 1 per cent is nearly always permissible in calculations of radio-frequency circuits. This is because the effective circuit constants at radio frequencies are very seldom known to an accuracy that involves an error of less than 1 per cent, and because many resonant-circuit formulas, such as Eqs. (22) and (24), contain a term $\left(\omega L - \frac{1}{\omega C}\right)$ which involves the difference of two large and nearly equal quantities. In order to obtain this difference without more than 1 per cent error five-place logarithms must be employed. Slide-rule calculations of $\left(\omega L - \frac{1}{\omega C}\right)$ will sometimes be in error as much as 100 per cent and are consequently of little use.

14. Parallel Resonance.—A parallel circuit consisting of an inductance branch in parallel with a capacity branch offers an impedance of the character shown in Fig. 30. At very low frequencies the inductive branch draws a large lagging current while the leading current of the capacity branch is small, resulting in a large lagging line current and a

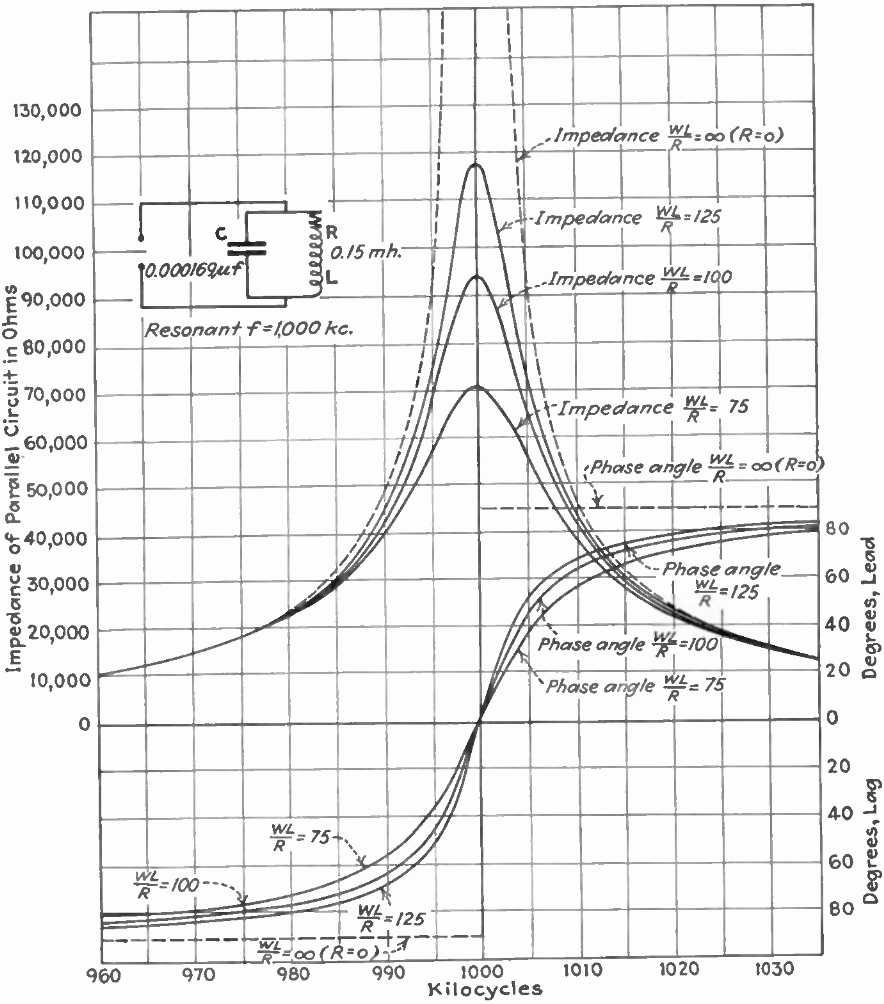


FIG. 30.—Magnitude and phase angle of impedance of a parallel circuit as a function of frequency. Curves are given for three representative values of circuit resistance and also for zero resistance.

low lagging circuit impedance. At high frequencies the inductance has a high reactance compared with the capacity, resulting in a large leading line current and a correspondingly low circuit impedance that is leading in phase. In between these two extremes there is a frequency at which the lagging current taken by the inductive branch and the leading current entering the capacity branch are equal, and being 180° out of phase they neutralize, leaving only a small resultant in-phase current flowing in the line. The impedance of the parallel circuit will then be a very high resistance, as is brought out in Fig. 30.[1]

The effect of circuit resistance upon the impedance of the parallel resonant circuit is very similar to the influence which resistance has upon the current flowing in a series resonant circuit, as is evident when Figs. 28 and 30 are compared. Increasing the resistance of a parallel circuit lowers and flattens the peak of the impedance curve without appreciably altering the sides, which are relatively independent of the circuit resistance. The dotted curve of Fig. 30 represents the impedance of the parallel circuit as a function of frequency when the circuit has no losses. It is apparent that for most practical purposes it is permissible to neglect the resistance in computing the impedance of a parallel resonant circuit at frequencies that differ from the resonant frequency by at least several per cent.

The resonant frequency of the parallel circuit is sometimes taken as the point of minimum line current, sometimes as the condition which makes the impedance a pure resistance, and sometimes as the frequency for which $\omega L = 1/\omega C$. These three definitions of resonance in parallel circuits lead to resonant frequencies that are different by such a very small fraction of 1 per cent when the circuit Q is at all large, that for all practical purposes the resonant frequency of a parallel circuit can be taken as the frequency that satisfies the relation.

$$\omega L = \frac{1}{\omega C} \tag{27a}$$

or

$$\text{Resonant frequency} = \frac{1}{2\pi\sqrt{LC}} \tag{27b}$$

With this definition the parallel resonant frequency of a tuned circuit is exactly the same as the series resonant frequency of the same circuit.

Analysis of Parallel Resonance.—In the analysis of the parallel resonant circuit of Fig. 30 the following notation will be used:

E = voltage applied to circuit

I_L = line current, *i.e.*, the current supplied to the circuit

[1] In obtaining a parallel resonance curve experimentally by measurements of applied voltage and line current, extreme care must be taken to insure that the applied voltage contains no harmonics. This is necessary because at resonance the circuit impedance is extremely high to the fundamental component of the applied voltage and very low to the harmonic components, with the result that even a small harmonic-voltage component will cause line currents that mask the small fundamental component.

I_1 = current in inductive branch
I_2 = current in capacity branch
Z = impedance of the two branches in parallel
f = frequency in cycles
f_o = resonant frequency
$\omega = 2\pi f$
$Q = \omega L/R$ of circuit.

The quantities L, C, and R refer to the inductance, capacity, and effective circuit resistance as indicated in Fig. 30.

The impedance of the parallel circuit is obtained by first deriving the expression for the admittance of the combination consisting of the two branches in parallel. This gives

$$\text{Admittance} = \frac{1}{Z} = j\omega C + \frac{1}{R + j\omega L}$$

Solving for Z gives the expression representing the impedance of the parallel circuit, namely

$$Z = \frac{j\omega L + R}{(1 - \omega^2 LC) + j\omega CR} \tag{28a}$$

Equation (28a) can be rearranged in a form more suitable for interpretation by dividing both numerator and denominator of the right-hand member by $j\omega C$ and then replacing the term L/C that results by its equivalent $(\omega_o L)/\omega_o C = (\omega_o L)^2 = (1/\omega_o C)^2$, which is the square of the reactance of one branch of the circuit at the resonant frequency. Making these changes in Eq. (28a) gives

$$Z = \frac{(\omega_o L)^2 + \dfrac{R}{j\omega C}}{R + j\left(\omega L - \dfrac{1}{\omega C}\right)} \tag{28b}$$

This equation, which is exact, can be simplified by neglecting $R/j\omega C$ in the numerator. The exact Eqs. (28a) and (28b) then take the approximate form

$$\left.\begin{aligned} Z &= \frac{(\omega_o L)^2}{R + j\left(\omega L - \dfrac{1}{\omega C}\right)} \\ Z &= \frac{(1/\omega_o C)^2}{R + j\left(\omega L - \dfrac{1}{\omega C}\right)} \end{aligned}\right\} \tag{28c}$$

At the frequency for which $\omega L = 1/\omega C$, which is the resonant frequency of the parallel circuit, the parallel impedance becomes a pure resistance having the value

$$\text{Parallel impedance at resonance} = \frac{(\omega_o L)^2}{R} \tag{29a}$$

$$= (\omega_o L)Q \tag{29b}$$

The parallel impedance at resonance is seen to be Q times the reactance of one branch of the circuit, or, in other words, there is a resonant rise of impedance of Q times in the parallel circuit at the resonant frequency.

The approximation made in deriving Eq. (28*c*) from Eq. (28*b*) introduces an error in the magnitude of Z that is roughly one part in $2Q^2$, or less than $\frac{1}{100}$ of 1 per cent with ordinary resonant circuits, and makes the computed phase angle of the circuit impedance too small, *i.e.*, too much leading, by an angle γ that has a tangent given by the equation

$$\tan \gamma = \frac{(R/\omega C)}{(\omega_o L)^2} \tag{30}$$

This correction can be made in results computed from the approximate formula (28*c*) if desired, but is usually unnecessary as the error is seldom greater than 0.5°.

Similarity of Parallel and Series Resonance.—Equation (28*c*) shows that the impedance of a parallel circuit is approximately equal to the square of the reactance of one branch of the circuit taken at resonance divided by the series impedance of the same circuit. This is apparent when it is observed that the denominator of Eqs. (28*b*) and (28*c*) is exactly the series impedance of the same circuit as given in Eq. (22). The impedance curve of a circuit connected for parallel resonance is accordingly of the same shape as the curve of current flowing when the same inductance, capacity and resistance are connected in series, because both the parallel impedance and the current in the series circuit are inversely proportional to the series impedance of the circuit. As a consequence of this the universal resonance curves of Fig. 29 can be employed to give the shape of the parallel impedance curve of any circuit to the same degree of approximation as contained in Eq. (28*c*). When used for this purpose the calibrations on the y-axis of the universal resonance curve represent the ratio of actual parallel impedance to the parallel impedance at resonance, instead of the ratio of actual current to resonant current as indicated on the figure, while the leading and lagging phase angles shown on the curve must be interchanged in the case of parallel circuits. The same working rules that gave the shape of the current curve for series resonance also apply to the impedance curve for parallel resonance; that is, when the fractional deviation of the frequency is $1/2Q$ off resonance the parallel impedance is 70 per cent of the value at resonance, and the phase angle is 45°, while if the fractional frequency deviation amounts to $1/Q$ of the resonance frequency these values become 45 per cent and 63.5° respectively.

Relations between Line and Branch Currents.—The line current I_L flowing into the parallel circuit is

$$
\left.\begin{aligned}
I_L &= \frac{E}{Z} \\
&= \frac{E}{\left[\dfrac{(\omega_o L)^2}{R + j\left(\omega L - \dfrac{1}{\omega C}\right)}\right]}
\end{aligned}\right| \tag{31a}
$$

At resonance the line current is a minimum and has the value

$$
\text{Line current at resonance} = \frac{E}{(\omega_o L)^2/R} = \frac{E}{(\omega_o L)Q} \tag{31b}
$$

The current I_1 in the inductive branch is

$$
I_1 = \frac{E}{R + j\omega L} \tag{32a}
$$

while the current I_2 in the condenser branch is

$$
I_2 = \frac{E}{(1/j\omega C)} \tag{32b}
$$

At resonance $\omega L = 1/\omega C$, and the branch currents are practically equal because R is small compared with ωL.

It is to be noted that although the line current is very small at resonance, the currents in the two branches are very large and do not go through any resonance curve. What actually happens in the parallel circuit at the resonant frequency is that there is a large current circulating between the inductance and the condenser, while the line current is just great enough to supply the circuit losses. Inasmuch as the resistance of the tuned circuit is low, the energy losses and hence the line current will be correspondingly small. Furthermore, as the losses in the circuit are proportional to the circuit resistance, the line current at resonance will also be proportional to the resistance R, as is also apparent from Eq. (31*b*). At frequencies other than the resonant frequency the line current increases because the line must supply some reactive energy to the parallel circuit in addition to the circuit losses.

The current that circulates between the branches of the parallel circuit at the resonant frequency depends only upon the applied voltage and the reactance of one branch, and is substantially independent of the circuit resistance. In contrast with this, the line current at resonance depends upon the circulating current and the Q of the circuit. Comparison of Eqs. (31*b*) and (32*a*) shows that for all practical purposes the branch currents at the resonant frequency are Q times the line current. *The parallel resonant circuit thus shows a resonant rise of current from line to branch of Q times.*

Components of Parallel Impedance.—The impedance of a parallel circuit can be broken up into resistance and reactance components with the aid of Eq. (28). When this is done for a typical case the result is as shown in Fig. 31. The resistance varies in much the same way with frequency as does the parallel impedance but is somewhat sharper. The circuit reactance reaches a maximum inductive value at a frequency less than resonance by a fractional deviation of $1/2Q$ (*i.e.*, difference between actual frequency and resonant frequency equal to resonant frequency

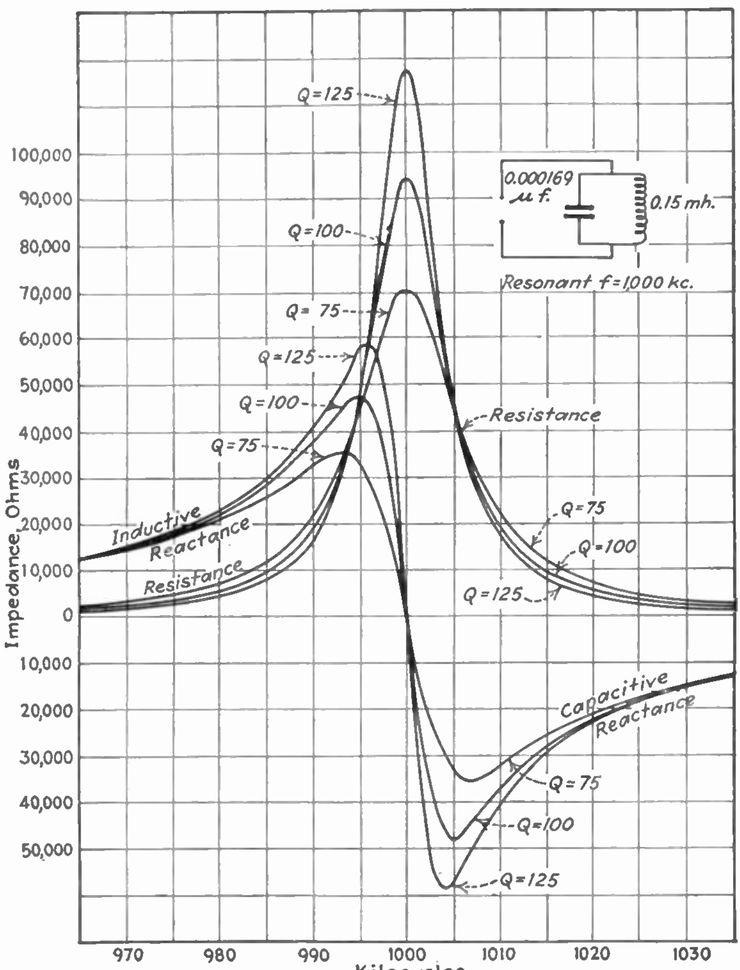


FIG. 31.—Resistance and reactance components of the parallel impedance of Fig. 30 shown as a function of frequency.

divided by $2Q$), then passes through zero at resonance and reaches a maximum capacitive value at a frequency exceeding resonance by a fractional deviation of $1/2Q$. These maximum values of the circuit reactance are both almost exactly one-half the impedance which the parallel circuit has at resonance (*i.e.*, one-half the maximum of the circuit resistance curve) and so are proportional to the circuit Q, as is clearly evident in Fig. 31. At frequencies somewhat different from resonance the reactance of the parallel circuit is substantially independent of the circuit Q while the resistance is inversely proportional to the circuit Q.

General Case of Parallel Resonance.—The discussion that has been given of parallel resonant circuits is a special case of the more general circuit in which Z_1 is the impedance of one parallel branch and Z_2 the impedance of the other. In this general circuit the parallel impedance Z is

$$Z = \frac{Z_1 Z_2}{Z_1 + Z_2} = \frac{Z_1 Z_2}{Z_S} \tag{33}$$

where Z_S is the series impedance of the circuit (*i.e.*, $Z_S = Z_1 + Z_2$). *The parallel impedance is therefore the product of the impedances of the separate branches divided by the impedance of the two branches in series.* This relation always holds true and is seen to reduce to Eqs. (28) and (29) when one branch is a capacitive reactance and the other is an inductive reactance with resistance. The properties of a parallel resonant circuit with inductive and capacitive branches, such as the circuit of Fig. 30, depend primarily upon the total series resistance and only to a slight extent upon the way in which the resistance is distributed between the branches, provided that the Q of the circuit is reasonably high. This is because Z_S of Eq. (33) is affected only by the total resistance and not by its location, while Z_1 and Z_2 are both large reactive quantities in a circuit having a reasonable Q and so are only slightly affected by the way in which the small circuit resistance is divided between them. Equations (28), (29), and (31) can therefore be used in practice with circuits having resistance in both branches, provided the R appearing in the equations is taken as the total series resistance of the circuit, and all the general conclusions that applied to the case with all resistance in the inductive branch also apply when the resistance is divided.

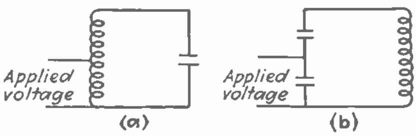


FIG. 32.—Parallel resonant circuits connected to make the parallel impedance at the circuit terminals less than the impedance across the complete circuit.

When the voltage is applied across only part of the inductive or part of the capacitive branch of a parallel resonant circuit, as illustrated in Fig. 32, the only effect on the impedance near the resonant frequency is to reduce the magnitude of the parallel impedance curve without changing its shape. The magnitude actually obtained will be approximately proportional to the square of the branch reactance between the points of voltage application. Thus the curve of parallel impedance for the case in Fig. 32*a* when the voltage is applied across only half of the inductance will have exactly the same shape as though the voltage had been applied across the entire inductance, but the impedance values will be only one-fourth as great. Tapping in at intermediate points on a branch of a parallel circuit accordingly serves to step down the parallel impedance to lower values without changing its characteristics near the parallel resonant frequency.

Parallel Resonance Effects in Inductance Coils.—The self-capacity associated with an inductance coil is in shunt with the inductance and thus makes the coil equivalent to a parallel resonant circuit. The result is that the apparent coil inductance as measured between the terminals increases with frequency until a maximum is reached just below the frequency at which the self-capacity is resonant with the coil inductance. The apparent inductance becomes zero at the parallel resonant frequency, *while for higher frequencies the coil has a capacitive reactance and is therefore equivalent to a small condenser.* The apparent resistance of the coil increases with frequency until a maximum is reached at the resonant frequency, beyond which the resistance rapidly diminishes. These effects are all direct consequences of the properties of parallel resonant circuits and can be readily deduced by an examination of Fig. 31 or of Eqs. (27) to (32).

The behavior of an inductance coil with self-capacity can be calculated just as one would determine the characteristics of any parallel circuit. The frequency at which the self-capacity and coil inductance are in parallel resonance is called the natural frequency of the inductance coil and represents the highest frequency at which the coil acts as an inductance. At frequencies that do not exceed 80 per cent of the natural frequency, Eq. (28a) can be rearranged to yield the following approximate results:[1]

$$\text{Apparent inductance of coil with self-capacity} = \frac{L}{1 - m^2} \qquad (34)$$

$$\text{Apparent resistance of coil with self-capacity} = \frac{R}{(1 - m^2)^2} \qquad (35)$$

where m is the ratio of actual frequency to natural resonant frequency of the coil while L and R are the true coil inductance and resistance, respectively. The apparent inductance and resistance given by these equations are the values one would find by actual measurement and represent the apparent but not the true quantities. Thus at a frequency one-half the natural frequency of the coil, $m = \frac{1}{2}$, and $(1 - m^2) = \frac{3}{4}$, and as a result of the coil capacity the apparent inductance and resistance as measured are respectively $\frac{4}{3}$ and $\frac{16}{9}$ of the actual values.

[1] Equations (34) and (35) are derived by rationalizing the denominator of Eq. (28a) and then neglecting the terms $(\omega CR)^2$ in the denominator and (ωRC) in the numerator, both of which are of negligible importance when the frequency is not too close to resonance. The real part of the result is the apparent resistance given in Eq. (35), and the reactive part can be simplified to give Eq. (34).

It is to be noted that the apparent Q of a coil is decreased by self-capacity. Because of the parallel resonance effects resulting from self-capacity it is customary to determine the inductance of radio coils at audio frequencies, where m is very small.

This discussion of the effect of self-capacity in inductance coils is approximate in that it assumes the self-capacity is lumped across the terminals of the coil as a single condenser, whereas it is actually distributed throughout the coil as shown in Fig. 14*a*. The principal effect of assuming the self-capacity lumped instead of distributed is to make the calculated natural frequency in error, but for frequencies that are appreciably lower than the natural period the approximation is very good.

15. Voltage and Current Distribution in Circuits with Distributed Constants.[1]—When the inductance, capacity, and resistance of a circuit are mixed together rather than being in separate lumps as in the case of the simple series and parallel circuits that have been considered, it is said that the circuit has distributed constants. Examples of circuits with distributed constants include telephone, telegraph, and power lines, as well as most types of radio antennas.

The distribution of voltage and current along a circuit with distributed constants, such as a transmission line, depends upon the impedance at the receiving end of the circuit and upon the distance from the receiving end to the point at which the voltage is applied to the circuit. The distributions with open- and short-circuited receivers are shown in Fig. 33 and are seen to involve very pronounced resonances. In considering these resonances it is convenient to measure distances along the circuit in terms of wave lengths. One wave length is twice the distance between adjacent minima and is given with high degree of accuracy by the equation

$$\left.\begin{array}{r}\text{Distance along circuit corre-}\\ \text{sponding to one wave length}\end{array}\right\} = \frac{1}{f\sqrt{LC}} \tag{36}$$

in which L and C are the series inductance and shunt capacity respectively per unit length of circuit, and f is the frequency of the applied voltage. Where the circuit consists of one or more straight wires in space, as is the case with an antenna or a 60-cycle power line, the distance corresponding to one wave length found from Eq. (36) will always be almost exactly the same as the wave length of radio waves of the same frequency.

[1] The subject of circuits with distributed constants is so extensive as to require an entire book devoted to this one topic if the treatment is to be adequate. The remarks made in this section should be considered as giving only a superficial treatment of those properties of such circuits that are of fundamental importance in radio communication. For further information relative to circuits with distributed constants the reader is referred to any good book on the theory of telephone or power lines.

Voltage and Current Distribution with Open- and Short-circuited Receiver.—When the receiving end of a circuit with distributed constants is open the voltage distribution is such that the voltage goes through minima at distances from the receiver corresponding to an odd number of quarter wave lengths and goes through maxima at distances corresponding to an even number of quarter wave lengths, always measured from the receiver. The current distribution for the open-circuited receiver has its minimum values where the voltage is maximum, and its maximum values where the voltage is smallest. The ratio of these maximum to

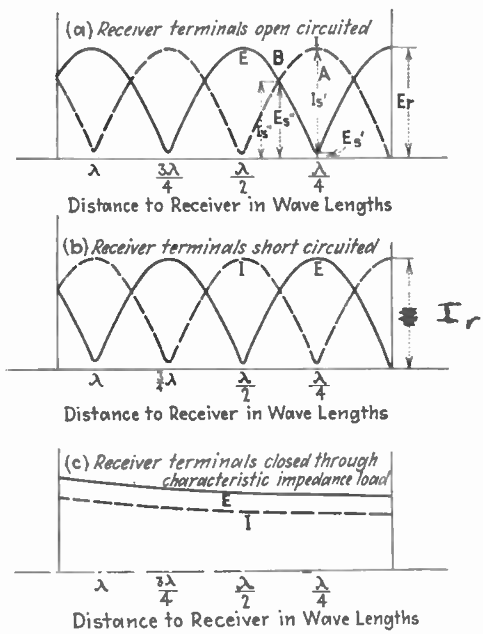


FIG. 33.—Typical voltage and current distributions for circuit with distributed constants, showing conditions existing with widely different receiver end conditions.

minimum values in a distribution curve, such as Fig. 33*a*, depends on the resistance per unit length of circuit and is greater the lower the resistance. Changing the frequency alters the distance representing one wave length and so changes the number of maxima and minima obtained in a given line length but does not otherwise affect the general character of the line behavior. The sending-end impedance of a circuit with distributed constants having an open receiver will be low when the line is an odd number of quarter wave lengths long because then the sending-end voltage is small and the current is high, while the impedance will be high when the line is an even number of quarter wave lengths in length.

The voltage that must be applied at the sending end of a circuit with distributed constants to develop a given receiving-end voltage (E_r in

Fig. 33*a*) on open circuit depends upon the length of the line in wave lengths. The ratio of receiver voltage to sending-end voltage is equal to the ratio of E_r in Fig. 33*a* to the height of the voltage curve in Fig. 33*a* at a distance from the receiver corresponding to the length of the line. Thus if the line is a quarter wave length long the sending end is at A, and an applied voltage E_s' will produce an open-circuit receiver voltage E_r, while if the line is three-eighth wave lengths long the sending end is at B, and an applied voltage E_s'' will be necessary to give the receiver voltage of E_r. The currents that will flow into two such lines will be I_s' and I_s'' respectively, as shown in the figure. It is seen therefore that when the circuit is exactly some odd number of quarter wave lengths in length there will be a resonant rise of voltage in the line, while if it is exactly an even number of quarter wave lengths long there will be no place along the line where the voltage exceeds the sending-end potential.

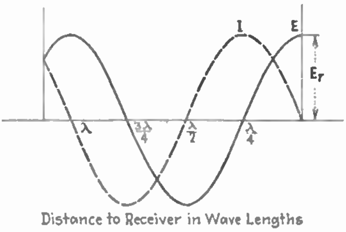


FIG. 34.—Schematic method of representing voltage and current distribution of Case *a* in Fig. 33. This diagram indicates the 180° phase shift on opposite sides of the minima and is an accurate representation of the conditions existing along the circuit except where the voltage and current are shown going through zero.

When the receiving end of a circuit with distributed constants is short-circuited, the voltage and current distribution are as shown in Fig. 33*b*. It will be noted that the voltage distribution in the short-circuited case is of exactly the same character as the current in the open-circuited receiver case, and that the current distribution with the receiver shorted is exactly the same as the voltage distribution with open receiver terminals. By making allowances for this interchange of voltage and current, everything that was said about the open-circuited receiver also applies to the case of the short-circuited receiver. Thus when the receiving terminals are short-circuited the voltage along the line is low at points that are an even number of quarter wave lengths from the receiver, and the current is low at all odd quarter wave lengths from the receiver. The sending-end impedance is low when the line length is an even number of quarter wave lengths long, and high at lengths that are an odd number of quarter wave lengths distant from receiver.

In the cases of both the open- and short-circuited receiver the voltage and current are substantially 90° out of phase at all places along the line except in the immediate vicinity of the quarter-wave-length points where the phase angle rapidly shifts from 90° on one side of unity power factor to 90° on the other side of unity power factor. The voltages on opposite sides of a voltage minimum are therefore substantially 180° out of phase, as are also the currents on opposite sides of a current minimum. In order to show this change of phase the voltage and current distributions in circuits with distributed constants are frequently drawn as shown in Fig. 34, which corresponds to *a* of Fig. 33 except that adjacent maxima are shown on opposite sides of the base line to indicate the reversal of phase. It will be noted that while this method of representation gives the distribution of most of the line accurately, it fails to show the conditions actually existing at the minima since the curves of Fig. 34 go through zero where the axis is crossed, whereas in reality the minima never reach zero if the line has any losses, and where there are losses the 180° phase shift takes place gradually, instead of all at one spot. In many circumstances, notably in the calculation of radiation from a circuit with distributed constants, the small amplitude at the minima is of negligible importance, and the convenient approximate representation of Fig. 34 yields results that are entirely satisfactory from a quantitative point of view. It will be noted that the distribution curves at *a* and *b* of Fig. 33 would be half sine waves if the minima reached zero. As a result the curves of Fig. 34 are sine wave in character, and this fact can be taken advantage of in solving problems involving distributions in circuits with distributed constants, provided of course that the effect of the finite though small minima can be neglected.

Voltage and Current Distribution with Characteristic Impedance Load.—When the receiving end of a circuit with distributed constants is closed through a resistance of $\sqrt{L/C}$ ohms, where L and C are respectively the inductance and capacity per unit length of circuit, the voltage and current distribution is as shown in Fig. 33*c* and the impedance at the sending end is a constant resistance of $\sqrt{L/C}$ ohms for all frequencies. This particular value of receiver resistance is called the characteristic impedance of the line and destroys all resonances, no matter what line length or frequency is being considered. When the circuit with distributed constants is a quarter wave length or greater in length and is to be used for the transmission of electrical energy, the receiver load impedance should equal the characteristic impedance in order to destroy resonances and prevent undesirably high currents and voltages from existing at places along the line. When the load resistance is greater than the characteristic impedance but is not infinite, the distribution contains resonances similar to those with the open receiver but the minima are not as small. Load resistances less than the characteristic impedance give distributions similar in general character to those existing with a short-circuited receiver but differ in that the amplitude at the minima is greater.

16. Inductively Coupled Circuits—Theory.—When mutual inductance exists between coils that are in separate circuits, these circuits are said to be inductively coupled. The effect of the mutual inductance is to make possible the transfer of energy from one circuit to the other by transformer action. That is, an alternating current flowing in one circuit produces magnetic flux that induces a voltage in the coupled circuit, resulting in induced currents, and a transfer of energy from the first or primary circuit to the coupled or secondary circuit. Several types of inductively coupled circuits commonly encountered in radio work are shown in Fig. 35.

The behavior of inductively coupled circuits is somewhat complex, but it can be readily calculated with the aid of the following rules.

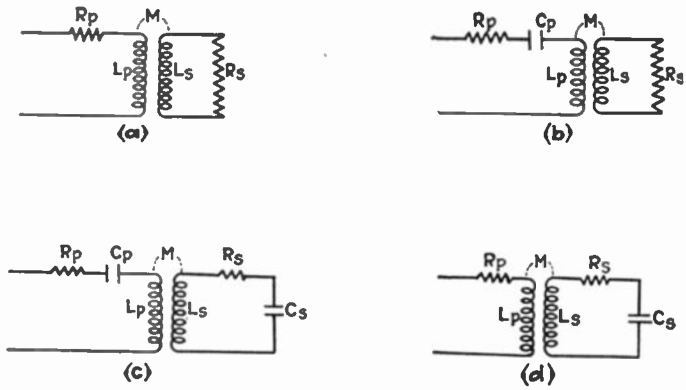


FIG. 35.—Various types of inductively coupled circuits commonly encountered in radio work.

Rule 1. *As far as the primary circuit is concerned the effect which the presence of the coupled secondary circuit has is exactly as though an impedance* $(\omega M)^2/Z_s$ *had been added in series with the primary,*[1] where

M = mutual inductance
ω = $2\pi f$
Z_s = series impedance of secondary circuit when considered by itself.

The equivalent impedance $(\omega M)^2/Z_s$ which the presence of the secondary adds to the primary circuit is called the coupled impedance, and since Z_s is a vector quantity having both magnitude and phase, the coupled impedance is also a vector quantity, having resistance and reactance components.

[1] This can be demonstrated by writing down the circuit equations for the primary and secondary. These equations are

$$E = I_pZ_p + j\omega MI_s$$
$$\text{Induced voltage} = -j\omega MI_p = I_sZ_s$$

where Z_p is the series impedance of the primary and E is the voltage applied to the primary. Solving this pair of equations to eliminate I_s gives

$$E = I_p\left[Z_p + \frac{(\omega M)^2}{Z_s}\right] \tag{37}$$

This relation shows that the effective primary impedance with secondary present is $Z_p + \dfrac{(\omega M)^2}{Z_s}$, of which the second term represents the coupled impedance arising from the presence of the secondary.

Rule 2. *The voltage induced in the secondary circuit by a primary current of I_p has a magnitude of $\omega M I_p$ and lags the current that produces it by 90°. In complex quantity notation the induced voltage is $-j\omega M I_p$.*

Rule 3. *The secondary current is exactly the same current that would flow if the induced voltage were applied in series with the secondary, and the primary were absent.* The secondary current therefore has a magnitude $\omega M I_p/Z_s$, and in complex quantity representation is given by $-j\omega M I_p/Z_s$.

These three rules hold for all frequencies and for all types of primary and secondary circuits, both tuned and untuned. The procedure to follow in computing the behavior of a coupled circuit is *first*, to determine the primary current with the aid of Rule 1; *second*, to compute the voltage induced in the secondary knowing the primary current using Rule 2; *finally*, to calculate the secondary current from the induced voltage by means of Rule 3. The following set of formulas will enable these operations to be systematically carried out.

$$\left.\begin{array}{l}\text{Impedance coupled into primary}\\ \quad\text{circuit by presence of the secondary}\end{array}\right\} = \frac{(\omega M)^2}{Z_s} \tag{38}$$

$$\text{Equivalent primary impedance} = Z_p + (\omega M)^2/Z_s \tag{39}$$

$$\text{Primary current} = I_p = \frac{E}{Z_p + \dfrac{(\omega M)^2}{Z_s}} \tag{40}$$

$$\text{Voltage induced in secondary} = -j\omega M I_p \tag{41}$$

$$\text{Secondary current} = \frac{-j\omega M I_p}{Z_s} \tag{42}$$

In these equations:

M = mutual inductance between primary and secondary
$\omega = 2\pi f$
Z_s = series impedance of secondary circuit considered as though primary were removed
Z_p = series impedance of primary circuit considered as though secondary were removed
E = applied voltage.

The primary and secondary impedances Z_p and Z_s, respectively, are vector quantities, so that Eqs. (38) to (42) are consequently all vector equations. Thus in the case of the coupled circuits of Fig. 35*c* one has:

$$Z_p = R_p + j\left(\omega L_p - \frac{1}{\omega C_p}\right)$$

$$Z_s = R_s + j\left(\omega L_s - \frac{1}{\omega C_s}\right)$$

For other circuits the impedances Z_p and Z_s are given by expressions similar to these, but differing in detail.

Action of the Coupled Impedance.—Many of the important properties of coupled circuits can be determined by examining the nature of the coupled impedance $(\omega M)^2/Z_s$. When the mutual inductance M is very small the impedance coupled into the primary by the presence of the secondary is correspondingly small, and the primary current is very nearly the same as though no secondary were present. The coupled impedance is also very small when the secondary impedance Z_s is large because even though a large mutual inductance induces a high voltage in the secondary, the secondary current produced is small as a result of the high impedance, and little energy transfer takes place. When the secondary impedance Z_s is low, and the mutual inductance is not too small, the coupled impedance $(\omega M)^2/Z_s$ is large, and the voltage and current relations in the primary circuit are affected to a considerable extent by the presence of the coupled secondary. In particular, when the secondary is a series resonant circuit, as in Figs. 35*c* and 35*d*, the secondary impedance Z_s will be very low at the resonant frequency of the secondary, so that at this frequency the coupled impedance will be very high, and the presence of the secondary will be an important factor in determining the primary current.

The coupled impedance has the same phase angle as the secondary impedance Z_s, with the exception that the sign of the angle is reversed; that is, if the secondary impedance is inductive and has an angle of 30° lagging, the impedance coupled in series with the primary circuit by the action of the secondary has a phase angle of 30° leading. The physical significance of this change from lagging to leading is that coupling a secondary circuit having an inductive reactance to a primary circuit is equivalent to neutralizing some of the inductive reactance already possessed by the primary, and this is done electrically by postulating a capacitive reactance of suitable magnitude in series with the inductance to be neutralized. When the secondary impedance Z_s is a pure resistance the coupled impedance will also be a resistance. This is a very important special case because when the secondary is a tuned circuit, as in Figs. 35*c* and 35*d*, the secondary impedance will be a resistance at resonance.

The energy consumed by the secondary circuit is the energy represented by the primary current flowing through the resistance component of the coupled impedance. In the same way the reactive volt-amperes transferred from the primary to the secondary are the volt-amperes developed by the primary current flowing through the reactive part of the coupled impedance. By expressing the secondary impedance Z_s in terms of its real and reactive components R_s and X_s respectively, one can rewrite Eq. (38) in the following form:

$$\text{Coupled impedance} = \frac{(\omega M)^2 R_s}{R_s^2 + X_s^2} - j\frac{(\omega M)^2 X_s}{R_s^2 + X_s^2} \qquad (43)$$

in which the resistance component of the coupled impedance is $\dfrac{(\omega M)^2 R_s}{R_s^2 + X_s^2}$, and the reactance component is $-j\dfrac{(\omega M)^2 X_s}{R_s^2 + X_s^2}$. The effect which the secondary has on the primary circuit is then exactly as though this resistance and this reactance had been inserted in series with the primary circuit, and as though this resistance and reactance consumed the energy and the reactive volt-amperes transferred to the secondary.

If the voltage acting in the primary is constant, the secondary current is a maximum when the frequency and mutual inductance are such as simultaneously to satisfy the conditions: (1) coupled resistance equal to the primary resistance, and (2) coupled reactance equal in magnitude to the primary reactance but opposite in sign. It is possible to derive formulas for various interesting situations, such as the mutual inductance giving maximum secondary current, and so on, but the general equations that result are not very useful because they are too complicated to permit of easy visualization or ready computation.[1]

It will be noted that although two inductance coils between which mutual inductance exists constitute a transformer, the inductively coupled circuit is not treated here in terms of leakage reactances, turn ratio, magnetizing current, and so on, as are power transformers. The reason for this is that in transformers used for radio work the leakage reactances are very high, causing the ratio of voltage transformation to be entirely different from the turn ratio. Another way of looking at the difference is to remember that the ordinary power transformer may have a leakage reactance of 2 per cent, but that it is not uncommon for a coupled circuit used in radio work to have a leakage reactance of over 99 per cent. As a result of these differences it is much easier to analyze coupled circuits used in radio in terms of mutual inductance rather than by turn ratio and leakage reactance.

17. Inductively Coupled Circuits with Tuned Secondary.—Coupled circuits of this type are of such importance in radio communication as to warrant detailed consideration. The impedance which a tuned secondary circuit couples into a primary has all of the essential characteristics of the impedance of this same tuned secondary circuit when acting as a parallel resonant circuit. The extent to which this is true can be seen by comparing the parallel resonant impedance curve of Fig. 30 with the impedance which this same circuit will couple into a primary, as given in Fig. 36. The similarity in shape of these two curves is striking, and the only essential difference is in the relative magnitude of the impedances involved.

The characteristics of the coupled impedance can be deduced from an examination of Eq. (38), which states that the coupled impedance is

[1] Formulas for a large number of such relations are to be found in G. W. Pierce, "Electric Oscillations and Electric Waves," Chap. XI.

equal to $(\omega M)^2/Z_s$. The quantity ωM entering into this expression changes only gradually with frequency and is substantially constant over the limited frequency range in which resonance phenomena show their important characteristics. Over such limited frequency ranges the

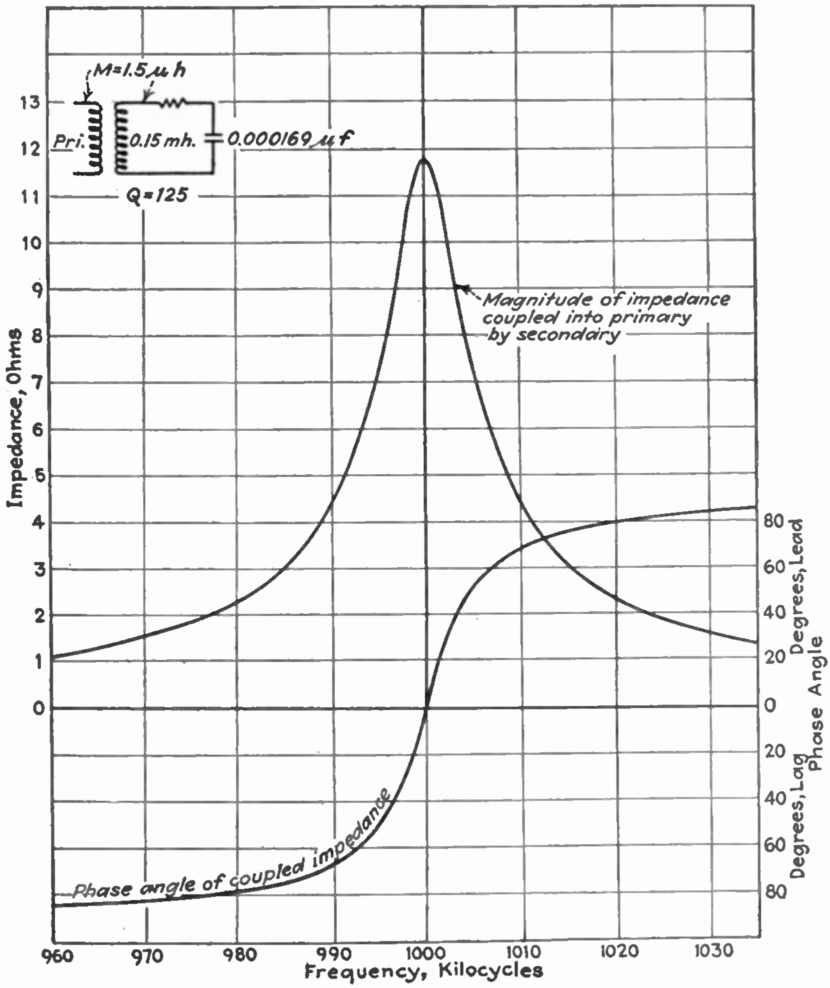


FIG. 36.—Magnitude and phase angle of the coupled impedance which a typical tuned secondary circuit couples into the primary. It will be observed that these curves have almost exactly the same shape as the corresponding characteristics obtained with parallel resonance.

coupled impedance is hence nearly inversely proportional to Z_s, the series impedance of the secondary circuit. It will be recalled that the impedance of a parallel resonant circuit was also shown to be almost exactly inversely proportional to the series impedance of the circuit, and this common property explains the similarity in the curves for coupled and parallel impedance. The coupled impedance produced by a tuned secondary is a pure resistance at resonance, inductive at frequencies less than resonance and capacitive above, just as is the case with the impedance of a parallel circuit. The differences between the two do not become appreciable until one goes to frequencies far off resonance.

Tuned Secondary Coupled to Untuned Primary Consisting of an Inductance in Series with a Resistance.—This case is illustrated in Fig.

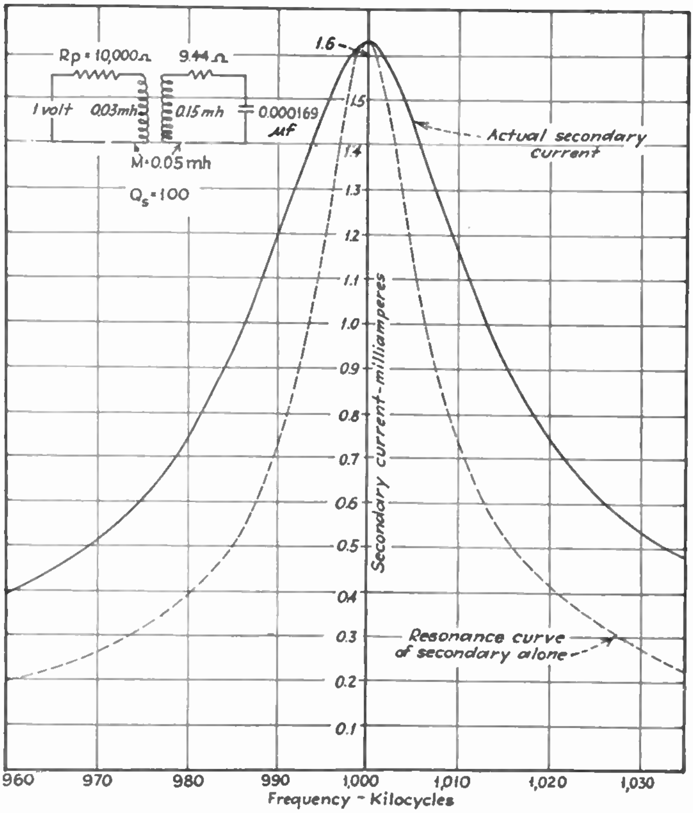


FIG. 37.—Resonance curve of secondary current in a coupled circuit of the type employed in tuned radio-frequency amplifiers, together with resonance curve of the secondary circuit alone, shown to indicate how the resonance curve of the secondary is broadened by being coupled to the primary.

35*d* and is very important because when the primary resistance R_p represents the plate resistance of a vacuum tube this is the equivalent circuit of the transformer-coupled tuned radio-frequency amplifier. The current produced in the secondary of a typical circuit of this type by a constant applied voltage is shown as a function of frequency in Fig. 37. The important feature of this curve is that while similar in general character to the series resonance curve of the secondary circuit, which is given in Fig. 28, it is not as sharp as that of the tuned circuit taken alone because the primary resistance R_p has the effect of increasing the equivalent resistance of the tuned secondary circuit. This, as well as several other features of the circuit will be discussed in detail in connection with tuned radio-frequency amplifiers.

Two Resonant Circuits Separately Tuned to the Same Frequency and Coupled Together—General Behavior.—Circuits of this type can be expected to show the type of behavior illustrated in Fig. 38, which gives the primary and secondary currents of a representative circuit as a function of frequency for different coefficients of coupling when the applied voltage is constant.

At very loose coupling (small coefficient of coupling) the primary current curve is practically the same as though there were no secondary present, the secondary current maximum is small, and the secondary current curve is much more peaked than the resonance curve of either the primary or secondary circuits. These characteristics are shown by the curves for $k = 0.002$ in Fig. 38. With very small coefficients of coupling the secondary current curve has a shape that is very nearly proportional to the product of the resonance curves of the primary and secondary circuits.[1] The maximum primary and secondary currents occur at the same frequency, which is the frequency to which the two circuits are tuned.

As the coefficient of coupling, and hence the mutual inductance, is increased to somewhat greater values the impedance which the secondary couples into the primary becomes of sufficient magnitude to alter the primary current appreciably. At resonance the effect is to decrease the primary current to values less than would be obtained with the secondary removed, while at frequencies above and below resonance the primary current is increased by the presence of the secondary. Both of these influences tend to broaden the resonance curve of primary current. At the same time, the increased coupling causes the voltage induced in the secondary, and hence the secondary current, to become much greater than with very loose coupling, and the resonance curve of secondary current is not quite as sharp because the broader curve of primary current makes the induced voltage change less rapidly with frequency. These characteristics of primary and secondary currents are shown in Fig. 38 by the curves for $k = 0.005$.

The effects which the increased coupling has on the primary current curve are results of the nature of the coupled impedance developed by a

[1] This is because the primary current follows the resonance curve of the primary and so induces a voltage in the secondary that varies with frequency in much the same way that the primary current does. The voltage induced in the secondary is therefore maximum at the resonant frequency of the primary (which is also the resonant frequency of the secondary in this case) and is much smaller for either higher or lower frequencies. This variation in induced voltage accordingly accentuates the sharpness of the secondary current resonance in the manner stated.

tuned circuit. At resonance the secondary couples a resistance into the primary, increasing the effective resistance of the primary circuit and thereby reducing the primary current. At frequencies appreciably below the common resonant frequency of the two coupled circuits the series reactance of the primary is capacitive, while the coupled impedance is largely inductive and therefore tends to neutralize part of the primary reactance that would exist if no secondary were present. The result is

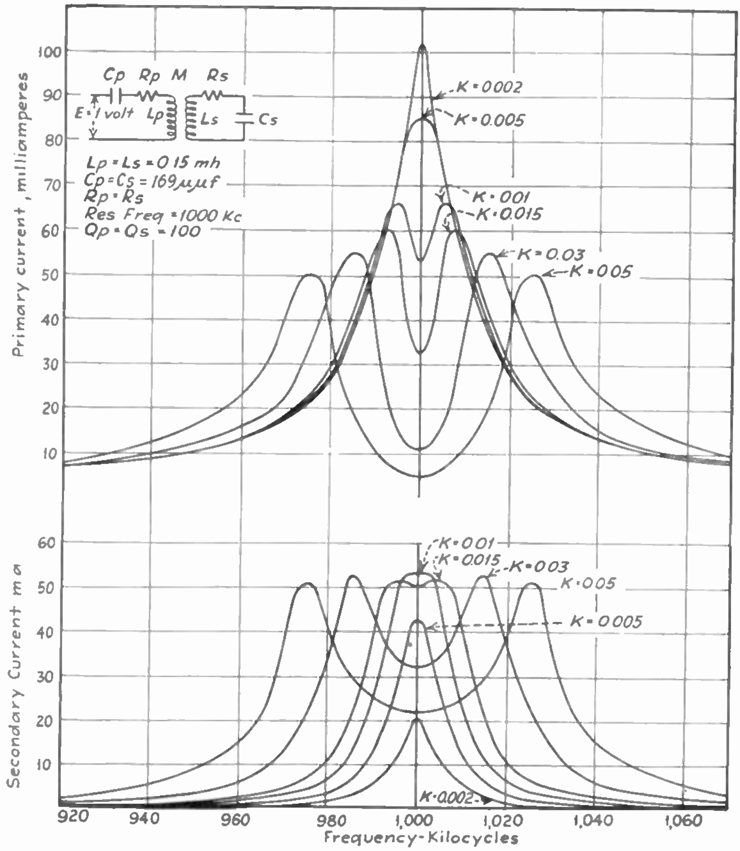


Fig. 38.—Curves for two circuits separately tuned to the same frequency and coupled together, showing variation of primary and secondary currents with frequency for a number of coefficients of coupling.

that the presence of the secondary lowers the effective primary reactance and increases the primary current. At frequencies somewhat above resonance a similar situation exists because then the coupled reactance is largely capacitive, while the reactance of the primary circuit taken alone is inductive, and the result is a lowered effective primary impedance and an increase in the primary current.

Critical Coupling When Two Circuits Tuned to the Same Frequency Are Coupled Together.—These effects of increased coupling continue until the resistance coupled into the primary circuit at the resonant frequency of the two circuits is equal to the primary resistance. This condition exists when the mutual inductance M satisfies the relation

$$\frac{(\omega M)^2}{R_s} = R_p \tag{44a}$$

or

$$\omega M = \sqrt{R_p R_s} \tag{44b}$$

The quantity ω in these equations is to be evaluated at the common resonant frequency. The coefficient of coupling satisfying Eq. (44b) is called the *critical coupling* by reason of the fact that it satisfies the conditions of maximum transfer of energy to the secondary and so gives the maximum possible secondary current that can be obtained. All values of mutual inductance either greater or less than that for critical coupling will result in a lower peak for the curve of secondary current.

When critical coupling exists the sharpness of the curve of secondary current is intermediate between that of the resonance curve of the secondary considered as a series circuit with constant applied voltage, and the curve of secondary current at extremely small couplings. The secondary current peak occurs at the resonant frequency of the two circuits and has a value $E/2\sqrt{R_p R_s}$, which, as has been stated, is the maximum secondary current possible under any condition. The primary current curve now shows a double hump, the current being less at resonance than at frequencies either slightly higher or lower. This behavior is caused by the coupled reactance neutralizing the primary reactance at frequencies slightly off resonance. The curve for $k = 0.01$ in Fig. 38 shows the type of characteristic that can be expected with critical coupling.

The coefficient of coupling k required to give critical coupling can be expressed in terms of the resonance ratios $Q_p = \omega L_p/R_p$ and $Q_s = \omega L_s/R_s$ of the two tuned circuits by substituting these two relations into Eq. (44). When this is done the critical coefficient of coupling is found to be:

$$\text{Critical } k = \frac{M}{\sqrt{L_p L_s}} = \frac{1}{\sqrt{Q_p Q_s}} \tag{45}$$

The critical k with tuned circuits having Q's in the neighborhood of 100 to 200, is roughly 1 per cent to ½ per cent, which represents a rather small mutual inductance.

As the coupling is increased beyond the critical value the two primary peaks become more pronounced and more widely separated, as is shown by the curves for $k = 0.015$, $k = 0.03$, and $k = 0.05$ in Fig. 38. This behavior results from the nature of the impedance which the secondary couples into the primary as a result of the increased mutual inductance. The two humps in the primary current represent resonances produced when the coupled reactance exactly neutralizes the primary reactance, and are spaced farther apart as the coupling is increased because the frequency off resonance at which the coupled and primary reactances are of equal magnitudes is greater with larger mutual inductances.

Secondary Current When Two Circuits Tuned to the Same Frequency Are Coupled Together.—Along with these changes of primary current that take place with increasing mutual inductance there are corresponding changes in the secondary current curve as shown in Fig. 38. The secondary current curve follows through the same general sequence of changes as does the primary current characteristic, first becoming lower and flatter and then separating into a two-humped curve with a minimum between the humps at the frequency to which the circuits are tuned. The principal difference between the curves of primary and secondary current is that the drop in current between the two humps is less pronounced in the latter case. The voltage induced in the secondary has the same general characteristics as the primary current curve and will therefore have two humps when the primary current does, but the low impedance of the secondary at its resonant frequency tends to counteract the low voltage that is induced at this frequency, with the result that the humped characteristic appears in the primary at smaller values of k than in the secondary.

The peaks of primary and secondary currents occur at practically the same frequencies and are very nearly symmetrically located on each side of the resonant frequency of the two circuits. The height of the current peaks decreases as the coupling is increased, but the two companion humps always have approximately the same maximum.

The secondary current curve begins to show signs of double humps when the coefficient of coupling is so related to the circuit resistance as to be about one and a half times the critical coupling, and these humps become quite pronounced when the coupling is twice the critical value. The frequency at which these humps appear is substantially independent of the circuit resistance, particularly if the resistance is low enough to give distinct double peaks, and so can be obtained with good accuracy by assuming the circuits to have zero resistance. The peaks then exist at frequencies which make the equivalent primary impedance zero, that is, when

$$j\left(\omega L_p - \frac{1}{\omega C_p}\right) + \frac{(\omega M)^2}{j\left(\omega L_s - \frac{1}{\omega C_s}\right)} = 0$$

Simplifying this equation by substituting $\omega_0 = 1/\sqrt{LC}$ and $k = (M)/\sqrt{L_pL_s}$ yields the very convenient relation

$$\frac{\text{Frequency at coupling hump}}{\text{Resonant frequency of tuned circuits}} = \frac{1}{\sqrt{1 \pm k}} \qquad (46)$$

This equation is exact when the circuit resistances are zero and gives results satisfactory for most practical purposes whenever well-defined double humps of secondary current are present. The plus sign denotes the lower frequency and the minus sign the higher frequency peak.

After the location of the coupling humps has been obtained from Eq. (46) an exact calculation of the currents at these frequencies by Eqs. (38) to (42) will give the height of the peaks, while a similar calculation

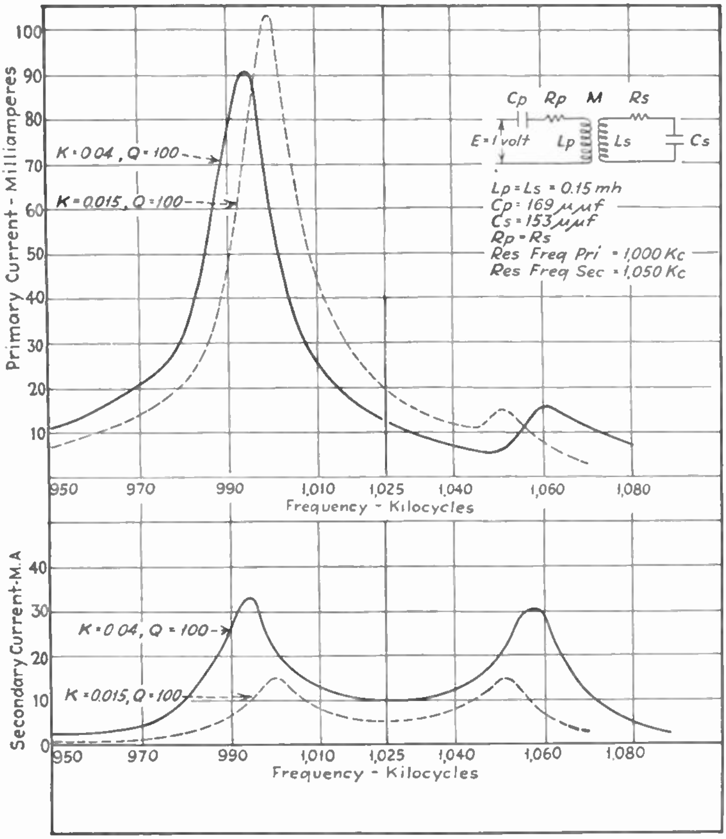


FIG. 39.—Curves of primary and secondary current obtained with coupled circuit in which primary and secondary are tuned to frequencies that differ considerably.

at the frequency to which the two coupled circuits are tuned will determine the current midway between the two humps. This information enables a fair estimate to be made of the characteristics of the coupled circuit in the vicinity of the resonant frequency and, when supplemented by calculations made at frequencies sufficiently removed from resonance as to make it permissible to neglect circuit resistances, enables an accurate sketch to be made of a double-peaked resonance curve without the necessity of extensive computations.

Two Coupled Resonant Circuits Tuned to Different Frequencies.—When the primary and secondary circuits are tuned to frequencies that differ considerably one can expect the type of behavior illustrated in Fig. 39. Both primary and secondary current curves have humps at frequencies that approximate the resonant frequencies of the individual circuits, with the approximation being closer as the mutual inductance is decreased.

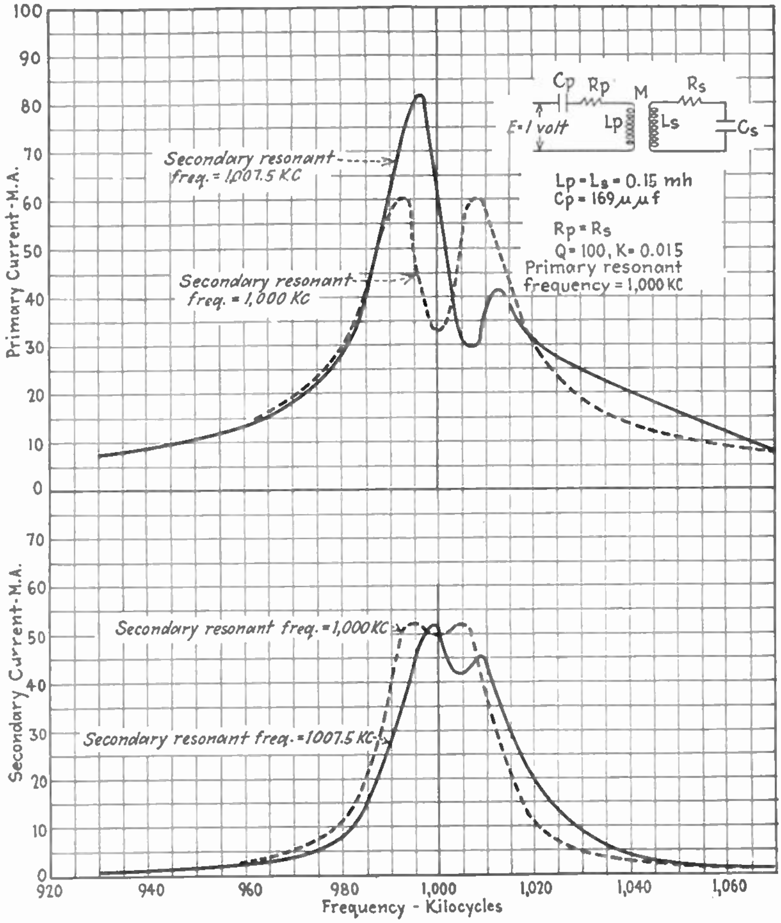


FIG. 40.—Curves of primary and secondary current obtained with coupled circuit in which primary and secondary are tuned to frequencies that differ by only a small amount.

The two humps of current will generally be of unequal magnitude, with the primary current tending to have its largest peak near the resonant frequency of the primary, and the secondary current tending to be maximum near the resonant frequency of the secondary.

When two coupled circuits are tuned to frequencies that differ by only a small amount, as might be the case if one of the circuits in Fig. 38 was slightly mistuned, the effect is to make the two current humps of unequal height. The frequencies at which the humps occur will also be changed slightly by the mistuning. These effects are clearly shown in Fig. 40, which is typical of what can be expected.

18. Mutual Inductance Effects Produced by Masses of Conducting Material and Short-circuited Turns.—The most important case coming under this heading is that of a coil surrounded by a non-magnetic shield of the type described in Sec. 11. Such a shield represents a secondary circuit made of resistance and inductance, which is inductively coupled to the coil. The effective coil resistance under such circumstances is the sum of its actual resistance and the coupled resistance, and will thus be greater than when the shield is removed. The presence of the shield also reduces the effective inductance of the coil since the secondary is an inductive reactance and hence couples a capacitive reactance into the coil circuit that neutralizes a portion of the actual coil inductance. The magnitude of these effects depends upon the mutual inductance between coil and shield.

This effect which a conducting mass has in lowering the effective inductance is often made use of to vary the inductance of large coils used in high-power radio transmitters. This is done by placing a copper disk or ring in the magnetic field of the coil with provision for adjusting the location of the conducting mass with respect to the lines of flux in such a way as to vary the coupling with the coil.

A short-circuited turn in an inductance coil represents a low-impedance secondary having a lagging reactance, just as does a shield or a mass of conducting material. A short-circuited turn hence lowers the effective coil inductance and raises the resistance somewhat. In large inductance coils used in radio transmitters the inductance is often varied in steps by short circuiting turns, with fine adjustment obtained by means of an adjustable copper ring or disk. While this arrangement introduces extra losses when the inductance is low, the reduction in effective Q is not serious, and by being short-circuited the unused turns are prevented from forming subsidiary resonant circuits.

Any metal object in the magnetic field of a coil also represents an inductively coupled secondary circuit composed of resistance and inductance in series. Binding posts, screws, condensers, metal panels, and similar objects either on or near the coil are examples of such secondaries, and can increase the effective coil resistance appreciably unless care is taken to place them where the magnetic field is weak, *i.e.*, where the mutual inductance between coil and object is low.

19. Other Types of Coupling. *Complex Coupling.*—In addition to inductive coupling, energy may be transferred between two circuits by means of other forms of coupling such as those shown in Figs. 10 and 41. Several cases of complex coupling are included in Fig. 41, such as *a*, where inductive coupling is shown combined with direct coupling, and at b and c, which are complicated forms of capacitive coupling. It would also be possible to couple two circuits by the aid of a common resistance, but this is very seldom done.

The various coupled circuits shown in Figs. 10 and 41 have characteristics similar in all respects to those of the inductively coupled circuit and can be analyzed by slightly modifying the formulas for inductive coupling. Thus in circuit b of Fig. 10 the primary current I_1 and the

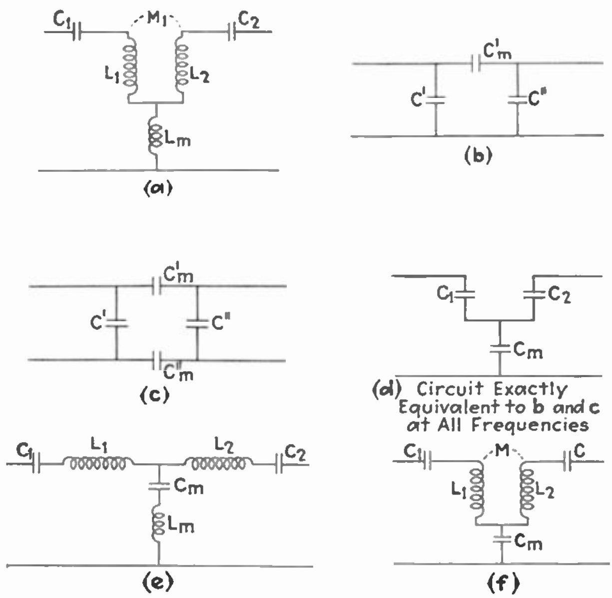


FIG. 41.—Some of the methods that may be used in coupling circuits.

secondary current I_2 are exactly the same as though this was an inductively coupled circuit in which

$$M = L_m$$
$$L_p = L_1 + L_m$$
$$L_s = L_2 + L_m$$

With these definitions all the properties of the inductively coupled circuit apply to Fig. 10b without change. The current through the coupling inductance L_m is equal to the difference between the primary and secondary currents, *i.e.*, $I_m = I_1 - I_2$.

Circuit a of Fig. 41 is also exactly equivalent to an inductively coupled circuit in which

$$M = L_m + M_1$$
$$L_p = L_1 + L_m$$
$$L_s = L_2 + L_m$$

where M_1 is the mutual inductance between L_1 and L_2 and may be either positive or negative depending on the direction of the turns in L_1 and L_2.

Capacitive Coupling.—Capacitively coupled circuits also have characteristics similar to the inductively coupled circuit. Thus circuit c of Fig. 10 can be analyzed with the equations of the inductively coupled circuit if these equations are modified by substituting $1/j\omega C_m$ for $j\omega M$ and $1/(\omega C_m)^2$ for $(\omega M)^2$ in Eqs. (38) to (42). The equivalent primary circuit is considered as having a capacity of $C_1C_m/(C_1 + C_m)$ while the equivalent capacity of the secondary is $C_2C_m/(C_2 + C_m)$. There are therefore two humps of secondary current if the coupling is large, *i.e.*, if condenser C_m is small, and there is only one peak of secondary current when the coupling is small, *i.e.*, when condenser C_m is large.

The capacitively coupled circuits shown at b and c of Fig. 41 are equivalent to each other and can be reduced to the circuit shown at d of Fig. 41. The conversion from the Δ of condensers shown at b to the Y of condensers given at d is carried out by the use of Eq. (50), which in this special case yields the conversion formulas:

$$C_1 = \frac{C'C'' + C'C_m' + C_m'C''}{C''}$$

$$C_2 = \frac{C'C'' + C'C_m' + C_m'C''}{C'}$$

$$C_m = \frac{C'C'' + C''C_m' + C_m'C''}{C_m'}$$

The notation is indicated in Fig. 41. It will be observed that this conversion is independent of frequency.

Circuits having combined electromagnetic and electrostatic coupling, such as those at e and f of Fig. 41, behave as ordinary coupled circuits except that the coefficient of coupling varies with frequency. Thus in the case of circuit e the circuit is capacitively coupled at low frequencies and inductively coupled at high frequencies because the coupling combination of C_m in series with L_m has capacitive and inductive reactance under these respective conditions. In between, at the resonant frequency of L_m and C_m, there is no coupling, and $k = 0$. The arrangement shown at f acts similarly as a circuit with a coefficient of coupling that varies with frequency. Depending upon the relative direction in which primary and secondary are wound in f, the mutual inductance M may introduce an inductive coupling that either aids or opposes the capacitive coupling. Circuits having combined electrostatic and electromagnetic coupling find practical application where it is desired to obtain a coefficient of coupling that varies with frequency. The proper combination of electromagnetic and electrostatic coupling will often eliminate mechanically adjustable coupling that would otherwise be necessary.[1]

[1] A number of practical applications of combined electromagnetic and electrostatic coupling are described in Edward H. Loftin and S. Young White, Combined Electromagnetic and Electrostatic Coupling, and Some Uses of the Combination, *Proc. I.R.E.*, vol. 14, p. 605, October, 1926.

20. Band-pass Filters.—If two resonant circuits tuned to the same frequency are suitably coupled, the secondary current will be substantially constant over a band of frequencies in the vicinity of resonance, while currents of all other frequencies will be discriminated against sharply. A pair of coupled circuits adjusted to give this result is called a band-pass filter because a band of frequencies is transmitted about equally well, while currents of all other frequencies are suppressed. This is in contrast with the resonance curves of the series and parallel circuits, which have rounded tops that do not give uniform response over a range of frequencies. Band-pass characteristics can also be exhibited by the primary as well as by the secondary current in a system comprising two coupled circuits tuned to the same frequency. Typical band-pass characteristics are exhibited in Fig. 38 by the curve of secondary current for $k = 0.015$ and of primary current for $k = 0.005$. The band-pass effect in the case of primary current is not so pronounced as in the secondary and so is practically never used.

The important characteristics of a band-pass filter are the width of the pass band, the uniformity of response to the different frequencies within the band, and the response to the frequencies lying within the band compared with the response to those that are without. When a band-pass filter is formed by coupling two tuned circuits, the width of the pass band depends primarily upon the coefficient of coupling and very little upon other factors, while the uniformity of response within the pass band is fixed by the relation which the Q of the resonant circuits bears to the coefficient of coupling. When the Q and the coefficient of coupling are fixed by these considerations the relative response to frequencies within and without the pass band is therefore indirectly determined and is not subject to independent control.

The behavior of a typical band-pass filter with variations in circuit Q and coefficient of coupling is well brought out by the curves of Fig. 42, which exhibit the characteristics displayed by all such band-pass circuits. The effect which the circuit Q has on the width of the pass band and the uniformity of response within the band is well brought out by the series of curves at *a* of Fig. 42, which are for the same coefficient of coupling but varying circuit resistance. When the resistance is low, *i.e.*, Q is large, very pronounced double humps appear, with frequencies at the edge of the pass band receiving much greater transmission than those near the center. In contrast with this, a high resistance, *i.e.*, low Q, causes the band-pass characteristic to be replaced by a curve with a rounded top similar to the resonance curve of a simple series or parallel circuit. In between these two extremes is a circuit Q giving the desired band-pass characteristic with substantially equal response to a band of frequencies. This proper value of circuit Q depends upon the coefficient of coupling and will be less as the coupling is made greater. The width of the pass band is seen to be substantially independent of the circuit resistance and has a value that approximates very closely the separation of the two humps existing when the circuit resistance is zero.

The effect which the coefficient of coupling has on the width of the pass band is demonstrated by the curves shown at b of Fig. 42, which

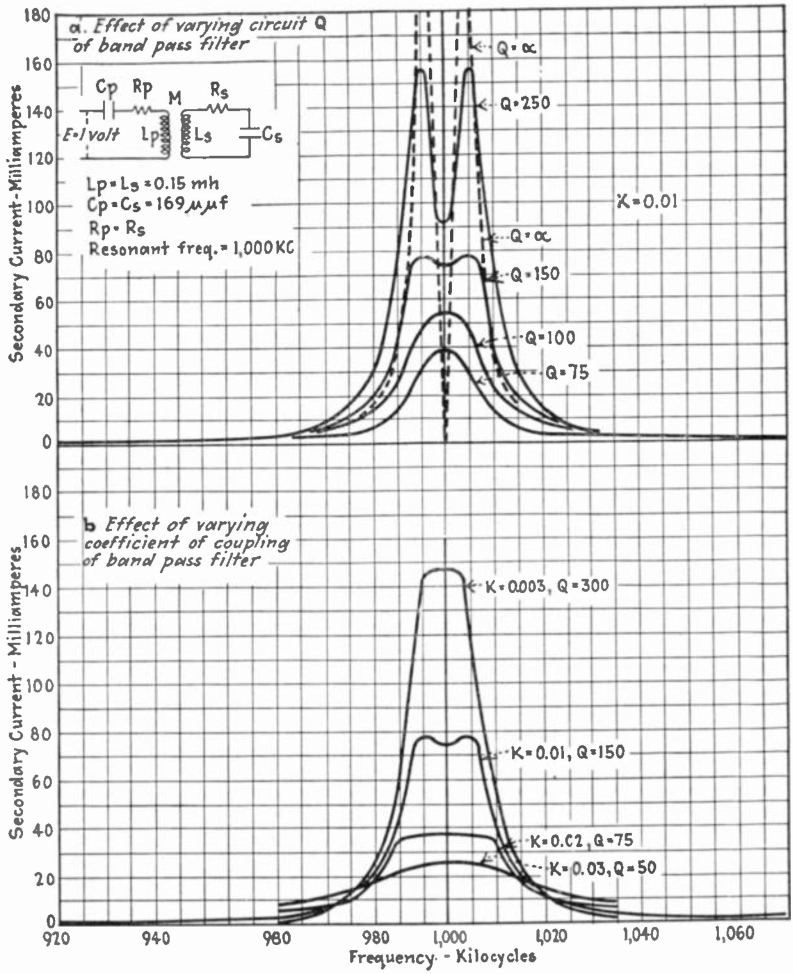


FIG. 42.—Characteristics of band-pass filter, showing (a) effect of circuit Q on uniformity of response within the pass band, and (b) effect of coefficient of coupling upon the width of the pass band and the response within the pass band when the proper circuit Q is used.

show the characteristics with different degrees of coupling when the resistance with each value of coupling is the value giving the best band-pass characteristics. The width of the pass band is very nearly directly proportional to the coefficient of coupling, while the response within the pass band is nearly inversely proportional to the band width.

Design of Band-pass Filters.—The design of band-pass filters can be carried out by simple approximate methods which give an accuracy sufficient for most practical purposes and at the same time avoid the not difficult but very lengthy exact calculations employing Eqs. (38) to (42). If the width of the pass band is considered as the spacing between the coupling humps obtained with zero circuit resistance, then the width of the pass band is related to the common resonant frequency of the two tuned circuits and the coefficient of coupling k in the following way, which is derived from Eq. (46).

$$\frac{\text{Width of pass band}}{\text{Resonant frequency of tuned circuits}} = \frac{\sqrt{1+k} - \sqrt{1-k}}{\sqrt{1-k^2}} \tag{47a}$$

When k is small this can be replaced by the approximate relation

$$\frac{\text{Width of pass band}}{\text{Resonant frequency of tuned circuits}} = k \tag{47b}$$

The error in this equation amounts to approximately k parts in unity, *i.e.*, 2 per cent error when $k = 0.02$, and so is sufficiently small in most practical cases to permit the use of Eq. (47*b*) in determining the coupling required to give a desired width of pass band. The width of the pass band is seen to be very nearly directly proportional to the coefficient of coupling and to the resonant frequency.

The nature of the response to different frequencies within the pass band can be readily estimated by comparing the secondary current at the center of the pass band, which is the frequency to which the circuits are tuned, with the secondary current at the frequencies given by Eq. (46) where the two coupling humps exist when the circuit resistance is zero. If the currents at these three frequencies are approximately equal the response throughout the pass band will be substantially constant. A manipulation of Eqs. (38) to (42) shows that this condition is approximately realized by satisfying the relation

$$\sqrt{Q_pQ_s} = \frac{1.5}{k} \tag{48}$$

The quantity $\sqrt{Q_pQ_s}$ can be considered as the effective Q of the combination and is the Q that is required when primary and secondary circuits are identical. This relation is not highly critical and can be departed from by as much as 10 per cent without destroying the band-pass characteristics. Where Q_p of the primary and Q_s of the secondary are the same, the Q required to give uniform response to the frequencies within the pass band is seen to be approximately inversely proportional to the coefficient of coupling and is hence approximately inversely proportional to the ratio of the width of pass band to the resonant frequency of the circuits.

When the condition for uniform response to the pass frequencies as given by Eq. (48) is realized it is found that the ratio which the voltage developed across the secondary condenser bears to the voltage applied in series with the primary is approximately

$$\frac{\text{Voltage across secondary condenser}}{\text{Voltage applied to primary}} = \frac{\sqrt{Q_pQ_s}}{2} = \frac{\text{effective } Q}{2} \tag{49}$$

It will be noted that the resonant rise of voltage is just half of that given by one circuit having the same effective Q and considered alone. This is the price that must be paid to secure the band-pass effect, and the return is a resonance curve having a substantially flat top, combined with sides somewhat steeper than those obtained using series resonance with a single circuit. Since the effective Q that must exist if the band-pass effect is to be realized is roughly inversely proportional to the width of the pass band, the resonant rise of voltage that can be realized from a pair of tuned circuits arranged to give band-pass characteristics is inversely proportional to the width of the pass band. This property of band-pass filters is brought out by the curves of Fig. 42*b*.

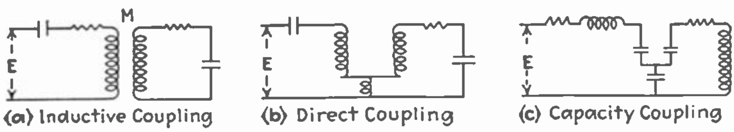


FIG. 43.—Various types of coupling employed in band-pass circuits.

The results discussed in connection with Fig. 42 and Eqs. (47), (48), and (49) are exactly the same irrespective of the method of coupling employed to give the coefficient k. The most commonly used methods are the inductive-, direct-, and capacitive-coupled arrangements shown in Fig. 43. The only differences in the performance of these circuits arise from variations in the coefficient of coupling as the common resonant frequency of the two tuned circuits is changed to alter the location of the pass band.

Variation of Band with Resonant Frequency.—Where a band-pass filter is used for tuning purposes instead of the usual simple series resonant circuit, the position of the pass band is ordinarily controlled by the use of identical primary and secondary variable condensers that are either geared together or mounted on a common shaft, so that the capacity of the two circuits remains identical as the common resonant frequency is changed. It would be possible to adjust the location of the pass band by varying the circuit inductances, but this is seldom done because of the difficulties involved in obtaining variable inductances having low radio-frequency resistance. When tuning is accomplished by varying the circuit capacity the coefficient of coupling is independent of the frequency to which the circuits are tuned if inductive or direct coupling is used, but is inversely proportional to the square of the frequency with capacitive coupling.[1] Band-pass filters that are tuned by variable condensers will therefore have a width of pass band that is approximately proportional to the resonant frequency of the circuits when inductive or direct coupling is used, and a band width approximately inversely proportional to the resonant frequency when capacitive coupling is employed. These conclusions follow directly from Eq. (47*b*), which shows that the width of the pass band is nearly proportional to the frequency at the center of the band and also to the coefficient of coupling.

If the width of the pass band is to be constant irrespective of the frequency at which the band is located it is necessary for the coefficient of coupling to be inversely proportional to the resonant frequency. Such a coefficient can be obtained by varying the coupling with the same control that adjusts the primary and secondary circuits and can be approximated by a combined inductive and capacitive coupling, such as shown at *e* and *f* in Fig. 41. The effective Q required to give uniform response to the different frequencies in the pass band is seen from Eq. (48) to be inversely proportional to the coefficient of coupling. When the pass band has a constant width the required Q will hence be proportional to the resonant frequency, which results in a resonant rise of voltage that is also inversely proportional to frequency. It is accordingly impossible to obtain a constant-width flat-topped pass band that has a resonant rise of voltage independent of the frequency at which the band is located. This is true irrespective of the type of coupling used or the method employed to change the location of the pass band.

It is absolutely essential that the primary and secondary circuits of a band-pass filter be tuned to exactly the same frequency, even though these circuits need not have the same ratio of inductance to capacity. If the two circuits are resonant to slightly different frequencies the result is to destroy the uniformity of transmission of the different frequencies within the pass band. This point was discussed in Sec. 17 and is well illustrated by Fig. 40, which shows how the flat-topped character of the response curve is completely lost if the two circuits are slightly detuned.

Uses of Band-pass Filters.—Band-pass filters are sometimes used for tuning in place of series resonant circuits and have the advantage of giving a uniform response to a band of frequencies, while discriminating very sharply against all frequencies outside this band. These are very important advantages when the signal to be received consists of a band of frequencies, as do radio-telephone signals. The disadvantages of the band-pass filter compared with the simple series resonant circuit, are: first, that the band-pass filter is much the more complicated, being in

[1] This is readily checked by observing that the circuit capacity varies inversely as the square of the resonant frequency, so that with a constant capacity of coupling condenser the definition of coefficient of coupling leads to the result stated.

effect two series circuits which must be adjusted to operate in *exact* synchronism; and second, that if the width of the pass band is appreciable the circuits must have a low Q in order to obtain uniform response in the pass band, and this results in a low resonance rise of voltage, even though the selectivity obtained is excellent. When very high selectivity is desired from a band-pass filter additional tuned circuits are used. Thus a third circuit can be coupled to the secondary, and still another to this third circuit if desired. The analysis of these more complicated band-pass filters is somewhat more complicated than the simple form that has been discussed and can most conveniently be carried out with the aid of hyperbolic functions. Without going into this analysis it may be said that the same general properties are found irrespective of the number of tuned circuits involved; that is to say, the width of the pass band depends on the coefficients of coupling while the uniformity of response within the pass band depends upon the circuit Q's.

Band-pass circuits show band-pass characteristics in connection with their parallel impedance. That is, the impedance which is offered to a voltage applied across the primary condenser of a circuit consisting of two circuits tuned to the same frequency and coupled together varies with frequency in very much the same way as does the current resulting from a voltage applied in series with the primary. This is because the series impedance of the primary varies in magnitude inversely as the primary current, while the parallel impedance of the primary varies inversely as the series impedance of the primary and therefore directly as the primary current. While interesting, these band-pass-impedance effects have found little or no practical application and therefore will not be discussed further.

21. The Analysis of Complex Circuits.—Complex networks made up of combinations of inductances, resistances, and condensers are most satisfactorily handled by being reduced to simpler equivalent networks. This simplification is carried out by adding impedances that are in series to obtain an equivalent impedance of the series combination, and adding the admittances of parallel impedances in order to obtain an equivalent admittance of the parallel combination. The procedure is exactly the same as used in the analysis of low-frequency alternating-current circuits and need not be considered in detail at this time.

The Δ-Y Transformation.—Bridge circuits of the type shown in Fig. 44*a* cannot be simplified by the usual method of combining impedances and admittances, and are most easily analyzed by replacing one of the Δ's of impedances in the circuit by an equivalent Y of impedances. Thus the impedance triangle $Z_1Z_2Z_3$ of Fig. 44*a*, which cannot be calculated by combining networks in series and parallel, can be replaced by the Y of impedances shown at Fig. 44*b* which can be made exactly equivalent to the original network as far as the terminals A, B, and C are concerned; that is, the currents and voltages at A, B, and C are the same with the equivalent Y of impedances as in the original network.[1]

The Y that is equivalent to a given Δ can be obtained by giving the impedances composing the Y the following values:

$$\left.\begin{aligned} Z' &= \frac{Z_1 Z_3}{Z_1 + Z_2 + Z_3} \\ Z'' &= \frac{Z_1 Z_2}{Z_1 + Z_2 + Z_3} \\ Z''' &= \frac{Z_2 Z_3}{Z_1 + Z_2 + Z_3} \end{aligned}\right\} \tag{50}$$

The notation is given in Fig. 44.

It is also possible to analyze complicated networks, including bridge circuits such as shown in Fig. 44a, by means of Kirchhoff's laws, which

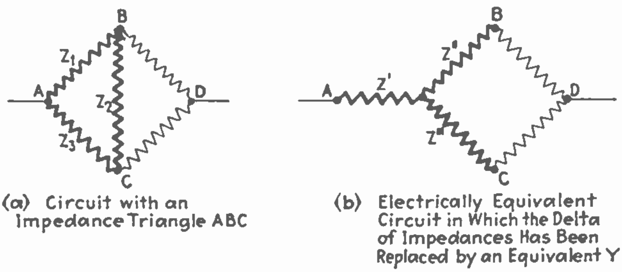


FIG. 44.—Circuits showing how an impedance triangle ABC can be replaced by an equivalent Y of impedances.

when applied to alternating-current circuits state: (1) the vector sum of all currents flowing into a junction is zero, and (2) the vector sum of the voltages acting around any closed loop is equal to the vector sum of the voltage drops around the loop. The number of independent equations that can be written according to law 1 is one less than the number of junctions because the equation for the last junction can be obtained by combining the other current equations. These current equations plus the voltage equations for the independent loops always equal the number of unknown currents, thus enabling these currents to be determined by a simultaneous solution. The use of Kirchhoff's laws has the disadvantage of requiring the simultaneous solution of a number of vector equations, each of which may be rather complicated, which makes the computations long and increases the likelihood of errors. It is ordinarily preferable to solve a network by combining impedances and admittances rather than by Kirchhoff's laws and this can always be done except in some circuits involving mutual inductance.

[1] This transformation was originated by A. E. Kennelly and published in his paper The Equivalence of Triangles and Three-pointed Stars in Conducting Networks. *Elec. World and Eng.*, vol. 34, p. 413, 1899.

Circuits Containing Two or More Mutual Inductances.—Circuits involving complex arrangements of mutual inductance must ordinarily be analyzed by means of Kirchhoff's laws. When these laws are applied to coupled circuits it must be remembered that every current flowing through an inductance induces a voltage in every other coupled inductance. Thus, in Fig. 45, I_1 induces voltages in circuits 2 and 3, I_2 induces voltages in circuits 1 and 3, and I_3 induces voltages in circuits 1 and 2. These induced voltages have a magnitude $-j\omega MI$ where M is the mutual inductance and I the current involved. The way in which

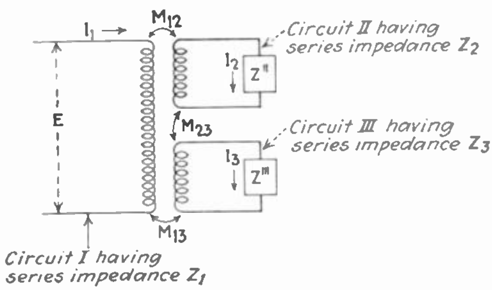


FIG. 45.—A complex coupled circuit involving three mutual inductances.

the analysis of complex circuits with mutual inductance is carried out can be seen from the following three equations that determine the behavior of the circuit of Fig. 45.

$$\begin{aligned} E - (j\omega M_{12}I_2 + j\omega M_{13}I_3) &= Z_1I_1 \\ (-j\omega M_{12}I_1 - j\omega M_{23}I_3) &= Z_2I_2 \\ (-j\omega M_{13}I_1 - j\omega M_{23}I_2) &= Z_3I_3 \end{aligned}$$

In each of these equations the term in parenthesis represents the induced voltages, and it will be noted that each equation states that the sum total of voltages acting in a circuit is equal to the voltage drop resulting from the circuit current flowing through the circuit impedance Z_1, Z_2, or Z_3, as the case may be. Simultaneous solution of these equations will give expressions for each of the three currents. The various mutual inductances M_{12}, M_{23}, and M_{13} involved in Fig. 45 may be either plus or minus depending upon the relative direction in which the turns are wound.

Complex circuits may frequently be analyzed qualitatively by inspection. Thus with the circuit of Fig. 46 there will be two frequencies at which the line current will be very great, namely, the frequencies at which one or the other branch is series resonant. In between these two frequencies the reactance of one branch is inductive while that of the other is capacitive, so that there is some frequency in this intermediate range where parallel resonance exists and the line current is extremely small.

When several voltages are simultaneously applied to a circuit the simplest procedure is to determine the currents and voltages resulting from each component of the applied potential just as though this component existed alone, and then to add the separate effects of the different

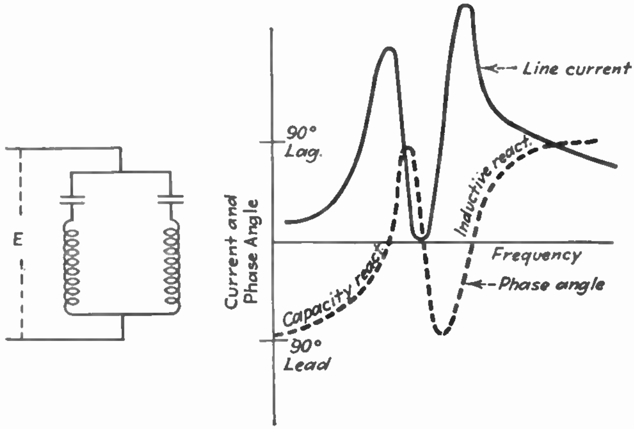


FIG. 46.—Complicated parallel circuit together with sketch showing the current entering the circuit when a constant voltage of varying frequency is applied.

parts of the applied voltage. This procedure can be followed both when the applied force is composed of components having different frequencies, and when there are several voltages of either the same or different frequencies applied at different points in the network.

CHAPTER IV

FUNDAMENTAL PROPERTIES OF VACUUM TUBES

22. Vacuum Tubes.—Vacuum tubes are used to generate the radio-frequency power required by a radio transmitter, to control the energy radiated, to amplify the weak radio-frequency signals obtained at the receiver, to rectify the signal, to amplify this rectified signal, and so on. The amplifying properties of vacuum tubes have also made possible the long-distance telephone, the talking picture, the modern phonograph, public-address systems, and many other things. Altogether it may be said that the vacuum tube is probably the most important single piece of equipment that has come into electrical engineering since the beginning of the century.

In its usual form the vacuum tube consists of a cathode capable of emitting electrons when heated, an anode (often called the plate) that attracts the electrons emitted from the cathode, and some means of controlling the flow of electrons from the cathode to the anode. These electrodes are enclosed in a gas-tight space that is evacuated to a degree which usually represents the highest vacuum that it is practicable to obtain, in order to permit the electrons to flow unimpeded from cathode to anode. There are in addition modified forms of vacuum tubes, such as those having only a cathode and anode, tubes with more than three electrodes, and tubes into which small quantities of gas have been intentionally introduced to produce certain operating characteristics.

Tubes are classified as diodes, triodes, tetrodes, pentodes, etc., according to whether there are two, three, four, five, etc., electrodes present. Thus a tube with only cathode and anode is a diode, while the addition of a control electrode (a grid) converts it into a triode.

23. Electrons and Ions.—Electrons can be considered as minute negatively charged particles that are constituents of all matter. They have a mass of 9×10^{-28} gram, which is $1/1840$ of that of a hydrogen atom, a charge of 1.59×10^{-19} coulomb, and are always identical irrespective of the source from which derived.[1] Atoms are composed of one or more such electrons associated with a much heavier nucleus that has a positive charge equal to the number of the negatively charged electrons contained in the atom, so that an atom with its full quota of

[1] Recent studies have shown that in addition to the usual properties of a moving charged body, electrons in motion possess wave characteristics, which however are not of practical importance as far as vacuum-tube technique is concerned.

electrons is electrically neutral. The differences between chemical elements arise from differences in the nucleus and in the number of associated electrons but not from variations in the character of electrons, which are always the same.

Positive ions represent atoms or molecules that have lost one or more electrons and so have become charged bodies having the weight of the atom or molecule concerned and a charge equal to that of the lost electrons. Unlike electrons, positive ions are not all alike and may differ in charge, or weight, or both. They are much heavier than electrons and resemble the molecule or atom from which derived. Ions are designated according to their origin, such as mercury ions, hydrogen ions, etc.

Electrons and ions are produced by separating the constituent parts of the atom or molecule in such a way as to produce molecules that are deficient in electrons and free electrons. There are a number of ways in which this separation may be accomplished. Thus in a gas when a swiftly moving ion or electron collides with a molecule, the impact may be sufficiently intense to knock one or more electrons out of the molecule, producing one or more free electrons and leaving a positive ion. This method of producing ions and electrons is known as *ionization by collision* and occurs in all vacuum tubes in which gas is present. Again if a solid body is sufficiently hot some of the electrons that make up the solid material will escape from its surface into the surrounding space, thus giving free electrons which are said to be obtained by *thermionic emission*. When ultra-violet light or x-rays strike a solid body or a gas, electrons will be emitted even at normal temperatures, and, with certain substances, notably potassium, caesium, and other alkaline earths, visible light will cause electrons to be emitted into the space surrounding the material. Electrons obtained in this way by the use of visible light are said to result from the *photoelectric effect*. Electrons can also be obtained from solid materials as a result of impact of quickly moving electrons or ions, which can knock electrons out of a solid body when striking with sufficient velocity. Electrons obtained in this way are said to result from *secondary electron emission* because it is necessary to have some primary source of electrons (or ions) before the secondary electron emission can be obtained. Finally it is possible to pull electrons directly out of solid substances by an intense electrostatic field at the surface of the material.

24. Motions of Electrons and Ions.—Electrons and ions are charged particles and so have forces exerted upon them by an electrostatic field in the same way that any other charged body does. The electrons, being negatively charged, tend to travel toward the positive or anode electrode while the positively charged ions travel in the other direction toward the negative or cathode electrode. The force exerted upon a charged particle by an electrostatic field is proportional to the product of the charge e of the particle and the voltage gradient G of the electrostatic field. Expressed in the form of an equation this relation is

$$\text{Force in dynes} = \left(\begin{array}{c}\text{gradient } G \text{ in} \\ \text{volts per centimeter}\end{array}\right)(\text{charge } e \text{ in coulombs})10^7 \quad (51)$$
$$= Ge \times 10^7$$

This force upon the ion or electron is in the direction of the electrostatic-flux lines at the point where the charge is located and is toward or away from the positive terminal depending on whether a negative or positive charge, respectively, is involved. The force which the field exerts on the charged particle causes an acceleration in the direction of the field at a rate that can be calculated by the ordinary laws of mechanics where the velocity does not approach that of light. That is

$$\text{Acceleration in centimeters per second per second} = \frac{\text{force in dynes}}{\text{weight in grams}} \quad (52)$$

It is to be noted that as long as there is an electrostatic field acting in the direction in which the electron or ion is moving that the velocity will increase because an acceleration will be produced as long as there is an electrostatic field present.

The velocity which an electron or ion acquires in being acted upon by an electrostatic field can be calculated in the usual way by the laws of mechanics. The amount of energy that a body with a charge of e coulombs gains in traveling between two points between which a difference of potential of V volts exists is equal to Ve joules, as can be readily proven by integrating Eq. (51). This energy is all converted into kinetic energy of motion, so that the velocity can be obtained from the formula

$$\left.\begin{array}{l}\text{Kinetic energy in ergs} = Ve10^7 = \tfrac{1}{2}mv^2 \\ \text{or} \\ v = \text{velocity in centimeters per second corre-} \\ \qquad \text{sponding to voltage } V = \sqrt{\dfrac{2Ve10^7}{m}}\end{array}\right\} \quad (53)$$

The velocity with which electrons and ions move in even moderate-strength fields is very great. Thus an electron in falling through a potential difference of 10 volts will gain a velocity of 1160 miles per second. These velocities are so great that from a practical point of view it is much simpler to express the velocity in terms of the difference of potential through which the electron (or ion) has fallen in gaining its speed. Thus when it is stated that an electron has a velocity of 10 volts one means the velocity that an electron would obtain in falling through a voltage drop of 10 volts. Since the velocity which a charged particle gains in falling through a difference of voltage is inversely proportional to the square root of the weight of the particle, the velocity represented by a given voltage will depend upon the weight of the charged body and will be much greater with electrons than ions, particularly the heavy ions. Thus a mercury ion having a charge equal to that of an electron will move less than one six-hundredth as fast as an electron in the same electrostatic field.

Effect of a Magnetic Field on a Moving Electron.—Since a moving charge represents an electrical current, an ion or electron in motion has all of the properties of an electrical current. Most important of these, so far as vacuum tubes are concerned, is the force which a magnetic field at right angles to the direction of current flow (*i.e.*, line of travel of electron or ion) exerts on a moving charge and which is exerted at right angles to both the magnetic field and the line of flow of current (*i.e.*, charge). Since the current represented by a moving charge is equal to the product of the charge and its velocity, a charge of Q coulombs moving with a velocity v represents a current Qv, and the force which is exerted upon this moving charge by a magnetic field having an intensity of H gausses in a direction at right angles to the motion of the charge is $HQv/10$ dynes. In the case of an electron which has a charge of e the force is accordingly

$$\text{Force in dynes} = \frac{Hev}{10} \tag{54}$$

An electron shot with high velocity into a magnetic field will follow a path similar to that shown in Fig. 47. The acceleration which the forces of the magnetic field produce on such an electron are always at right angles to the direction in which the electron is traveling and will result in a circular path when the field is uniform. The radius of this circle is smaller the greater the strength of the magnetic field, and the more slowly the electron is moving through the field. When electrons or ions are under the simultaneous influence of both electrostatic and electromagnetic fields, the resultant force that acts upon the moving charge at any instant is the vector sum of the two component forces.

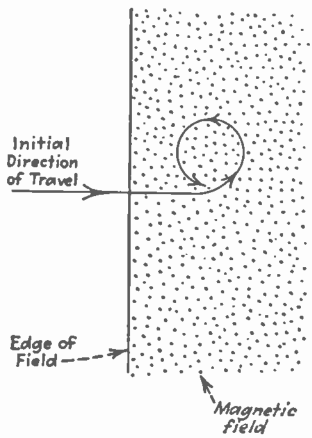


FIG. 47.—Path of electron that is projected with high velocity into a magnetic field. The moving electron is deflected in a direction that is at right angles to the magnetic field and at right angles to the direction in which the electron is traveling.

It can be shown that an electron or ion that is being accelerated will radiate a certain amount of energy in the form of electromagnetic waves. Thus when an electron moving at high velocity is suddenly stopped by impact against a metallic surface, x-rays of a wave length depending upon the velocity of the electron will be produced. An alternating acceleration, such as can be produced by an alternating voltage, will cause an electron to radiate electromagnetic waves of the type employed in radio communication. It is this property of electrons to radiate energy when accelerated that accounts for many of the properties of the Kennelly-Heaviside layer in the upper atmosphere.

25. Thermionic Emission of Electrons.—The electrical conductivity of metals is the result of electrons within the material that for the moment are not definitely attached to any particular molecule. These free electrons move about inside the conductor with a velocity that increases with temperature and exert a pressure just as does an ordinary gas. This pressure does not cause electrons to escape from the metal into the neighboring space, however, because there are attractive forces at the surface that tend to keep the electrons in the substance, and these forces are much greater in magnitude than the pressure of the "electron gas."

In order to escape from the surface of a conductor an electron must therefore do a certain amount of work to overcome the surface forces, and this energy can only be obtained from the kinetic energy possessed by the electron as a result of its motion. *Unless the kinetic energy of an electron exceeds the work that the electron must perform to overcome surface orces of the conductor the electron cannot escape.* For all known substances this energy which a free electron must have in order to be emitted from the material is so related to the kinetic energy possessed by the electrons that practically none escape at ordinary room temperatures. It is only at high temperatures, where the average kinetic energy possessed by the free electrons is large, that an appreciable number will have sufficient kinetic energy to escape through the surface of the material.

The process of electron emission from a solid substance is very similar to the evaporation of vapor from the surface of a liquid. In the case of the vapor the evaporated molecules represent molecules that obtained sufficient kinetic energy to overcome the restraining forces at the surface of the liquid, and the number of such molecules increases rapidly as the temperature is raised. The thermionic emission of electrons from hot bodies is seen to represent the same process and can be considered as an evaporation of electrons in which the energy the electron must give up in escaping corresponds to the latent heat of vaporization of a liquid.

The number of electrons evaporated per unit area of emitting surface is related to the absolute temperature T of the emitting material and a quantity b that is a measure of the work that an electron must perform in escaping through the surface, according to the equation:

$$I = AT^2\epsilon^{-\frac{b}{T}} \tag{55}$$

where I is the electron current in amperes per square centimeter, and A is a constant, the value of which may vary with the type of emitter. The temperature at which the electron current becomes appreciable is determined almost solely by the quantity b, which is accordingly the most important characteristic of an electron-emitting material. The emission is very sensitive to changes in b and to temperature, for these quantities appear in the exponent of Eq. (55), and slight variations in their value change the magnitude of the exponential term enormously. The value of A is therefore of secondary importance, for the effects of wide variations in A can be compensated for by minute temperature changes. The relationship between electron emission per unit area of emitter and the emitter temperature is given in Fig. 48 for several typical electron-emitting materials commonly used in vacuum tubes and shows

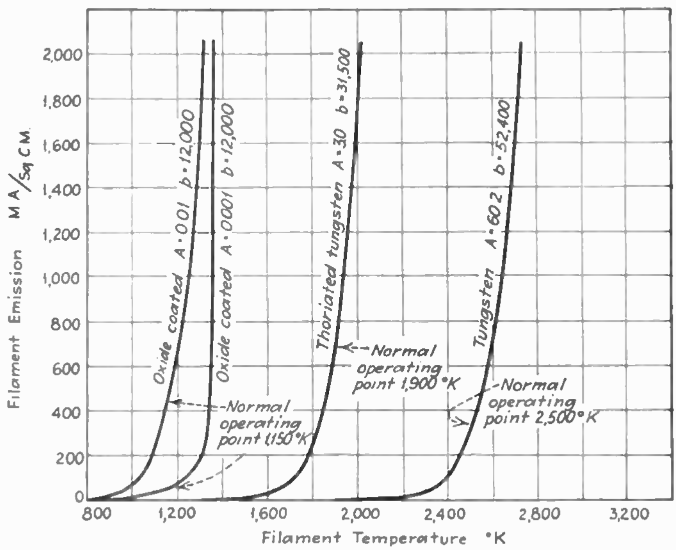


FIG. 48.—Variation of electron emission with absolute temperature for the three principal types of electron emitters, showing the extreme sensitiveness of the emission to temperature and to the value of the constant b.

how extremely sensitive electron emission is to temperature and to the value of the work function b, while being much less dependent on A.

26. Electron-emitting Materials.[1]—The properties of matter are such that thermionic emission of electrons does not become appreciable until temperatures in the order of 700 to 2500°K. are reached, the exact value depending upon the material involved. The high temperatures required limit the number of substances suitable for use as thermionic electron emitters to a very few, of which tungsten, thoriated-tungsten,

[1] This section is a very brief summary of the principal characteristics of the types of electron emitters employed in commercial vacuum tubes. For further information on electron emitters the reader should consult Saul Dushman, Thermionic Emission, *Reviews of Modern Physics*, vol. 2, p. 381, October, 1930. This article is a summary of what is known about thermionic emission of electrons and is written by a noted investigator in this field.

and oxide-coated emitters are the only ones commonly used in vacuum tubes.

Tungsten.—While a relatively poor emitter, tungsten can be operated at temperatures so high as to give good emission in spite of the large value of work function b which it possesses. Tungsten is extensively used as the electron-emitting cathode of high-power vacuum tubes because of its durability under the exacting conditions encountered in such tubes. The essential characteristics of tungsten emitters are shown in Table IV.

Thoriated Tungsten.—Thoriated-tungsten emitters have an electron emission that is many thousand times that of pure tungsten operated at the same temperature, and begin to emit an appreciable quantity at temperatures 500° to 600°K. lower than does pure tungsten. This increased electron emission is the result of a layer of thorium one molecule deep that forms on the surface of the thoriated tungsten and reduces the work that an electron must do to escape.[1] Thoriated-tungsten emitters consist of tungsten containing a reducing agent (ordinarily carbon), and a small quantity (1 to 2 per cent) of thorium oxide. Such cathodes must be activated by heating the emitter to approximately 2600° to 2800°K. for one or two minutes, which is called flashing, and then glowing for some minutes at an activating temperature of 2100° to 2200°K. The flashing raises the temperature to the point where the impregnated carbon reduces some of the thorium oxide to metallic thorium, and the subsequent treatment at the activating temperature allows this thorium to diffuse to the surface where it forms a layer one molecule deep that is the seat of the high electron emission. During operation of a thoriated-tungsten emitter, thorium is being continuously evaporated from the surface layer but is replenished by diffusion from the interior of the tungsten. The important characteristics of thoriated-tungsten emitters are given in Table IV.

Oxide-coated Emitter.—The oxide-coated emitter consists of a mixture of barium and strontium oxides coated on the surface of a suitable metal, such as platinum or a platinum alloy, and when suitably activated will emit large numbers of electrons at temperatures in the order of 1150°K. The electron emission of emitters of the oxide-coated type appears to arise from a layer of alkaline-earth metal, *i.e.*, metallic barium and strontium, that forms on the surface of the oxide coating, and is a maximum when this layer covers the entire surface of the oxide to a depth of approximately one molecule.[2] During operation of the emitter at its normal working temperature the surface metal that is evaporated is

[1] For further information relative to thoriated-tungsten emitters the reader is referred to Irving Langmuir, The Electron Emission from Thoriated-tungsten Filaments, *Phys. Rev.*, vol. 22, p. 357, 1923.

[2] A detailed discussion of the properties of oxide-coated emitters is to be found in Joseph A. Becker, Phenomena in Oxide-coated Filaments, *Phys. Rev.*, vol. 34, p. 1323, 1929.

replenished by diffusion of additional molecules from the interior of the oxide coating.

Oxide-coated emitters when first formed must be activated (or "formed") before the full thermionic activity is realized. This activation consists in glowing the emitter for several minutes at a temperature of approximately 1500°K. and then applying a strong positive electrostatic potential gradient for a period of from 2 to 30 min. During this period the electron emission increases, and when it has reached a rather large value the electrostatic field and emitter temperature are readjusted to lower values, which are maintained for an additional period. The exact procedure for activating oxide-coated emitters is in the nature of an art that varies greatly under different circumstances but which in general follows the broad outline that has been given. The object of the activation process is to break down a fraction of the oxide coating into the alkaline-earth metal and to use this metal to build up a layer of metallic molecules on the surface of the oxide. The high temperature which is used at the start of the activation process causes some of the oxide coating to dissociate into ions, and the positive or metal ions so produced migrate inward when electrons are drawn from the emitter and are transformed into metal molecules at the interface between core and oxide. These molecules then diffuse to the surface through the porous oxide coating. The increase in electron emission that takes place during the forming process is caused by the gradual building up of the surface layer of alkaline-earth metal. The high thermionic activity of oxide-coated cathodes appears to be caused by this surface layer becoming positively charged as a result of losing electrons and then acting as a positively charged screen or grid that covers the surface of the oxide coating and helps pull electrons from its surface.[1] The essential characteristics of properly formed oxide-coated emitters are to be found in Table IV.

TABLE IV.—CHARACTERISTICS OF ELECTRON EMITTERS*

Type of emitter	Constants in equation $i = AT^2\epsilon^{-\frac{b}{T}}$		Normal operating temperature, degrees Kelvin	Efficiency at operating temperature in milliamperes emission per watt of heating power
	A	b		
Tungsten	60.2	52,400	2450 to 2600	3 to 15
Thoriated tungsten	3.0	31,500	1900	62.5
Oxide coated	0.01 to 0.001	12,000	1100 to 1170	50 to 125

* Most of the data in this table are taken from the paper by Saul Dushman, Thermionic Emission, *Reviews of Modern Physics*, vol. 2, p. 381, October, 1930. This article also contains data on many other types of emitters which have been studied by physicists but have not found commercial use.

[1] An excellent and easily readable discussion of the mechanism of electron emission in oxide-coated cathodes is to be found in J. A. Becker, The Role of Barium in Vacuum Tubes, *Bell Laboratories Rec.*, vol. 9, p. 54, October, 1930.

Power Required to Heat Emitter.—The electron-emitting cathodes of vacuum tubes are heated electrically, either by forming the emitter into a filament that is raised to the necessary temperature by the passage through it of a suitable current, or by using a cylindrical cathode that is heated either by conduction or radiation from an internal heater con-

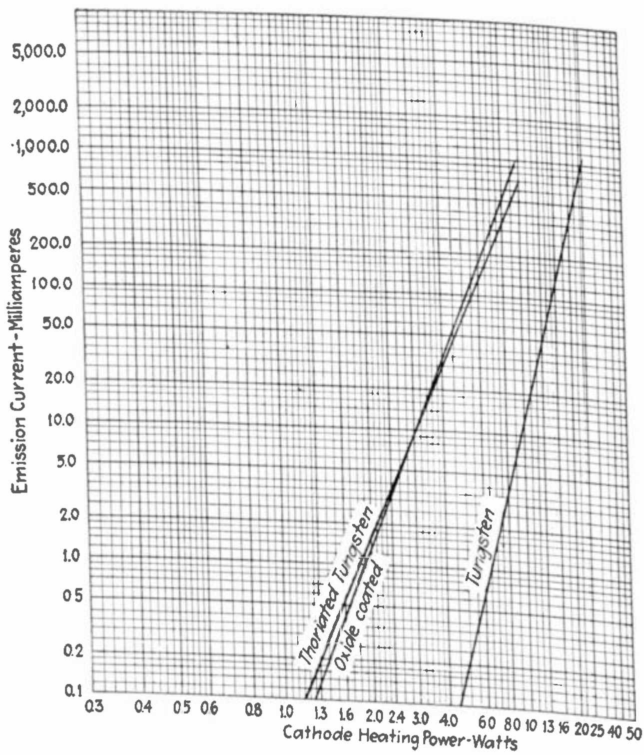


FIG. 49a.—Relation between cathode heating power and electron emission for three representative cathodes, plotted on power-emission paper. These curves are straight lines and so can be drawn when only three points are known, and can be extrapolated as far as desired.

sisting of an incandescent tungsten filament. Filament cathodes may be of the tungsten, thoriated-tungsten or oxide-coated type, while heater-type cathodes always employ an oxide-coated emitter because of the impossibility of obtaining by indirect heating the high temperatures required by other emitters.

Most of the power required to maintain the cathode at a given temperature represents energy sent out from the cathode in the form of radiant heat. A small amount of power is lost by heat energy being conducted away from the cathode along the support and lead wires, but other than this the loss of heat by conduction is negligible because the cathode is in a very high vacuum. When an object is at a temperature considerably higher than surrounding objects the heat energy radiated is proportional to the fourth power of the absolute temperature, and so is given by the relation

$$\text{Radiated energy per square centimeter of surface} = KT^4 \qquad (56)$$

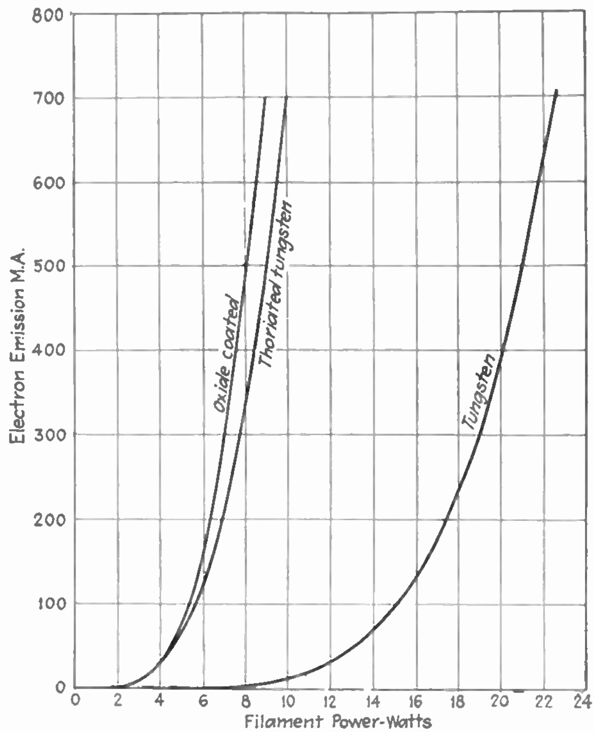


FIG. 49b.—Relation between cathode heating power and electron emission for three representative cathodes, plotted in rectangular coordinates. The very sudden rise in these curves as contrasted with the same data plotted on Fig. 49a emphasizes the usefulness of the power-emission paper.

where T is the temperature in degrees Kelvin; and K is a constant, the exact value of which depends on the surface conditions. As a consequence of this relation the electrical power required to heat the cathode increases very rapidly with temperature, and emitters such as tungsten, that must operate at high temperature, require considerably more heating power in proportion to electron emission than do the low-temperature emitters of the oxide-coated type. These differences are demonstrated by the column of milliamperes electron emission per watt of cathode heating power given in Table IV for different emitters.

Equation (56) can be substituted in Eq. (55) to give an equation relating electron emission to the cathode heating power. From this equation the curvilinear system of coordinates shown in Fig. 49a can be derived on which the electron emission plots as a straight-line function of the heating power. Coordinate paper of this type, which is known as power-emission paper, is very useful in the comparison of different cathodes because the cathode heating power required to give three values of electron emission, such as 0.1, 1.0, and 10 milliamp., suffices to determine the straight line on the power-emission paper and to permit accurate extrapolation to emissions that could not be maintained for even a few seconds without damage to the tube. Power-emission curves for the cathodes of Fig. 48 are shown on power-emission paper in Fig. 49a, and when contrasted with similar curves on rectangular coordinates as shown in Fig. 49b, illustrate the usefulness of the curvilinear coordinate system in straightening out the plotted emission characteristic.

Operating Temperature and Life.—The temperature at which tungsten emitters are operated is a compromise between high electron emission per watt of heating power, which means a high operating temperature and a short life; and a low emission per watt of heating power, which goes with a low operating temperature and a long life. The actual temperature selected is the highest value that will give a life expectation of 1000 to 2000 hr. The life is determined by the rate at which tungsten evaporates from the surface of the filament, and, since a given depth of evaporation is less important when the filament is thick, large-diameter filaments can be operated at slightly higher temperatures than thin filaments and so give greater electron emission per watt of heating power.

The life of thoriated-tungsten filaments is determined by the supply of thorium in the tungsten, and by the rate at which the thorium is evaporated from the surface layer. Filaments of this type never burn out, as do those of tungsten, except as a result of an accident. The proper operating temperature for thoriated tungsten is very close to 1900°K., at which the total life expectancy is at least several thousand hours of service. If the operation is at a temperature much above 2000°K. the thorium evaporates from the surface layer faster than this layer is replenished by diffusion of thorium from the interior of the tungsten, with the result that the surface is soon denuded of its thorium layer, and the electron emission becomes that of pure tungsten. A thoriated-tungsten filament that has been in service for many hours will begin to show low electron emission as a result of exhaustion of the metallic thorium produced in the tungsten at the time of activation. The original electron emission can be restored by flashing the filament for 20 to 30 sec. at three to four times normal voltage, and then burning for 30 to 60 min. at an over voltage of 25 to 40 per cent. This process, which is often termed rejuvenation, reactivates the filament by reducing additional thorium oxide to metallic thorium and reforming the surface layer of thorium, and can be repeated until the supply of thorium oxide is exhausted.

The life of oxide-coated cathodes is limited by the supply of active electron-emitting material in the cathode. The life of oxide-coated cathodes is very great, often in excess of 5000 hr., and many tubes with oxide-coated cathodes have been operated continuously for more than three years before the electron emission became seriously reduced. Oxide-coated filaments do not require occasional rejuvenation because the activation process of electrolysis and diffusion goes on in them continuously during use. The operating temperature is a compromise between life and efficiency of emission (*i.e.*, electron emission per watt of heating power) and is not as critical as in the case of tungsten and thoriated-tungsten cathodes.

Uses of Different Emitters.—Pure tungsten, thoriated-tungsten, and oxide-coated cathodes each have a special field of usefulness. Oxide-coated emitters have longer life and at least as great emission per watt of heating power as the other types, but their high-emission property is destroyed by bombardment from high-velocity positive ions. This prevents the use of oxide-coated cathodes in large tubes which are operated at high plate voltages and limits their field of application to low power tubes, where they are superior to other types. Thoriated tungsten cathodes find their most important use in moderate power tubes, such as those intended to deliver a few hundred watts when operating as an oscillator. For such tubes thoriated-tungsten cathodes will give somewhat better service than those of the oxide-coated type. Pure tungsten cathodes are used only where the other, more efficient types of emitters are unsatisfactory, which is in the very large tubes operating at high plate voltages. The requirements under these conditions are very severe because the high velocity with which positive ions produced from residual gas strike the cathode causes rapid disintegration of oxide-coated cathodes and strips off the surface layer of thorium from thoriated-tungsten filaments much faster than it can be replaced by diffusion of thorium to the surface.

Velocity of Emission.—The electrons emitted from a hot cathode come out with a velocity that represents the difference between the kinetic energy possessed by the electron just before emission and the energy that must be given up to escape. Since the energy of the different electrons within the emitter is not the same, the velocity of emission will be different for different electrons, and while usually much less than one volt, there are a few electrons emitted at much higher speeds. Experiment shows that the velocities of emission are distributed according to Maxwell's law for the distribution of velocities in a gas composed of electrons and having the temperature of the emitting cathode.[1] The

[1] See L. H. Germer, The Distribution of Initial Velocities among Thermionic Electrons, *Phys. Rev.*, vol. 25, p. 795, 1925.

average velocity of emission accordingly increases with the cathode temperature, just as does the average velocity of gas molecules.

27. Current Flow in a Two-electrode Tube—Space-charge Effects.—When an electron-emitting cathode is surrounded by a positive anode (*i.e.*, plate electrode) to form a two-electrode vacuum tube (or diode), the relation between the plate current (*i.e.*, the number of emitted electrons that are attracted to the anode) and the plate potential has the character shown in Fig. 50, which gives the results obtained in a typical tube for several cathode temperatures. It is seen that at high anode

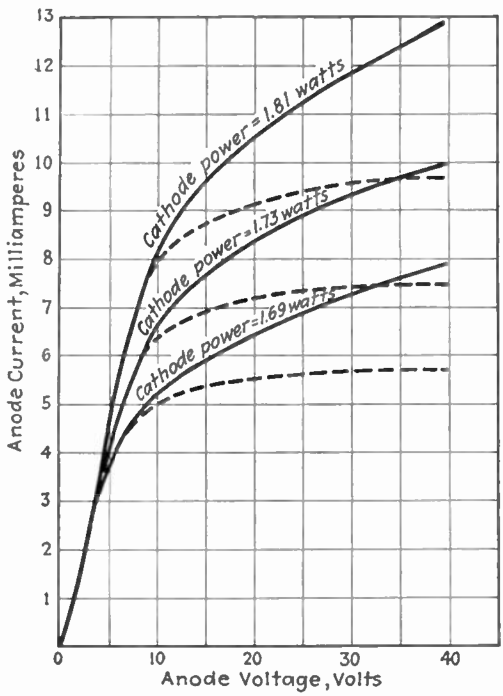


FIG. 50.—Anode current as a function of anode voltage in a two-electrode tube for three cathode temperatures. The solid lines are the characteristics actually obtained using an oxide-coated cathode, while the dotted lines show the type of curve that is given by tungsten and thoriated-tungsten cathodes.

voltages the electron current is largely independent of anode voltage, being determined primarily by the cathode temperature, while at low anode voltages the current is controlled by the anode voltage and is independent of cathode temperature. When the plate is negative it repels electrons and the plate current is then zero.

The behavior observed at high anode voltages is a result of the fact that a high anode potential draws the electrons away from the filament as fast as they are emitted, which makes the anode current equal the total electron emission from the cathode. Under these conditions the anode current is given by Eq. (55), and the tube is said to be operating at voltage saturation.[1]

Space-charge Effects.—At low plate (*i.e.*, anode) voltages the anode current is limited by the repelling effect which the negative electrons already in the space between anode and cathode have on the electrons just being emitted from the cathode. The electrons in the interelectrode space constitute a negative space charge (*i.e.*, a negative charge distributed in space) and at any instant the number of electrons that are in transit between electrodes cannot exceed the number that will produce a negative space charge that completely neutralizes the attraction which the positive plate exerts upon the electrons just leaving the cathode. All electrons in excess of the number necessary to neutralize the effect of the plate voltage are repelled back into the cathode by the negative space charge of the electrons in transit, so that the anode current will be independent of the electron-emitting power of the cathode, provided the cathode is capable of emitting enough electrons to produce a full space charge. When a full space charge is present the plate current depends upon the plate voltage, since with higher voltages the electrons travel from cathode to anode more rapidly, making the rate of arrival greater in proportion to the total number in the space between anode and cathode at any instant. Increasing the plate voltage thus causes the electron flow to increase until a point is reached where the total electron emission of the cathode is being drawn to the plate, after which further increases in voltage will produce practically no additional current because of voltage saturation. The anode current for a given anode voltage has its greatest possible value when the current flow is limited by the space charge, and will be less than the maximum if voltage saturation is present. On the other hand with a given cathode temperature the maximum current is obtained when the plate potential is sufficiently high to give voltage saturation with the electron emission present.

The energy that is delivered to the tube by the source of anode voltage is first expended in accelerating the electrons traveling from cathode to anode and so is converted into kinetic energy. When these swiftly moving electrons strike the anode this kinetic energy is then transformed into heat as a result of the impact and appears at the anode in the form of heat that must be radiated to the walls of the tube.

When the anode current is limited by space charge the negative charge of the electrons in transit between cathode and plate will be sufficient to give the space in the immediate vicinity of the cathode a

[1] The sharpness with which voltage-saturation effects appear differs greatly with the type of emitter. Thus the anode current with cathodes of tungsten or thoriated tungsten have a characteristic such as shown by the dotted lines in Fig. 50, in which the saturation effect is almost complete, while in emitters of the oxide-coated type, the saturation effect takes place more gradually, as shown by the solid lines of Fig. 50.

slight negative potential with respect to the cathode. The electrons emitted from the cathode are projected out into this negative field with an emission velocity which will vary with different electrons. The negative field next to the cathode causes the emitted electrons to slow down as they move away from the cathode, and those having a low velocity of emission are driven back into the cathode. Only those electrons having the highest velocities of emission will be sent out with sufficient force to penetrate through the negative field near the cathode and reach the region where they are drawn toward the positive plate. The remainder, that is those electrons having low emission velocities, will be brought to a stop by the negative field adjacent to the cathode and will fall back into the cathode.

Anode Current When Limited by Space Charge.—When limited by space charge the anode current received from any portion of the cathode is proportional to the $3/2$ power of the voltage between the plate and that part of the cathode contributing the current. In heater-type cathodes where the cathode is an equipotential surface the total plate current is consequently given by the equation

$$\text{Plate current} = KE_p^{3/2} \tag{57}$$

where K is a constant determined by the geometry of the tube, and E_p is the anode (plate) voltage with respect to the cathode.[1,2] In filament-type tubes the voltage drop produced in the cathode by the heating current causes different parts of the filament to have different potentials with respect to the anode, so that while the current from each part of the filament is proportional to the $3/2$ power of the voltage with respect to that part, the total current is not proportional to the $3/2$ power of the potential of the anode. In filament-type tubes it is customary to refer the plate voltage to the negative side of the filament so that the plate is considered to be at a potential which, with respect to most of the cathode, is less than the anode voltage as measured with respect to the negative end of the filament. The result is that when limited by space charge the total anode current varies at a power of the anode voltage, as measured with respect to the negative side of the filament, that is greater than the $3/2$ power, but which approaches the $3/2$ power as the

[1] When the anode voltage is low and when very precise results are to be obtained the voltage E_p appearing in Eq. (57) must be interpreted to mean the actual anode voltage plus a correction to take into account the contact potential existing between plate and cathode and also the effective velocity of emission of the electrons. Each of these corrections ordinarily amounts to less than one volt and can be neglected where the anode voltage is moderately high unless precision results are desired.

[2] The equations giving the anode current when limited by space charge have been worked out for various structures by Langmuir, Child, Schottky, and others, and have been found to differ only in the value of the constant K appearing in Eq. (57).

voltage drop in the cathode becomes small compared with the anode voltage.[1]

28. Action of the Grid.—The grid is a screen-like electrode placed between the cathode and plate for the purpose of controlling the flow of electrons to the plate. The grid is normally operated at a negative potential with respect to the cathode and so attracts no electrons, but the extent to which it is negative affects the electrostatic field in the vicinity of the cathode and so controls the number of electrons that pass between the grid wires and on to the plate.

The grid may consist of any type of open-mesh structure that provides holes of ample size for the passage of electrons, and which at the same time has an influence on the electrostatic field near the cathode. The most common form of grid is a coil of fine wire with widely spaced turns, as shown in Figs. 51*a* and 51*b*. The cross section of this coil is circular when a straight filament cathode or a heater cathode is used, but is usually oval when a filament in the form of a V or W is used. Several other types of grid construction in common use are also shown in Fig. 51. All of these forms of grid construction function in the same way, and the choice between them is determined by such design considerations as mechanical rigidity, cost of construction, etc.

The grid electrode controls the flow of electrons to the plate because the electrostatic field between the plate and cathode, and particularly near the cathode, is affected by the grid potential. This is shown in

[1] The way in which the total anode current varies with anode voltage when the potential drop in the filament cathode is taken into account can be readily worked out by setting up a simple differential equation based on Eq. (57). The results of such a solution give the following:

Case 1. Anode voltage less than voltage drop E_f in filament:

$$i = K\frac{2}{5E_f}E_p^{5/2} \tag{58a}$$

Case 2. Anode voltage greater than voltage drop E_f in filament:

$$\left.\begin{aligned} i &= K\frac{2}{5E_f}\left[E_p^{5/2} - (E_p - E_f)^{5/2}\right] \\ &= KE_p^{3/2}\left[1 - \frac{3}{4}\frac{E_f}{E_p} + \frac{3}{24}\left(\frac{E_f}{E_p}\right)^2 \cdots\right] \end{aligned}\right\} \tag{58b}$$

This series converges so rapidly that when the plate voltage exceeds twice the filament drop, one can with close approximation write

$$i = KE^{3/2}\left[1 - \frac{3}{4}\frac{E_f}{E_p}\right] \tag{58c}$$

The effect of voltage drop in the cathode has the effect of causing the plate current to increase faster than the 3/2 power of the voltage but at a rate that never exceeds the 5/2 power. The departure from the 3/2 power becomes less as the ratio E_f/E_p is reduced.

Fig. 52, which gives the electrostatic field that exists between plate and cathode for several values of grid voltage (no space charge present). When the grid is at zero potential with respect to the cathode the positive potential of the plate produces a stray electrostatic field near the cathode, which, while somewhat weaker than would be the case with the grid removed, is still not zero because the grid is not a perfect shield. As the grid is made negative it produces an electrostatic field between cathode and grid that opposes the stray field produced by the plate potential, and thereby weakens the electrostatic field in the vicinity of the cathode,

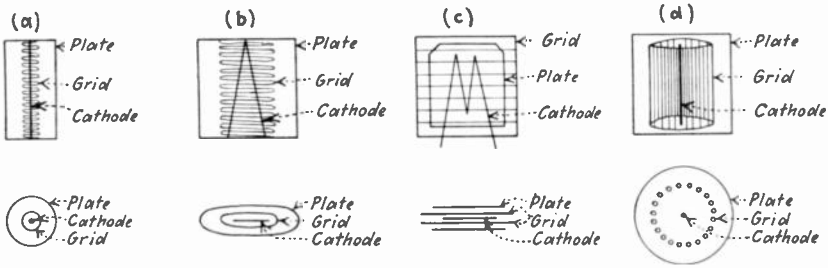


FIG. 51.—Grid, plate, and cathode structures of a number of typical tubes. It will be observed that in every case the grid is a screen-like electrode that affects the electrostatic field near the cathode while permitting electrons to flow to the plate.

as shown at *b* in Fig. 52. When the grid is made sufficiently negative the stray electrostatic field produced between the cathode and grid by the positive anode is entirely neutralized by the negative grid, as shown at *c* in Fig. 52. In this last case there is no electrostatic field to draw the emitted electrons away from the cathode, and the space current will be zero.

The number of electrons that reach the anode is determined almost solely by the electrostatic field near the cathode and is affected hardly at all by the field in the rest of the interelectrode space. This is because the electrons near the cathode are moving very slowly compared with the electrons that have traveled some distance toward the plate, with the result that the volume density of electrons in proportion to the rate of flow is large near the cathode and low in the remainder of the interelectrode space. The total space charge of the electrons in transit toward the plate is therefore made up almost solely by the electrons in the immediate vicinity of the cathode, and once an electron has traveled beyond this region it reaches the plate so quickly as to contribute to the space charge for only a brief additional time interval. The result is that the space current in a three-electrode vacuum tube is for all practical purposes determined by the electrostatic field which the combined action of the grid and plate potentials produce near the cathode.

When the grid structure is symmetrical it can be shown from the theory of electrostatics that the field at the surface of the cathode is proportional to the quantity $\left(E_g + \frac{E_p}{\mu}\right)$, where E_g and E_p are the grid and anode (plate) voltages, respectively, with respect to the cathode, and μ is a constant that is determined by the geometry of the tube and is independent of the grid and plate voltages.[1] The constant μ is known as the *amplification factor* of the tube and is a measure of the relative effectiveness of grid and plate voltages in producing electrostatic fields at the surface of the cathode.

Quantitative Effect of Grid Potential on Space Current.—The space current in a three-electrode tube varies with $\left(E_g + \frac{E_p}{\mu}\right)$ in exactly the same way that the space current in a two-electrode tube varies with the plate voltage, since in both cases the current flow is determined by the electrostatic field near the cathode, and this electrostatic field is in turn proportional to $\left(E_g + \frac{E_p}{\mu}\right)$ when a grid is present, and to E_p when the only electrode is a plate. With no voltage drop in the cathode the space current is therefore proportional to $\left(E_g + \frac{E_p}{\mu}\right)^{3/2}$, and when the grid is negative all of this current goes to the plate, so that

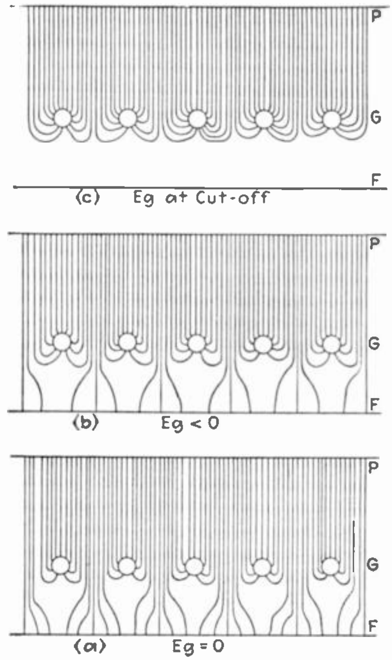


FIG. 52.—Electrostatic field produced between plate and cathode with different grid potentials, showing how the electrostatic field in the vicinity of the cathode can be controlled by the potential of the grid. These curves take into account only those fields produced by the electrode potentials, and do not include the field developed by the space charge of electrons which is superimposed upon the fields shown.

$$\text{Plate current} = K\left(E_g + \frac{E_p}{\mu}\right)^{3/2} \tag{59}$$

where K is a constant determined by the tube dimensions. It will be noted that this equation is analogous in all respects to Eq. (57) and that

[1] The quantity $\left(E_g + \frac{E_p}{\mu}\right)$ represents the grid voltage that will produce the same electrostatic field at the surface of the cathode when the plate is at zero potential as is actually produced by the combined action of the plate and grid potentials E_p and E_g and so can be considered as the effective anode voltage. The combined effect of anode and grid voltages is also the same as though the grid were the only anode and was at a potential of $\frac{(E_p + \mu E_g)}{1 + \mu}$.

by interpreting $\left(E_g + \frac{E_p}{\mu}\right)$ to be the effective anode voltage they are identical. In filament-type tubes there is a voltage drop in the cathode so that different parts of the cathode are at different potentials with respect to the grid and plate. The result is that with filament-type cathodes $\left(E_g + \frac{E_p}{\mu}\right)$ must be corrected exactly as Eq. (57) was modified to take into account the effect of the voltage drop in the filament of two-electrode tubes. It is customary to measure the grid and plate

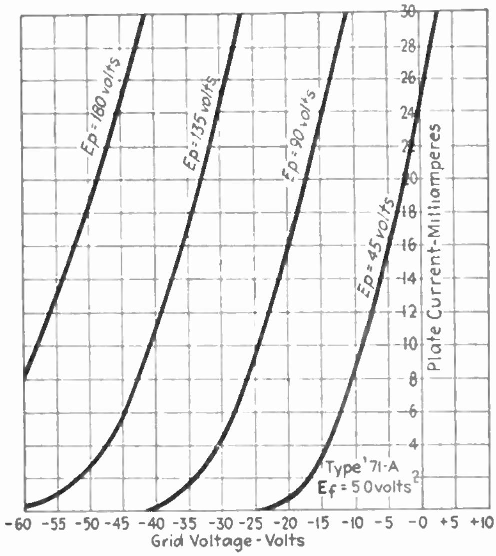


FIG. 53.—Relationship between grid voltage and plate current for several values of plate voltage in a typical three-electrode tube. Note that the only effect of changing the plate voltage is to displace the curves without changing the shape.

potentials with respect to the potential of the negative side of the filament, and in terms of this notation the space current of three-electrode tubes with filament-type cathodes is given by Eq. (58) provided E_f and E_p in Eq. (58) are replaced by $E_f(1 + 1/\mu)$ and $(E_g + E_p/\mu)$ respectively.

A grid maintained negative with respect to all parts of the cathode draws no electrons and so controls the plate current without consuming any power. It is this property which gives the three-electrode vacuum tube the power to amplify and to generate oscillations. If the grid of the vacuum tube is allowed to go positive it attracts large numbers of electrons and consumes a considerable amount of power, with the result that when the grid is positive the ratio of energy controlled in the plate circuit to energy consumed at the grid is small, and little if any amplification results.[1]

[1] The plate current appearing in Eq. (59) represents the total space current, which is the sum of the grid and plate currents. When the grid is negative, the space current

29. Characteristic Curves of Triodes.—The most important characteristics of vacuum tubes with grid, plate, and cathode electrodes are the relationships between: (1) plate current and plate voltage with constant grid voltage, and (2) plate current and grid voltage with constant plate voltage. Examples of such curves are shown in Figs. 53, 54, and 55. It is to be noted that the curves showing plate current as a function of grid voltage have the same shape as those showing plate current with varying plate voltage, the only difference being in the scales involved and in the location of the curves with reference to the axes. Furthermore the individual curves in the family showing plate current as a

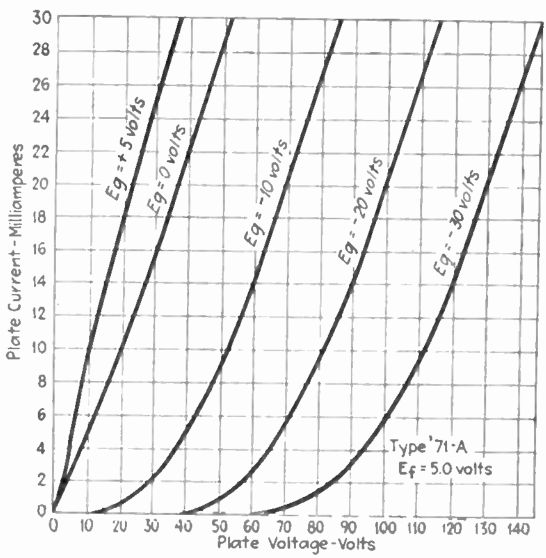


FIG. 54.—Relationship between plate voltage and plate current for several values of grid voltage for the same tube as in Fig. 53. Note that the only effect of changing the grid voltage (provided the grid is at least slightly negative) is to displace the curves without changing the shape, and that these curves have the same shape as those of Fig. 53.

function of grid voltage for different values of plate voltage are of the same shape, differing only in that the curves for different plate voltages are displaced along the grid-voltage axis. The same is true of the plate current-plate voltage family, where the only effect of changing the grid voltage is to displace the curve along the plate-voltage axis. It will also be noted that all of these curves have the same general shape as the plate voltage-plate current curves of the two-electrode tube.

These various properties of the characteristic curves of three-electrode vacuum tubes result from the fact that the plate current is the same

is then the plate current; but when the grid is positive, the current given by Eq. (59) represents the sum of the grid and plate currents.

function of $(E_g + E_p/\mu)$ that the plate current of a two-electrode tube is of plate voltage. This means that the plate current is determined only by the value $(E_g + E_p/\mu)$ and not by the particular combination of grid and plate voltages involved. A voltage of ΔE_g added to the grid potential therefore produces the same plate-current increment as would be produced by an increment in the plate potential of $\mu \Delta E_g$, and the effect of a grid-voltage increment ΔE_g can be neutralized by a plate-

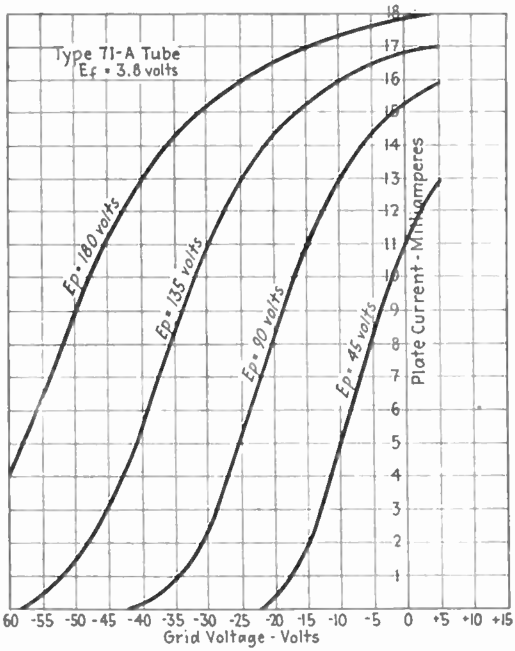


FIG. 55.—Grid-voltage plate-current curves differing from those of Fig. 53 only in that the cathode temperature has been lowered to the point where voltage saturation begins to appear at the larger plate currents, causing the tops of the curves to bend over.

potential increment of $-\mu \Delta E_g$. In a similar manner it can be shown that changing the plate voltage ΔE_p has the same effect on the plate current as an increment $\Delta E_p/\mu$ in the grid voltage.

The range covered by Figs. 53 and 54 lies in the region where the anode current is limited by space charge. Under these circumstances the space current varies as a power of $\left(E_g + \dfrac{E_p}{\mu}\right)$ that ranges from $3/2$ when the effective anode voltage is large compared with $\left(1 + \dfrac{1}{\mu}\right)$ times the voltage drop in the filament, to $5/2$ when the effective anode voltage is less than $\left(1 + \dfrac{1}{\mu}\right)$ times the voltage drop in the filament. Figure 55 shows the situation that exists when the electron emission is sufficiently low to bring in voltage saturation. It is seen that the anode current is still a function of $\left(E_g + \frac{E_p}{\mu}\right)$ exactly as in Fig. 54, but the shape of the curves is now different as a result of voltage saturation. The curves of Figs. 53 and 54 would show similar saturation effects if extended to higher values of I_p.

The plate current of a three-element vacuum tube becomes zero when the grid is at a sufficiently negative potential to just neutralize the stray electrostatic field produced at the surface of the cathode by the positive plate acting through the meshes of the grid. The grid potential at which this condition is realized is known as the "cut-off" grid potential and is equal to $-E_p/\mu$ volts, *i.e.*, it is the potential which makes $\left(E_g + \frac{E_p}{\mu}\right)$ equal to zero.

30. Vacuum-tube Constants.—The principal operating characteristics of a vacuum tube can be expressed in terms of the amplification factor μ, the dynamic plate resistance R_p (generally called simply the plate resistance), and the mutual conductance G_m. A knowledge of these permits a quantitative calculation to be made of vacuum-tube performance under many conditions and permits a qualitative estimate of the behavior under still other circumstances.

Amplification Factor.—The amplification factor μ is the ratio of the effectiveness of the grid and plate voltages in producing electrostatic fields at the cathode surface. The amplification factor is therefore determined by the geometry of the system comprising the grid, plate, and cathode electrodes, and its calculation in terms of the dimensions involved is a problem of pure electrostatics. Where the geometrical proportions can be readily introduced into an equation, as for example is the case when the plate and cathode are concentric cylinders and the grid is composed of a series of bars parallel to the central axis as in Fig. 51*d*, it is possible to derive formulas that will give the amplification factor with accuracy. Such ideal conditions are never completely realized in practice, however, because of supporting elements that distort the electrostatic field, and the amplification factor in practical tubes is obtained by empirical formulas that are based upon modifications of the ideal cases.[1] The amplification factor depends primarily upon the

[1] A comprehensive treatment of amplification-factor computations is given by Yuziro Kusunose, Calculation of Characteristics and the Design of Triodes, *Proc. I.R.E.*, vol. 17, p. 1706, October, 1929. For additional information the reader is referred to the following: R. W. King, Calculation of the Constants of Three-electrode Thermionic Vacuum Tube, *Phys. Rev.*, vol. 15, p. 256, April, 1920; John M. Miller, The Dependence of the Amplification Constant and Internal Plate Circuit Resistance

grid structure and will be increased by anything that causes the grid to shield the cathode more completely from the plate. Thus larger grid wires or a closer spacing of the grid wires will increase the amplification factor, as does also increasing the distance between the grid and plate. The amplification factor of ordinary three-electrode tubes ranges from about 3 as the minimum to about 40 as the practical maximum, the exact value depending upon the purpose for which the tube was designed.

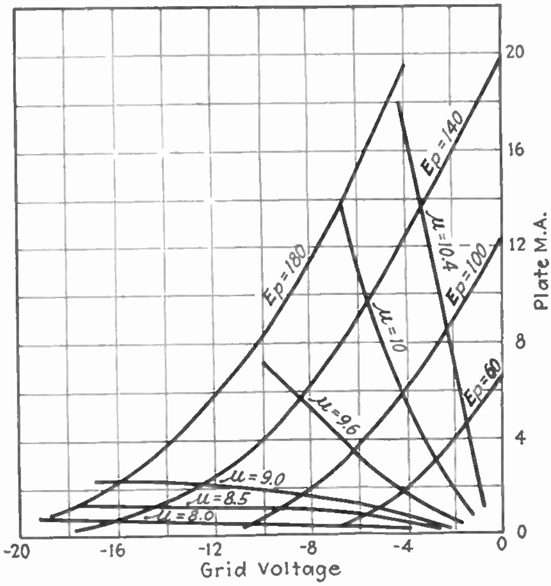


FIG. 56a.—Curves showing variations in amplification factor in a tube having no voltage drop in the cathode (heater-type cathode). The amplification factor is relatively constant over most of the region of negative grid potential but tends to become less as the plate current becomes less, particularly when the grid is highly negative at the same time.

If the relative effects of the grid and plate voltages in producing electrostatic field at the cathode were the same for all parts of the cathode, the amplification factor μ would be absolutely independent of plate, grid, and filament voltages. In commercial tubes various mechanical requirements, such as the necessity of supporting wires and the inevitable imperfections in construction, result in dissymmetry that causes different parts of the tube to have somewhat different amplification factors. The over-all amplification factor of such a combination will vary with plate, grid, and filament voltages and will tend to become lower as cut-off is approached because as the grid becomes more negative those parts having the highest value of μ will reach cut-off first, leaving only the low μ parts of the tube contributing to the space current. The way in

of Three-electrode Vacuum Tubes upon the Structural Dimensions, *Proc. I.R.E.*, vol. 8, p. 64, February, 1920.

which the amplification factor varies over the characteristic curves of a typical tube is shown in Fig. 56*a*. It is seen that over the main part of the characteristic, which also represents the usual operating range, the amplification factor does not vary by more than 10 to 15 per cent, but that at very low plate currents the variation in the amplification factor is greater. The results shown in this figure are typical of a large number of tubes that have been investigated.[1]

Dynamic Plate Resistance.—The dynamic plate resistance of a vacuum tube represents the resistance which the plate circuit offers to a small increment of plate voltage. Thus when an increment of plate voltage ΔE_p produces an increment in the plate current of Δi_p the dynamic plate resistance is given by the relation

$$\text{Dynamic plate resistance} = R_p = \frac{\Delta E_p}{\Delta I_p} = \frac{dE_p}{dI_p} \tag{60}$$

The plate resistance is therefore the reciprocal of the slope of the plate current-plate voltage characteristic shown in Fig. 54 and depends upon the grid and plate voltages at the operating point under consideration. It is important to remember that the plate resistance is determined by the slope of the plate voltage-current curve, being lowest at points where the slope is greatest, *and is not equal to the ratio of total plate voltage to total plate current.*

In any particular tube the dynamic plate resistance depends primarily upon the plate current and only to a small extent upon the combination of grid and plate voltages used to produce this current. The dynamic plate resistance furthermore becomes progressively lower as the plate current is increased. This behavior is a result of the fact that the different plate voltage-plate current curves in Fig. 54 differ only in being displaced along the plate-voltage axis, and because the slope of each curve increases as the plate current becomes greater. The result is that the plate resistance R_p is the same for all curves with a given current, and becomes less as the current is increased. This ideal is not entirely realized in practical tubes as a result of the same irregularities that prevent the amplification factor from being constant, so that in commercial tubes the plate resistance is to a certain extent affected by the combination of grid and plate voltages involved, as well as the value of plate current. The characteristics of an actual commercial tube are shown in Fig. 56*b* for a representative case.

The effect of a voltage drop in the cathode is to increase the plate resistance to a value considerably higher than that which would be obtained with a similar tube having no voltage drop in the cathode.

[1] A more complete discussion of the causes of the dependence of the amplification factor upon electrode voltages is to be found in F. E. Terman and A. L. Cook, Note on Variations in the Amplification Factor of Triodes, *Proc. I.R.E.*, vol. 18, p. 1044, June, 1930.

This is particularly pronounced in tubes having a high amplification factor, where the voltage drop in the filament is generally comparable with the effective anode voltage. Thus a tube having an amplification factor of 30 when operating with a grid voltage of -1.5 and a plate voltage of 120 will have an effective anode potential of $-1.5 + {}^{120}\!/_{30}$, or 2.5 volts. If such a tube has a 5-volt filament it is apparent that the effective anode potential will be positive to only approximately one half of the filament, leaving the other half inoperative and increasing the plate resistance accordingly. With tubes of low amplification factor, or in tubes operating at very high plate voltages, this effect is usually negligible.

The plate resistance of a vacuum tube depends upon the dimensions and relative positions of the cathode, plate, and grid. It becomes less as the effective cathode area is increased, *i.e.*, as the cathode is made longer or of larger diameter, and as the distance between the cathode and the other electrodes is reduced. The dynamic plate resistance can be computed with accuracy by theoretical formulas only for the same simple and symmetrical geometrical conditions for which the amplification factor can be determined by exact methods. In commercial tubes it is necessary to use empirical modifications of these formulas to take into account the effect of dissymmetries in the geometry. This is usually done by developing empirical rules for defining an effective cathode area that when substituted in the theoretical formula will give results in agreement with those actually observed.[1] Since the power required to heat the cathode depends upon the surface area of the cathode it may be said in a general way that the plate resistance is dependent upon the cathode heating power and is lowered by employing a larger cathode.

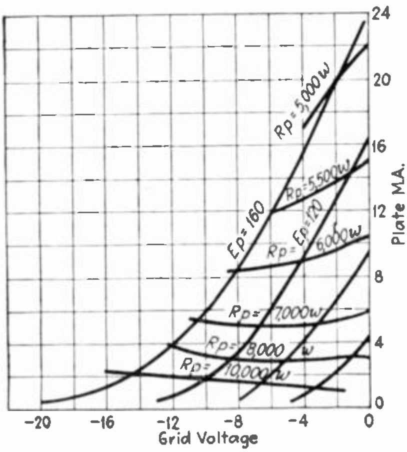


FIG. 56b.—Curves showing variations in the plate resistance of a tube having no voltage drop in the cathode. The resistance is approximately the same for a given plate current irrespective of the combination of plate and grid voltages used to produce this current.

The plate resistance of tubes differing only in grid structure decreases as the amplification factor is lowered. This is because the electrostatic field produced near the cathode by a given plate voltage is inversely proportional to the amplification factor. As a consequence a given increment of plate voltage produces an increment in the electrostatic

[1] For further information regarding such calculations see Yuziro Kusunose and R. W. King, *loc. cit.*

field at the surface of the cathode (and consequently an increment in the plate current) that varies inversely with the amplification factor. Tubes having no voltage drop in the cathode and differing only in the grid structure will have a plate resistance that is approximately directly proportional to the amplification factor of the tube. In the case of filament-type tubes this is only roughly true because as the amplification factor is increased by using a finer mesh grid structure, more and more of the tube becomes inoperative as a result of voltage drop in the filament, and the plate resistance will increase somewhat faster than the amplification factor.

Mutual Conductance.—The mutual conductance G_m is the ratio of amplification factor to plate resistance, that is

$$G_m = \frac{\mu}{R_p} \tag{61}$$

The mutual conductance is usually expressed in micromhos, and will range from 250 to 2000 for ordinary receiving tubes. The mutual conductance represents the rate of change of plate current with respect to a change in grid voltage. Thus if the grid voltage is changed by ΔE_g, the resulting plate-current change Δi_p is related to the mutual conductance by the equation

$$\left.\begin{aligned} \Delta i_p &= \Delta E_g G_m \\ G_m &= \frac{di_p}{dE_g} \end{aligned}\right\} \tag{62}$$

The mutual conductance is a rough indication of the design merit of a tube. This is because a low plate resistance and a high amplification factor are desired, and the mutual conductance measures the extent to which this feature is attained. Tubes of equal design merit but with slightly different values of amplification factor will have substantially the same value of mutual conductance under normal operating conditions. When tubes with widely different values of μ are compared the tendency is for the mutual conductance to be less for the high amplification-factor tubes, and this effect is especially pronounced when there is a voltage drop in the cathode.

In the table on page 116 are presented the most important characteristics of a large number of triodes used in radio receivers. This table includes all the more important three-electrode tube types in use in this country.

31. Constructional Features of Small Tubes.—The outstanding constructional features of small tubes are shown by the cut-away photograph of Fig. 57. The plate, grid, and cathode are supported from wires imbedded in the press, with certain of these wires passing through the

Table V.—Characteristics of Vacuum Tubes Used in Radio Receivers

Type	Type of cathode	Cathode heating power			Amplif. factor	Normal electrode voltages		Properties			Principal uses
		Volts	Amperes	Watts		Plate voltage	Negative grid bias	Plate current, ma	Plate resistance R_p, ohms	Mutual conductance G_m, micromhos	
'01A	Thoriated-tungsten filament	5.0	0.25	1.25	8.0	90 135	4.5 9.0	2.5 3.0	11,000 10,000	725 800	Transformer coupled amplification, detection, small oscillators
'40	Thoriated-tungsten filament	5.0	0.25	1.25	30	135 180	1.5 3.0	0.2 0.2	150,000 150,000	200 200	Resistance and impedance coupled amplification
'26	Oxide-coated filament	1.5	1.05	1.58	8.2	90 180	5.0 12.5	3.8 7.4	8,600 7,000	955 1,170	General-purpose tube similar to '01A but designed for alternating current on filament
'27	Oxide-coated heater	2.5	1.75	4.37	9.0	90 180	6.0 13.5	2.7 5.0	11,000 9,000	820 1,000	General-purpose tube similar to '01A and '26, but heater cathode
'99	Thoriated-tungsten filament	3.3	0.063	0.21	6.6	90	4.5	2.5	15,500	425	General-purpose tube similar to '01A but used where low filament power is important
'12	Oxide-coated filament	1.1	0.25	0.275	6.6	90 135	4.5 10.5	2.5 3.5	15,500 15,000	425 440	General-purpose tube similar to '99
'20	Thoriated-tungsten filament	3.3	0.132	0.436	3.3	135	22.5	6.5	6,300	525	Power-amplifier tube used where low filament power is important
'30	Oxide-coated filament	2.0	0.06	0.12	8.8	90	4.5	2.0	12,500	700	General-purpose tube similar to '99 and '12
'31	Oxide-coated filament	2.0	0.13	0.26	3.5	135	22.5	8.0	4,000	875	Power tube similar to '20
'12A	Oxide-coated filament	5.0	0.25	1.25	8.5	90 135	4.5 9.0	5.2 6.2	5,600 5,300	1,500 1,600	General-purpose tube similar to '01A
'71A	Oxide-coated filament	5.0	0.25	1.25	3.0	90 180	16.5 40.5	12.0 20.0	2,250 1,850	1,330 1,620	Power tube
'10	Thoriated-tungsten filament	7.5	1.25	9.38	8.0	250 425	18.0 35.0	10.0 18.0	6,000 5,000	1,330 1,600	Oscillator tube
'45	Oxide-coated filament	2.5	1.5	3.75	3.5	180 250	33.0 48.5	25.0 34.0	1,900 1,750	1,850 2,000	Power tube similar to '71A but greater power capacity
'50	Oxide-coated filament	7.5	1.25	9.38	3.8	250 450	41.0 80.0	28.0 55.0	2,100 1,800	1,800 2,100	Power tube similar to '45 but greater power capacity

glass and acting as leads. There is also usually some method provided for holding the top of the electrode structure in the proper position, which in the tube of Fig. 58 is a mica sheet. The technique of assembling the tube is similar to that employed in the manufacture of incandescent lights. The electrodes are formed to the proper shape and spot-welded to the supporting wires, which are then held in position while the press is made at the end of the stem. The bulb is now sealed to the stem, after which the tube is evacuated and based.

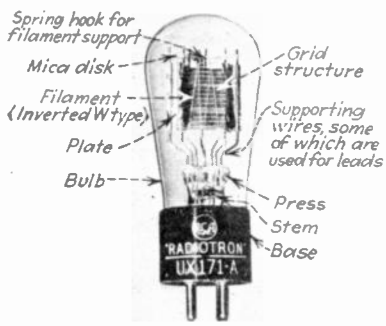


Fig. 57.—Photograph showing constructional details of typical three-electrode vacuum tube.

The plates of small tubes are usually formed of nickel since this material is easily worked and welds readily. The most common construction uses sheet (wire screen is also used sometimes), with the rigidity of the electrode increased by crimping the flat surfaces and turning over the edges to form a flange. The surface of the plate is often blackened to facilitate the radiation of the heat produced by the impact of electrons. The grid structure is nearly always made of molybdenum, which is able to withstand high temperatures, is stiff, and is readily welded. The grid is wound in the form of a coil of the proper shape and pitch and is then spot-welded to the supporting wires. The filament in filament-type tubes is strung in the form of an inverted V or inverted W on hooks which keep it under tension.

The air is removed from the tube by the use of a motor-driven oil pump supplemented by a molecular pump equipped with a liquid-air trap. In carrying out the evacuation it is necessary to remove not only all of the gas that is present within the space enclosed by the bulb, but also the occluded gas contained in the metal and glass, if the vacuum is to be maintained for any length of time. The removal of these occluded gases is carried out by baking the entire tube at a temperature just below the softening point of glass, while keeping the pump in operation, and by heating the electrodes to a high temperature either by electron bombardment or by placing the electrodes in the field of a coil carrying radio-frequency current. As a final step it is general practice to volatilize (or flash) a small quantity of some suitable substance, called a "getter," in the tube to remove the gas by either chemical or mechanical action. Magnesium is widely used for this purpose, but other materials, such as calcium, barium, phosphorus, etc., can also be employed. Some such getters also act as "keepers" in that they not only remove what gas is present in the tube at the time of flashing but also combine with any gas that may be liberated within the tube subsequently.

The amount of gas present in a vacuum tube can be estimated by measuring the grid current that flows when a high positive voltage is applied to the plate, and the grid is made negative. Under such conditions the electrons flowing to the plate ionize gas molecules by collision, producing positive ions that are attracted to the negative grid. The number of such ions is proportional to the space current and to the gas pressure, and so can be used as an indication of the degree of vacuum. A high vacuum gives a negligible grid current, while traces of current indicate traces of gas. When considerable gas is present the ionization becomes sufficiently intense to give a visible glow in the tube, and the tube is said to be "soft."

CHAPTER V

TRIODE AMPLIFIERS

32. Vacuum-tube Amplifiers.—The amplifying property of a vacuum tube results from the fact that the grid draws no electrons when maintained at a negative potential, while at the same time variations in the negative grid potential cause corresponding variations in the space current. In this way it is possible for a voltage representing practically no energy to control a relatively large space current and to develop a corresponding quantity of electrical energy in an anode circuit. The vacuum tube is capable of amplifying voltages of all frequencies up to the highest used in radio communication, and by employing a number of tubes in cascade can give almost any desired amount of amplification. The vacuum-tube amplifier is therefore one of the most important instruments in electrical engineering, having made possible the long-distance telephone, talking pictures, radio sets with loud-speakers, television, etc.

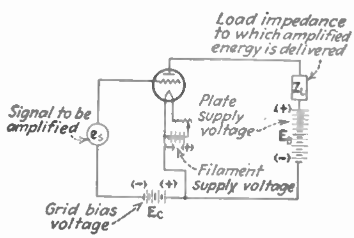


FIG. 58.—Basic circuit of vacuum-tube amplifier employing a three-electrode vacuum tube.

The basic circuit of the triode amplifier is shown in Fig. 58. The voltage E_c is for the purpose of maintaining the grid negative with respect to the cathode so that no electrons will be drawn to the grid, and is known as the grid bias or C-voltage. The signal to be amplified is represented in the figure by e_s and is applied to the grid in series with the grid-bias voltage. It is necessary that the amplitude of signal be so related to the grid-bias voltage that the instantaneous grid potential never becomes positive with respect to the cathode. Variations in the plate current that result from the action of the signal voltage on the grid must flow through the impedance Z_L which is in series with the plate supply voltage and is known as the load impedance. The energy which is supplied to the load impedance by the variations in the plate current represents the useful part of the amplified energy. The signal voltage applied to the grid supplies substantially no energy to the tube because a negative grid draws no electrons, but at the same time causes plate-current variations that deliver considerable quantities of energy to the load impedance. The result is that the output energy of the tube, *i.e.*, the energy delivered to the load impedance, can be made many times that required to produce the signal applied to the grid.

Methods of Classifying Amplifiers.—Amplifiers are classified in ways descriptive of their character and properties. The first classification is according to the frequency to be amplified, and leads to the three broad divisions known as audio frequency, radio frequency, and direct-current amplifiers. Audio-frequency amplifiers are intended for amplifying currents of audible frequencies, that is from about 15 cycles per second to approximately 10,000 cycles. Frequencies higher than 10,000 to 15,000 are considered as radio frequencies.

Amplifiers are also classified as to whether they handle a wide or narrow band of frequencies. A band of frequencies is considered wide or narrow in proportion to the ratio of the width of the band to the frequency at its center. Thus the group of frequencies lying between 100 and 5000 cycles is said to represent a wide band, while the frequency band from 1,000,100 to 1,005,000 cycles, which extends over the same frequency range, is narrow. When substantially equal amplification is to be obtained over a band that is wide according to this definition the amplifier is said to be untuned, while when the amplifier is associated with sharply resonant circuits so arranged that only a narrow band of frequencies is amplified it is spoken of as a tuned amplifier.

Amplifiers can also be divided into the following types, based upon the results which the amplifier is to produce:

1. Voltage amplifiers.[1]
2. Power amplifiers.
3. Linear or Class B amplifiers.
4. High-efficiency or Class C amplifiers.

The power amplifier has as its purpose the production of large quantities of undistorted output power and is used to develop the energy required to operate loud-speakers and similar equipment. Voltage amplifiers have as their object the production of the maximum possible voltage from a given signal applied to the grid. This is to be contrasted to power amplifiers where the object is to deliver as much power as possible to the load impedance. Linear amplifiers are used in the amplification of a single frequency where it is desired to obtain an output voltage proportional to the applied voltage without regard to any harmonics that may be produced in the amplifier and are a special form of power amplifier. The high efficiency or Class C amplifier is characterized by a high efficiency of conversion of direct-current plate energy to alternating-current output energy, but this amplifier is neither linear (*i.e.*, its output is not proportional to the input) nor distortionless (*i.e.*, its output is not an exact repetition of the signal voltage applied to the grid).

Each of these four types of amplifier has its own sphere of usefulness. Thus consider the problem of obtaining a large quantity of undistorted

[1] Voltage and power amplifiers giving an output wave that has the same shape as the signal are sometimes called Class A amplifiers.

power output from a signal voltage that is so small as to require more amplification than can be obtained from a single tube. Under such circumstances it is customary to use a number of amplifiers in cascade, each one of which amplifies the output of the preceding tube and delivers its output to another tube for additional amplification. In arranging such an amplifier the best results are obtained by making all the amplifying tubes except the last one operate as voltage amplifiers, while the last tube functions as a power amplifier. The power tube then has the maximum possible signal voltage applied to its grid and is therefore able to deliver the greatest amount of power. There is no object in making the intermediate tubes function as power amplifiers, since the purpose of these tubes is to increase the signal voltage delivered to the tube, the power output of which represents the output of the amplifier.

The type of amplifier is fixed primarily by the constants of the associated electrical circuits and by the grid and plate voltages employed. It is possible to make any particular tube function as any one of the four types of amplifiers, although the tube characteristics best suited for each type of amplifier are somewhat different.

33. Distortion in Amplifiers.—An ideal amplifier produces an output that exactly duplicates the input in all respects except magnitude. An actual amplifier can fall short of this ideal by failing to amplify the different frequency components of the input voltage equally well, by giving an output that is not exactly proportional to the amplitude of the input or by making the relative phases of the different frequency components in the output differ from the relative phase relations existing in the input. These effects are commonly referred to as frequency, amplitude (or non-linear), and phase distortion, respectively.

Frequency distortion (*i.e.*, unequal amplification of different frequencies contained in the signal) is particularly important in audio-frequency amplifiers and is more difficult to eliminate the wider the band of frequencies that must be amplified with substantial equality. Frequency distortion tends to be greater as the amplification per stage is increased, so if low distortion is desired it is necessary to make some sacrifice in the amplification. An example of frequency distortion is given in Fig. 59, in which the high-frequency component of the voltage to be amplified is discriminated against.

Amplitude distortion results from a non-linear relation between voltage and current in either the input (*i.e.*, grid) or the output (*i.e.*, plate) circuits of the amplifier. Non-linearity of the input circuit results when the grid is allowed to go positive during a portion of the signal-voltage cycle. This is because the grid-cathode resistance of the tube is relatively low when the grid is positive, and extremely high when negative, with the result that the voltage actually applied to the grid is usually distorted if the grid is not maintained negative at all times. Amplitude distortion also occurs in the output circuit as a result of the non-linear relation between grid voltage and plate current and, while not completely avoidable, can be kept small by proper attention to the conditions under which the amplifier operates. In particular the amplitude distortion introduced by the plate circuit will be least at points where the plate current-plate voltage characteristic as given in Fig. 53 has the least curvature, and when the signal voltage applied to the grid is small. The use of a high impedance load also tends to reduce amplitude distortion.

Amplitude or non-linear distortion results in the production of frequencies in the amplifier output that are not present in the input voltage applied to the grid. The most important of these distortion frequencies are harmonics of the frequencies contained in the input, and sum and difference frequencies formed by combinations of the signal components. An example of amplitude distortion is shown in Fig. 59.

Phase distortion results when the relative phase relations of the different frequency components contained in the signal voltage are disturbed in such a way as to make the wave form of the amplifier output differ from the wave shape of the signal without changing the magnitudes of the frequency components that are involved. This effect is illustrated by the top and bottom waves of Fig. 59, which contain the same percentage of third harmonic in different phase relations with respect to the fundamental. The resulting wave forms are seen to be very different, although the amount of energy in the harmonic is the same in both cases. In order that the original wave shape may be maintained it is necessary that the phase shift in the amplifier be exactly proportional to the frequency, and furthermore that the phase shift when extrapolated to zero frequency be zero or some integral multiple of π. The phase shift of an amplifier is related to the time required for the transmission of different frequencies through the amplifier, and when phase distortion exists it means that different frequencies are transmitted with different speeds and hence do not arrive

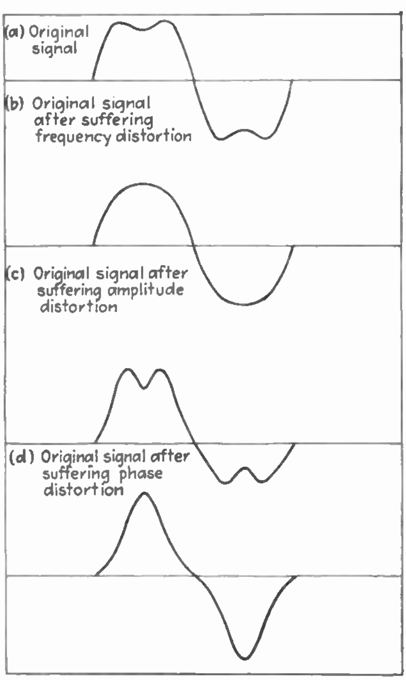


FIG. 59.—Series of waves showing the effects produced by frequency, amplitude, and phase distortion.

at the output at the same time. From a practical point of view, phase distortion is relatively unimportant in audio-frequency amplifiers because the phases can be altered over a wide range without producing a noticeable effect to the ear. It is only where the time of transmission has an order of magnitude comparable with the duration of the signal voltage applied that phase distortion becomes of importance. This situation is encountered in television circuits, and in long telephone and telegraph lines.

34. Equivalent Circuit of the Vacuum-tube Amplifier.—The variations produced in the plate current of a vacuum tube by the application of a signal voltage to the grid are exactly the same variations that would be produced in the plate current by a generator producing a voltage $-\mu e_s$ acting inside the tube from cathode toward the plate, in a circuit consisting of the tube plate resistance in series with the load impedance. The effect on the plate current of applying a signal voltage e_s to the grid is therefore exactly as though the plate-cathode circuit of the tube was a generator developing a voltage $-\mu e_s$ and having an internal resistance equal to the plate resistance of the tube. This leads to the equivalent circuit of the vacuum-tube amplifier which is shown in Fig. 60 and which is the basis of all amplifier designs and calculations.

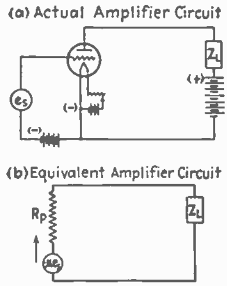


FIG. 60.—Equivalent circuit of the vacuum-tube amplifier, in which the effect on the plate current of the signal voltage applied to the grid is taken into account by considering the plate-cathode circuit of the tube to be a generator having an internal resistance equal to the tube plate resistance and developing a voltage of $-\mu e_s$ acting from cathode toward anode.

The equivalent circuit of the amplifier gives only those currents and voltage drops that are produced as a result of the application of a signal voltage upon the amplifier grid, and the actual currents and potentials existing in the plate circuit are the sum of the currents and potentials developed in the equivalent circuit and those existing in the amplifier when no signal is applied. Since the steady values that are present when no signal is applied are of no particular interest so far as the amplifier performance is concerned, it is usually unnecessary to superimpose them upon the results calculated on the basis of the equivalent circuit. In particular when the signal applied to the grid is an alternating voltage, as is usually the case, the equivalent circuit gives directly the alternating-current currents produced in the plate circuit by the signal voltage and which are superimposed upon the direct-current quantities present when no signal is applied.

The equivalent circuit gives the exact performance of the vacuum-tube amplifier provided the resistance R_p and the amplification factor μ that are used in setting up the equivalent circuit correspond to the values possessed by these tube constants at the plate and grid voltages existing at the instant in question.[1] The equivalent amplifier circuit has its principal usefulness when the grid- and plate-voltage variations during the signal-voltage cycle are so small that the plate resistance and amplification factor are substantially constant. Under these conditions the equivalent circuit can be set up using values of amplification factor and plate resistance that exist when no signal is being applied to the grid, and can then be solved as an ordinary electrical circuit. When the voltage applied to the grid is sufficiently large to cause the plate resistance R_p (or the amplification factor μ) to vary appreciably during the cycle, the equivalent circuit becomes non-linear (*i.e.*, amplitude distortion is present), and the mathematical solution of the circuit becomes too complicated for practical calculations. This presents no serious limitations to the use of the equivalent circuit, however, because amplifiers are nearly always operated in such a way that the non-linear distortion is small, and hence so that the assumptions of constant amplification factor and constant plate resistance are realized with an accuracy that is sufficient for all practical purposes.

35. Audio-frequency Voltage Amplifiers—Resistance Coupling.—In voltage amplifiers the object is to obtain from the amplifier output as much voltage as possible to be applied to the grid of the succeeding amplifier tube. One way of doing this is to place in the amplifier plate circuit a high-resistance load called the coupling resistance.

The circuit of a practical resistance-coupled amplifier is shown in Fig. 61*a* in which R_c is the resistance load across which the amplified

[1] A rigorous proof of the equivalent amplifier circuit of the vacuum tube is given by John R. Carson, A Theoretical Study of the Three-element Vacuum Tube, *Proc. I.R.E.*, vol. 7, p. 187, April, 1919. A somewhat less general demonstration is as follows: The plate current *flowing from cathode to plate* which a grid voltage increment e_s produces will be called i_p. This plate current i_p causes a voltage drop $-Z_L i_p$ in the plate load impedance Z_L, so that the application of e_s to the grid reduces the voltage actually applied to the plate of the tube by an amount $\Delta E_p = -Z_L i_p$, and the plate-current increment i_p that flows is the result of the joint action of the voltage increment e_s applied to the grid and the reduction $-Z_L i_p$ in the plate voltage. Since the voltage e_s produces the same effect on the electrostatic field adjacent to the cathode as an added plate voltage of μe_s, the current i_p is the same current that would result from an increment in the plate voltage of $(\mu e_s + Z_L i_p)$, and when this increment is small the current increment flowing toward the plate is the negative of the plate-voltage increment divided by the dynamic plate resistance R_p, so that

$$i_p = -\frac{(\mu e_s + Z_L i_p)}{R_p}$$

or

$$i_p = \frac{-\mu e_s}{R_p + Z_L}$$

This last equation is simply a mathematical statement of the equivalent amplifier circuit.

voltage is developed. The grid-leak resistance and coupling condenser shown in Fig. 61 are for the purpose of preventing the direct-current voltage applied to the plate of the amplifier tube from also being applied to the grid of the tube to which the amplified voltage is delivered. The coupling condenser should be large enough to offer a low reactance to

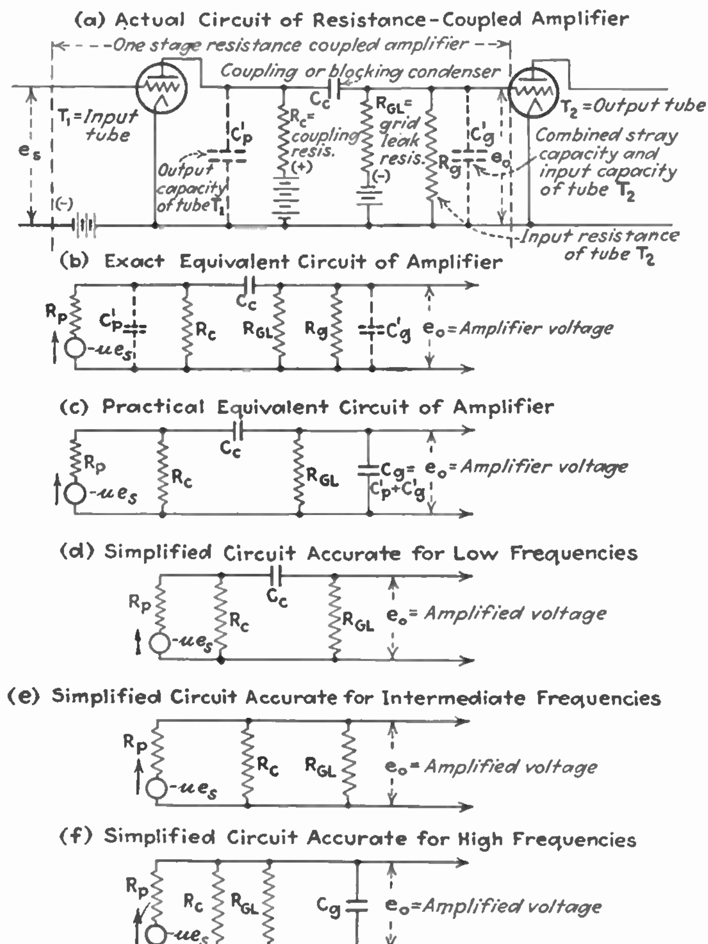


FIG. 61.—Circuit of resistance-coupled voltage amplifier, together with the equivalent circuit and simplifications of the equivalent circuit useful in making amplifier calculations.

the frequencies to be amplified, while the grid leak should have a very high resistance in order that the shunting effect of the grid leak and coupling condenser upon the coupling resistance may be small. The equivalent circuit of the resistance-coupled amplifier is shown in Fig. 61*b* and can be used to predict the amplification with a high degree of accuracy. The capacity C'_g and the resistance R_g represent the input capacity (plus stray capacities of the wiring) and input resistance respectively of the tube to which the amplified output is delivered, while C'_p represents the plate-cathode capacity of the amplifier tube plus stray wiring capacities that are in shunt with the plate resistance of the tube. For all ordinary purposes the capacities C'_p and C'_g can be combined into a single capacity as shown at Fig. 61*c*. It is also permissible under most circumstances to neglect the input resistance R_g of the tube to which the amplified output is applied because this resistance usually has an extremely high value compared with the impedance which it shunts.

The most important property of the resistance-coupled amplifier is the way in which the amplification varies with frequency. Such a characteristic is shown in Fig. 62 for a representative case and has as its distinguishing feature an amplification that is substantially constant

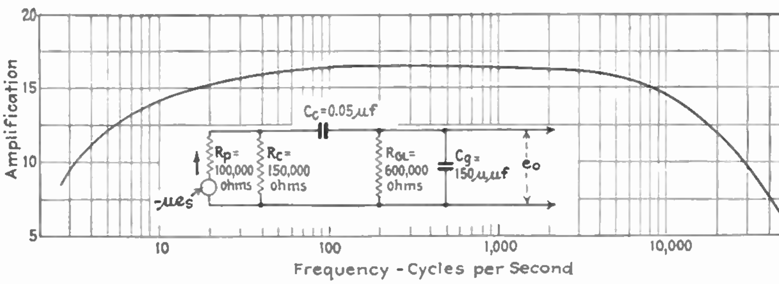


FIG. 62.—Variation of amplification with frequency in a typical resistance-coupled amplifier.

over a wide range of frequencies, but which drops off at both very low and very high frequencies. The falling off at very low frequencies is a result of the fact that the high reactance which the coupling condenser C_c offers to low frequencies consumes some of the low-frequency voltage that would otherwise be developed across the grid leak, while the reduction in amplification at high frequencies is caused by the low reactance of capacities C_p' and C_g' at high frequencies, which lowers the impedance across the terminals of R_c, with a consequent reduction in voltage developed at the output.

Maximum Amplification of Resistance-coupled Amplifier.—The maximum amplification of the resistance-coupled amplifier occurs at intermediate frequencies where the reactance of the coupling condenser C_c is so low as to consume negligible voltage, while the reactance of the tube and stray capacities in shunt with the coupling resistance is still so high as to cause negligible loss in amplification. Under such conditions the equivalent circuit of the resistance-coupled amplifier can be simplified by considering C_c to be a short circuit and by neglecting the tube and stray capacities C_p and C_g'. The resulting approximately equivalent circuit shown at Fig. 61*e* accurately represents the amplifier behavior at intermediate frequencies and can be used to calculate amplification at such frequencies by solving for the ratio of voltage developed across the grid leak (*i.e.*, output voltage) to the signal voltage e_s (*i.e.*, input voltage applied to the grid), with the result[1]

$$\left.\begin{array}{l}\text{Maximum amplification of}\\ \text{resistance-coupled amplifier}\end{array}\right\} = \frac{e_o}{e_s} = \mu\frac{R_{gl}R_c}{R_{gl}R_c + R_{gl}R_p + R_pR_c} \quad (63)$$

$$= \mu\frac{R_{eq}}{R_p + R_{eq}} \quad (63a)$$

where

μ = amplification factor of tube
R_c = coupling resistance
R_p = plate resistance of tube
R_{gl} = grid-leak resistance
R_{eq} = resistance of grid leak and coupling resistance in parallel.

This is the maximum amplification that can be obtained from the resistance-coupled amplifier and is seen to be proportional to the amplification factor μ of the tube and to increase as the ratio R_{eq}/R_p is increased. The amplification can never exceed the amplification factor of the tube and in practical cases will lie between one-half and three-fourths of the amplification factor.

Amplification at Low Frequencies.—At low frequencies the equivalent circuit of the resistance-coupled amplifier can be simplified by neglecting the high reactance of the small shunting capacity C_g, as shown at Fig. 61*d*, which can be used to calculate the low-frequency performance with a high degree of accuracy. While a little tedious and long, such calculations can be carried out in a straightforward manner without encountering any particular difficulties in evaluating the amplification e_o/e_s. It

[1] This formula can be derived by solving the equivalent circuit of Fig. 61*e* for e_o/e_s. Inspection of the circuit shows that

$$\frac{e_o}{e_s} = -\mu\frac{\text{resistance of } R_c \text{ and } R_{gl} \text{ in parallel}}{\text{total series resistance of equivalent circuit}}$$

The resistance R_{eq} of R_c and R_{gl} in parallel is $(R_cR_{gl})/(R_c + R_{gl})$, while the total circuit resistance is

$$\text{Total series resistance} = R_p + \frac{R_cR_{gl}}{R_c + R_{gl}}$$

$$= \frac{R_pR_c + R_pR_{gl} + R_cR_{gl}}{R_c + R_{gl}}$$

Substituting these values in the formula for e_o/e_s given above yields Eq. (63). The negative sign in the result signifies that the amplification reverses the phase and can be omitted in the formula giving the amplification.

is fortunately seldom necessary to make such calculations, however, because the low-frequency characteristics can with very little effort be estimated to an accuracy sufficient for most practical purposes. Thus at the frequency which makes the reactance of the coupling condenser equal the grid-leak resistance (*i.e.*, when $1/\omega C_c = R_{gl}$) the amplification will be slightly more than 70 per cent of the maximum amplification obtained from the amplifier as given by Eq. (63). This is because the voltage appearing across the grid leak when the reactance of the condenser equals the grid-leak resistance is 70 per cent (actually $1/\sqrt{2}$) of the voltage across the coupling resistance, and at this frequency the voltage across the coupling resistance R_c will be very close to the value at intermediate frequencies (actually slightly greater because the reactance of C_c increases the effective load impedance of the amplifier at low frequencies). A similar line of reasoning shows that at the frequency for which $1/\omega C_c = 2\ R_{gl}$ the amplification is approximately 50 per cent of the maximum value as given by Eq. (63) and cannot be less than 45 per cent of this maximum, while when the frequency is sufficiently high to make the reactance of the coupling condenser less than one-third of the grid-leak resistance the amplification is within 5 per cent of the value given by Eq. (63). These rules make it possible to estimate without calculation the low-frequency characteristics of a resistance-coupled amplifier to an accuracy that is sufficient for most purposes.

Amplification at High Frequencies.—At high frequencies the reactance of condensers becomes low so that, while the coupling condenser becomes of negligible importance, the tube and stray capacities C_g have an appreciable shunting effect on the coupling resistance, making the equivalent circuit of the resistance-coupled amplifier take a form at high frequencies that can be represented to a high degree of accuracy by the simplified equivalent circuit of Fig. 61*f*. The shunting effect of the capacity C_g lowers the impedance across the terminals of R_c and hence causes a reduction in amplification that becomes more pronounced as the frequency is increased. A solution of the equivalent circuit of Fig. 61*f* shows that the ratio of amplification obtained at high frequencies as obtained from Fig. 61*f* to the maximum amplification as given by Eq. (63) is[1]

[1] The derivation of this formula follows: Consider that R_{eq} represents the equivalent resistance of R_c and R_{gl} in parallel, and that X is the reactance of the shunting capacity C_g. The vector impedance Z which the plate circuit offers to the equivalent signal voltage $-\mu e_s$ is then

$$Z = R_p + \frac{1}{\dfrac{1}{R_{eq}} - \dfrac{1}{jX}} = R_p + \frac{jXR_{eq}}{jX - R_{eq}}$$

$$= \frac{R_p(jX - R_{eq}) + jXR_{eq}}{jX - R_{eq}}$$

$$\frac{\text{Actual amplification at high frequencies}}{\text{Maximum amplification}} = \frac{1}{\sqrt{1 + (R/X)^2}} \tag{64}$$

where

R = equivalent resistance of R_c, R_p, and R_{gl}, all in parallel
$X = 1/2\pi f C_g$ = reactance of capacity C_g.

The extent to which the amplification falls off at high frequencies is therefore determined by the *ratio which the equivalent resistance obtained by combining the plate resistance, coupling resistance, and grid-leak resistance in parallel bears to the reactance of the shunting capacity* C_g and is obtainable directly from Fig. 63 for any value of this ratio. This loss of amplification at high frequencies can be estimated by the fact that at the frequency which makes the reactance of the shunting condenser C_g equal three times the equivalent resistance of R_p, R_c, and R_{gl} in parallel the amplification drops to 95 per cent of the maximum given by Eq. (63), while at three times this frequency the amplification is $1/\sqrt{2}$, or 70.7 per cent of the maximum.

The signal current produced in the plate circuit by $-\mu e_s$ is therefore

$$i_s = \frac{-\mu e_s}{Z} = -\mu e_s \frac{(jX - R_{eq})}{R_p(jX - R_{eq}) + jXR_{eq}}$$

The output voltage e_o of the amplifier is the voltage which this current develops across the impedance formed by R_c, R_{gl}, and C_g in parallel, and so is

$$e_o = i_s \frac{jXR_{eq}}{jX - R_{eq}}$$

Substituting for i_s and simplifying shows the amplification e_o/e_s to be given by the expression

$$\frac{e_o}{e_s} = -\mu \frac{R_{eq}}{(R_p + R_{eq}) - \dfrac{R_pR_{eq}}{jX}}$$

The ratio of the actual amplification given by this equation, to the maximum amplification obtainable, as given in Eq. (63), is therefore

$$\frac{\text{Actual amplification}}{\text{Maximum amplification}} = \frac{R_p + R_{eq}}{R_p + R_{eq} - \dfrac{R_pR_{eq}}{jX}}$$

$$= \frac{1}{1 - \dfrac{R_pR_{eq}}{jX(R_p + R_{eq})}} = \frac{1}{1 - \dfrac{R}{jX}} \tag{65}$$

This last transformation is accomplished by noting that

$$R = \frac{R_pR_{eq}}{R_p + R_{eq}}$$

and gives the vector ratio of actual to maximum amplification. The absolute magnitude of this ratio is then Eq. (64).

Factors Affecting the Design of Resistance-coupled Amplifiers.—In the design of resistance-coupled amplifiers there are a number of factors to be considered in selecting the circuit elements. Since the maximum amplification is shown by Eq. (63) to be determined by the equivalent resistance of R_c and R_g, in parallel, it is necessary that both of these resistances be very large if the amplification is to be high. While there are no special difficulties encountered in using a high-resistance grid leak, the value of coupling resistance is limited by the fact that it must carry

Also good in lo freqs "R/X" is if $\frac{X_c}{R_g + \frac{R_c R_p}{R_c + R_p}}$

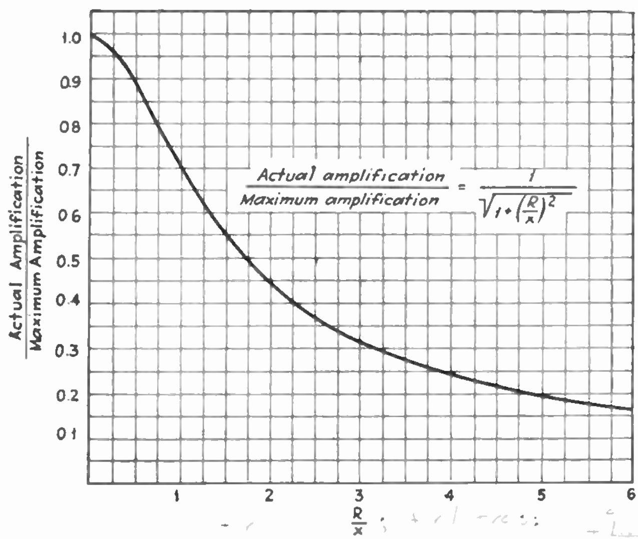


FIG. 63.—Plot of Eq. (64) giving ratio of actual amplification obtained at high frequencies with resistance coupling to the maximum amplification obtained at any frequency, expressed in terms of R/X, where X is the reactance at the frequency in question of the shunting capacity C_g across the amplifier output, and R is the equivalent resistance of grid-leak, plate, and coupling resistances in parallel.

the plate current of the tube and so consumes a direct-current voltage that must be supplied by the plate battery. High values of R_c therefore call for a high plate-supply voltage, and this consideration makes the maximum practical value of R_c lie in the neighborhood of two to three times the plate resistance of the tube. Higher coupling resistances give very little extra gain in amplification while requiring high plate voltage, while lower values are not warranted because the saving in plate-supply potential is offset by serious reduction in amplification. When the voltage actually applied to the plate is constant (*i.e.*, plate resistance R_p constant) the effect of changing the coupling resistance is primarily to alter the maximum amplification obtainable without appreciably changing the extent to which the amplification falls off at the high and low frequencies. This is brought out by Fig. 64*a* which is typical of the results that can be expected when the coupling resistance is varied, and which shows that the gain in amplification obtained by using very high coupling resistances as compared with moderate values is not so very great.

The point at which the amplification begins to fall off at low frequencies is determined almost solely by the product of grid-leak resistance R_{gl} and coupling-condenser capacity C_c (*i.e.*, by the ratio R_{gl}/X_c) which should be large if the region of full amplification is to extend to very low frequencies. Full amplification at low frequencies hence requires a high-resistance grid leak, or a large coupling condenser, or both. The results obtained by changing R_{gl} and C_c are clearly shown by curves *b* and *c* of Fig. 64, where it is seen that C_c affects the amplification only at very low frequencies and has no influence on the remainder of the characteristic, while varying the grid-leak resistance has in the main the same effect but is accompanied by a small change in the maximum amplification.

The extent to which the amplification drops off at high frequencies is determined by the magnitude of the capacity C_g in relation to the plate, coupling, and grid-leak resistances. The amplification will drop off at high frequencies unless the capacity C_g is small, or unless the resistance R formed by R_p, R_c, and R_{gl} in parallel is small as is seen in Fig. 64*a*, *b*, *d*. and *e*. In practical resistance-coupled amplifiers that are well designed the response at high frequencies is determined primarily by the amplification factor of the tube employed and will drop off more rapidly the higher the value of the amplification factor. One reason tending to bring this about is that a tube with a low amplification factor also has low plate resistance, and hence a low coupling resistance can be used, so that the shunting effect of C_g has less effect than when it is across a high coupling resistance. Furthermore the capacity C_g is determined largely by the amplification factor of the tube to which the amplifier output is delivered, and will be greater as the amplification factor of this tube increases. This is discussed in Sec. 43, where it is also shown that the input capacity of a tube depends upon the load impedance in the plate circuit and so will vary somewhat with frequency.

For most practical purposes C_g can be considered as approximately constant at a value that is given with fair accuracy by the relation

$$C_g = C_s + C_{pf} + C_{gf} + C_{gp}(1 + A) \tag{66}$$

where

C_g = equivalent capacity in shunt with coupling resistance
C_s = stray wiring capacity
C_{pf} = plate-cathode capacity of amplifier tube
C_{gf} = grid-cathode capacity of output tube
C_{gp} = grid-plate capacity of output tube
A = amplification of output tube from input signal to alternating-current voltage developed between plate and ground by load impedance (A cannot exceed μ of output tube).

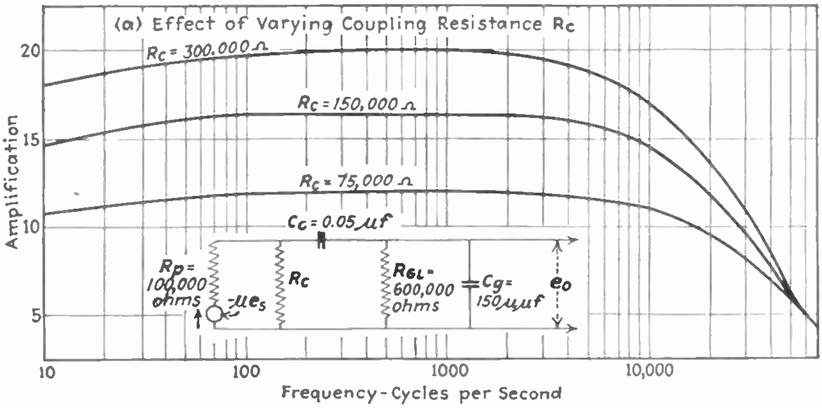


Fig. 64A.

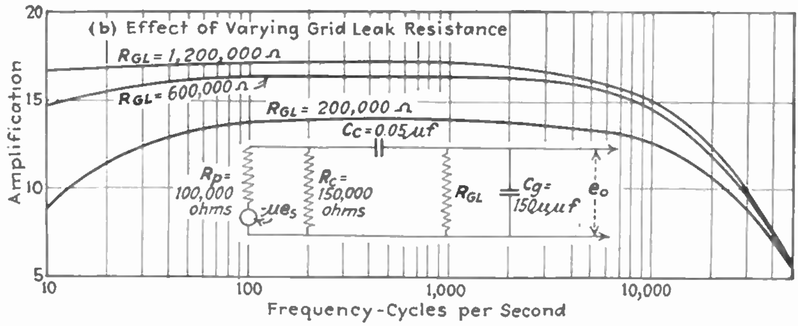


Fig. 64B.

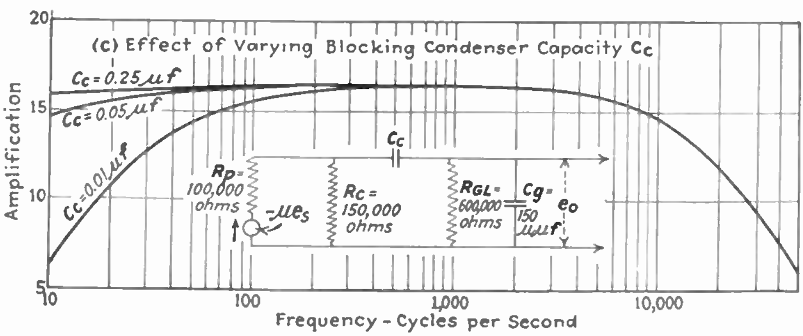


Fig. 64C.

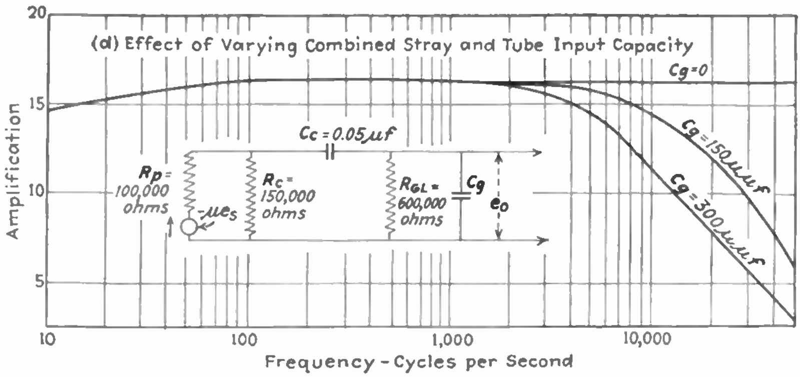


FIG. 64D.

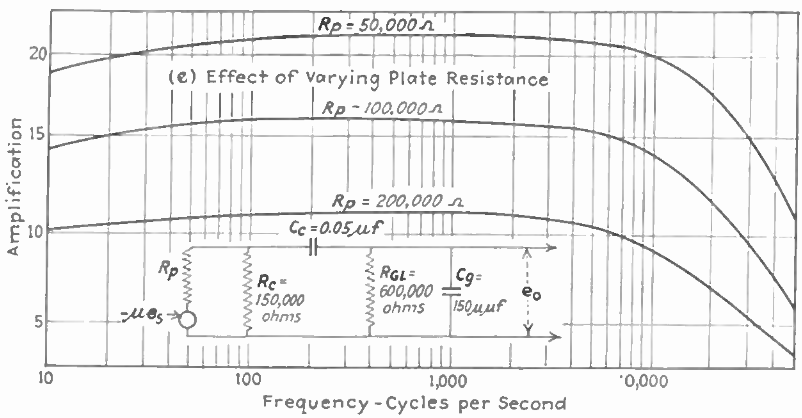


FIG. 64E.

FIG. 64.—Series of curves showing the effect which variations in circuit proportions have upon the way in which amplification varies with frequency in a resistance-coupled amplifier.

The capacities C_{pf}, C_{gf}, and C_{gp} usually range from 5 to 10 $\mu\mu$f in value while C_s depends upon the wiring and can be made small by proper attention to the arrangement of circuits. The result is that the last term in Eq. (66) is by far the most important and causes the capacity C_g to be almost proportional to the amplification factor of the output tube. A stage of resistance-coupled amplification delivering its output to a tube having a low amplification factor will therefore give a better high-frequency response than the same stage of amplification delivering its output to a high μ tube. Since most stages of amplification deliver their output to similar succeeding stages it is apparent that by using low μ tubes throughout, the response at high frequencies will fall off less from the maximum response than when the amplification factor is high.

These considerations show that though tubes with a low amplification factor give a relatively low amount of amplification per stage the amplification obtained is substantially constant up to very high frequencies. Tubes with a very high μ give a correspondingly large maximum amplification per stage but have a response that begins to drop off at moderately high frequencies. In audio-frequency amplifiers where full amplification must be maintained up to 5000 cycles, the highest amplification factor that it is feasible to use is in the neighborhood of 30 to 40, while lower values are not infrequently employed. Where it is necessary to maintain the amplification at still higher frequencies it is preferable to use tubes with values of μ in the neighborhood of 10 to 20.

The grid-bias potential E_c and the plate-supply battery voltage E_b should be selected with care if satisfactory results are to be obtained with resistance coupling. The grid bias is determined by the maximum signal voltage that the amplifier is to handle and must be sufficient to prevent this signal from making the grid positive, but the grid bias should not exceed this value appreciably because a low grid bias reduces the plate resistance of the tube and hence increases the amplification, as is brought out by Fig. 64*e*. The minimum grid bias that should be used is approximately -1 volt, since otherwise the velocity of emission of electrons will cause the grid to draw a small current. With the grid bias fixed in this way by the signal voltage the plate-supply potential E_b should have a potential as high as can be conveniently obtained in order to keep the plate resistance of the amplifier tube as low as possible. It is not advisable however to go to extremes in the matter of plate-supply voltage as the gains made by doing so are small. Amplitude distortion is seldom a factor in determining the circuit adjustments of resistance-coupled amplifiers because such amplifiers always operate with a relatively small signal voltage and with a large effective load impedance in the plate circuit.

Design Procedure.—In designing resistance-coupled amplifiers there is a fairly definite procedure that can be followed to advantage. The first step is to make a tentative selection of the tube to be employed, using the highest amplification factor consistent with the high-frequency response desired. The probable shunting capacity C_g and plate resistance R_p at the operating point are then estimated, and the coupling resistance R_c is given a value that is roughly twice R_p. The value of grid-leak resistance R_{gl} is then made three to six times R_c, after which the coupling-condenser capacity C_c is assigned the minimum value that will give the required low-frequency response with the grid-leak resistance already selected. The capacity C_c should be no larger than necessary to give the required low-frequency response because this insures the highest possible leakage resistance in C_c. Such leakage is to be avoided because it allows current to flow through R_{gl} from the source of plate voltage, as can be seen by examining Fig. 61, and puts a positive bias on the grid of the output tube. This bias voltage resulting from condenser leakage is proportional to R_{gl}, E_b, and to the leakage conductance of the condenser and can easily be sufficiently great to cause trouble. Resistance-coupled amplifiers using paper dielectric condensers will give satisfactory results only when the leakage resistance is extremely high for this type of condenser, and most commercial paper condensers have entirely

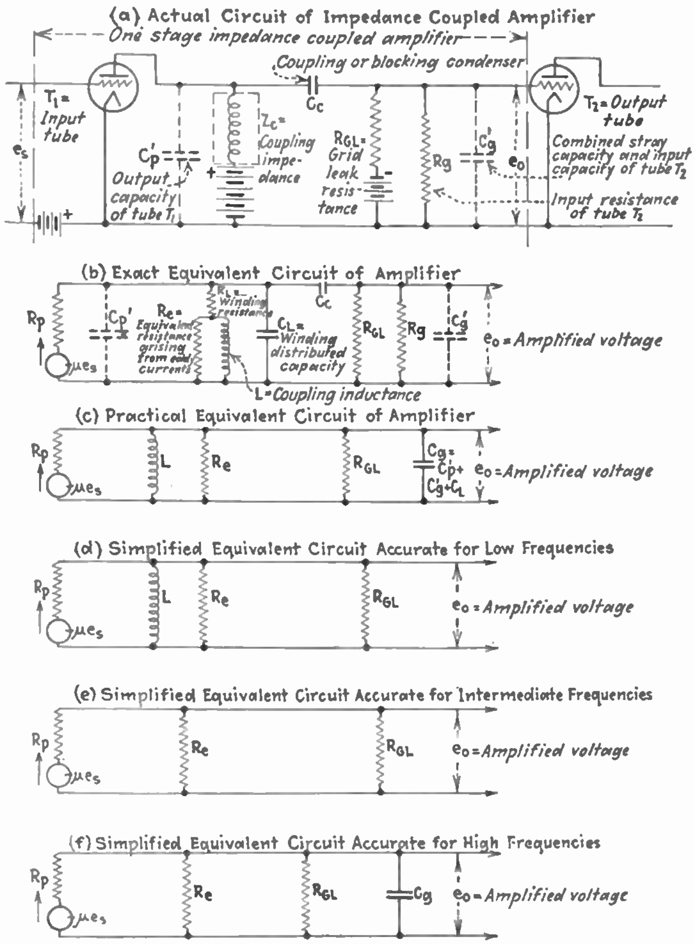


FIG. 65.—Circuit of impedance-coupled amplifier, together with the equivalent circuit and simplifications of the equivalent circuit useful in making amplifier calculations.

too much leakage to be suitable for coupling condensers in resistance-coupled amplifiers.

Resistance-coupled audio-frequency amplifiers are extensively used because they have the merit of combining low cost with an amplification that is very nearly constant over a very wide range of frequencies. On the other hand the amplification per stage is relatively small, and the performance is satisfactory only when the coupling condensers have high leakage resistance and when non-microphonic coupling and grid-leak resistances are used.

36. Audio-frequency Voltage Amplifiers—Impedance Coupling.—Instead of placing a resistance in series with the amplifier to develop the amplified voltage, as is done in the case of resistance coupling, it is possible to use an inductance as shown in Fig. 65. This arrangement is spoken of as impedance coupling. The amplification-frequency characteristic for a typical impedance-coupled amplifier is shown in Fig. 66 and is similar to that of the corresponding resistance-coupled-amplifier characteristic shown to facilitate comparison. It is seen that replacing the coupling resistance by a coupling impedance increases the maximum amplification but narrows the range of frequencies over which the amplification is constant, with the lower frequencies suffering most.

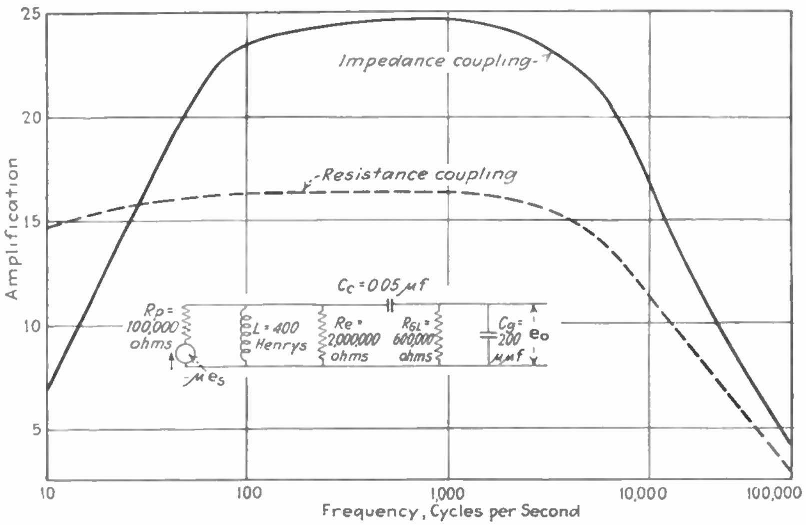


FIG. 66.—Variation of amplification with frequency in a typical impedance-coupled amplifier, compared with similar characteristics of corresponding resistance-coupled amplifier employing the same tube.

The characteristics of the impedance-coupled amplifier can be analyzed with the aid of the equivalent amplifier circuit, which for this type of amplifier has the form given in Fig. 65*b* and which is similar to that for resistance coupling except for the coupling inductance with its associated capacity and its resistances. The capacity C_L is the coil distributed capacity while R_L is the wire resistance of the coil, and R_e represents the eddy-current iron loss and is a very high resistance (usually in the order of megohms) in shunt with the coil. For all practical calculations the complicated circuit of Fig. 65*b* can be replaced by the simplified circuit shown at Fig. 65*c* in which C_g', C_p', and C_L have been lumped together, while C_c, R_g, and R_L have been neglected. The justification for these omissions is that R_L is ordinarily quite small compared with the other impedances that are in series with it, and R_g is so large as to have negligible effect, while C_c can be neglected because the low-frequency amplification in the typical impedance-coupled amplifier begins to fall off as a result of the low reactance of the coupling inductance long before there is appreciable loss of amplification as a result of voltage drop in the coupling condenser. With circuit proportions for which this situation is not true the coupling condenser must be included in the simplified equivalent circuit.

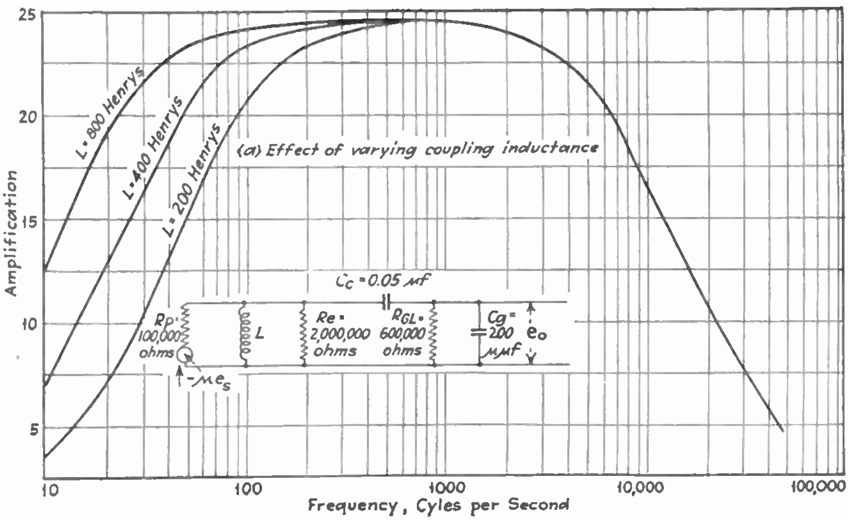


FIG. 67A.

Calculation of Amplification.—The maximum amplification of the impedance-coupled amplifier is obtained at intermediate frequencies centering about the region where C_g and L are in parallel resonance. For such frequencies the combined reactance of the coupling inductance L and capacity C_g is so high as to be substantially equivalent to an open circuit so that the amplifier performance can be represented with high accuracy by the simplified circuit shown at Fig. 65*e*, the solution of which gives the following relation:

$$\left.\begin{array}{l}\text{Maximum amplification of}\\ \text{impedance-coupled amplifier}\end{array}\right\} = \mu\frac{R_{eq}}{R_{eq} + R_p} \tag{67}$$

where

μ = amplification factor of amplifier tube
R_p = plate resistance of amplifier tube
R_{eq} = equivalent resistance of R_c and R_{gl} in parallel.

This maximum amplification with impedance coupling usually approaches the amplification factor of the tube much more closely than does the maximum amplification of the resistance-coupled amplifier.

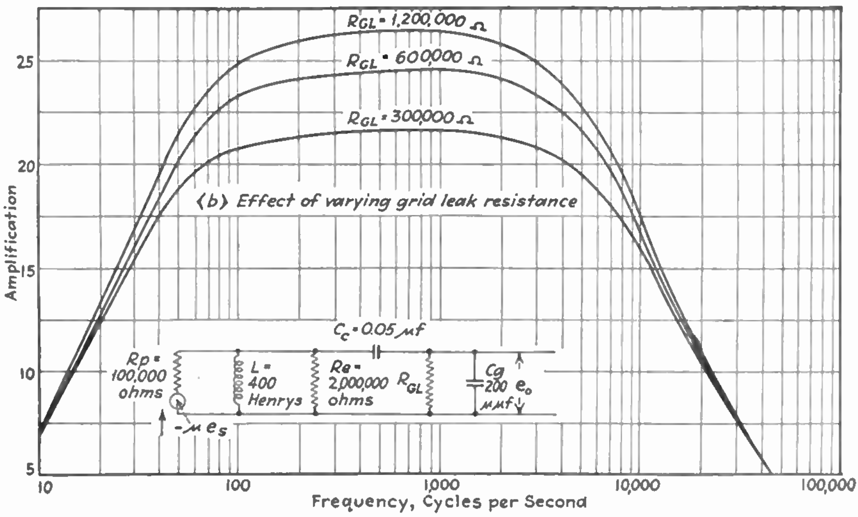


FIG. 67B.

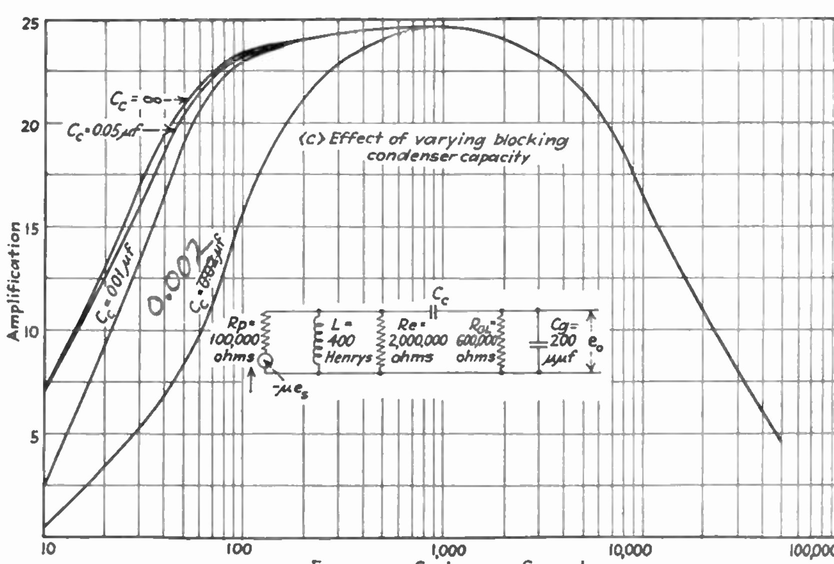


FIG. 67C.

At high frequencies the reactance of the coupling inductance L is so great as to be practically equivalent to an open circuit, while C_c has such a low reactance as to be substantially equivalent to a short circuit.

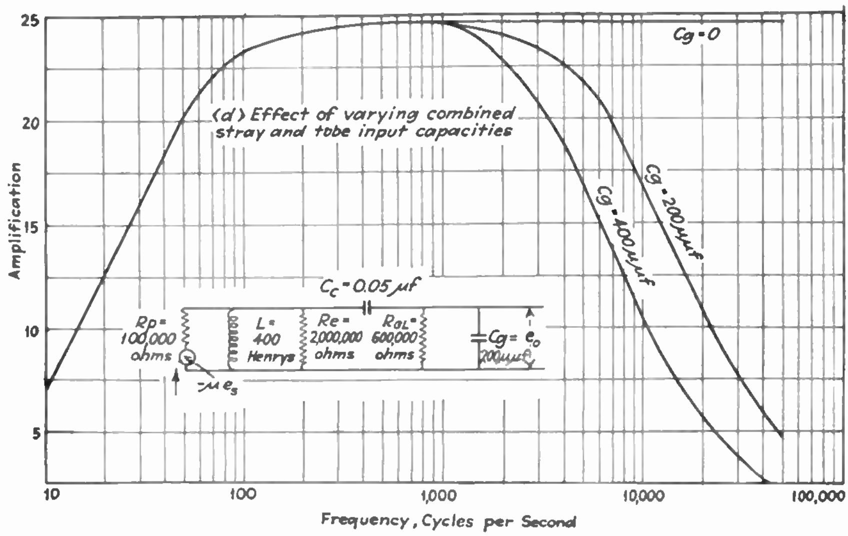


Fig. 67D.

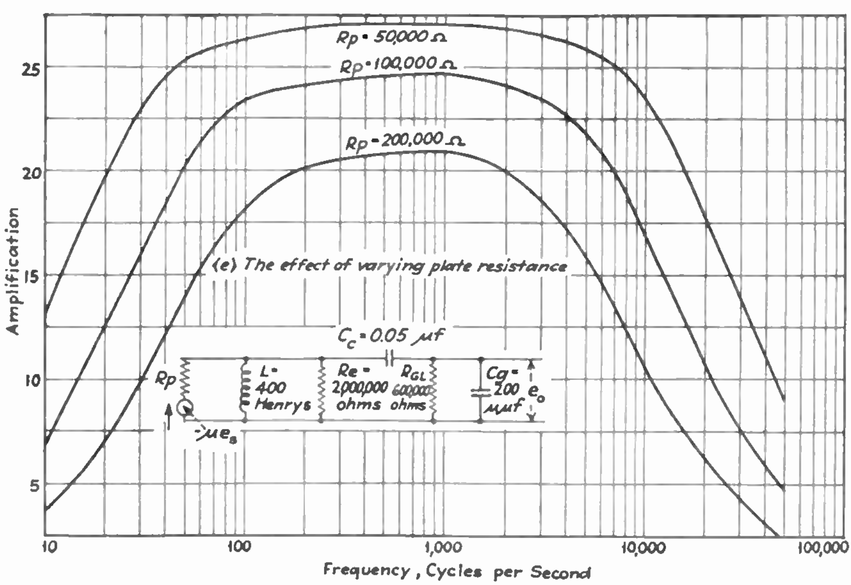


Fig. 67E.

Fig. 67.—Series of curves showing the effect which variations in circuit proportions have upon the way in which the amplification varies with frequency in an impedance-coupled amplifier.

The equivalent circuit of the impedance-coupled amplifier under these conditions is therefore represented with good accuracy by Fig. 65*f*. It will be observed that this equivalent circuit representing the behavior of the impedance-coupled amplifier at high frequencies is similar in all respects to the corresponding circuit representing the behavior of the resistance-coupled amplifier under like conditions. The way in which the impedance-coupled amplification falls off at high frequencies is therefore given by Eq. (64), if R in this equation is interpreted as representing the combined resistance of R_p, R_{gl}, and R_c in parallel. The ratio of actual amplification obtained at any high frequency to maximum amplification can also be obtained directly from Fig. 63 in terms of the ratio of X/R. The way in which the amplification falls off at high frequencies may be estimated by the fact that at the frequency which makes $1/\omega C_g = R$, the amplification is 70.7 per cent of the maximum value, while at one-third of this frequency the amplification is almost exactly 95 per cent of the maximum.

The amplification obtained with impedance coupling falls off at low frequencies because the low reactance of the coupling inductance at such frequencies causes only a small voltage to appear across the terminals of the inductance. At low frequencies the reactance of the shunting capacity C_g is so great as to have negligible effect, so that the equivalent circuit of the impedance-coupled amplifier is approximated by Fig. 65*d* at low frequencies. The coupling capacity C_c is not shown because it may be neglected with the circuit proportions usually employed but at the same time becomes of importance if the capacity C_c is very small, or if the frequency is very low. Neglecting the voltage loss in the coupling condenser the solution of the equivalent circuit, Fig. 65*d*, shows that the ratio of actual amplification obtained at low frequencies to the maximum amplification given by the amplifier is

$$\frac{\text{Actual amplification at low frequencies}}{\text{Maximum amplification}} = \frac{1}{\sqrt{1 + (R/X)^2}} \qquad (68)$$

where

R = equivalent resistance of R_p, R_{gl}, and R_e all in parallel

$X = \omega L$ = reactance of coupling inductance.

It will be noted that Eqs. (68) and (64) are identical, provided X and R are properly interpreted. The way in which the amplification drops off at low frequencies can therefore also be obtained directly in terms of ωL and R by the use of Fig. 63, and can be estimated by the fact that at the frequency which makes $\omega L = R$ the amplification is only 70.7 per cent of the maximum value and becomes 95 per cent of the maximum at three times this frequency.

Effect of Circuit Proportions on Amplification Characteristic.—The effect which variations in the circuit proportions have on the characteristics of the impedance-coupled amplifier can be deduced qualitatively from an inspection of the equivalent circuit of Fig. 65*b* and from the series of diagrams given in Fig. 67. Increasing the coupling inductance has the effect of extending the region of full amplification to lower frequencies but does not affect the amplification at higher frequencies, as is shown at Fig. 67*a*. The effect of changes in grid-leak resistance and coupling capacity are shown at Fig. 67*b* and *c* respectively. The value of grid-leak resistance affects the maximum amplification but otherwise has negligible importance, while variations in the coupling-condenser capacity have practically no effect at all until the coupling-condenser capacity is much smaller than would ordinarily be employed. Variations in the shunting capacity C_g, representing the stray-wiring capacity plus tube input and output capacities has much the same effect with impedance coupling as in the case of resistance coupling, as is apparent from Fig. 67*d*. The effect of altering the plate resistance of the amplifier tube is shown at Fig. 67*e* and affects both the amount of amplification and the extent to which the amplification falls off at high and low frequencies.

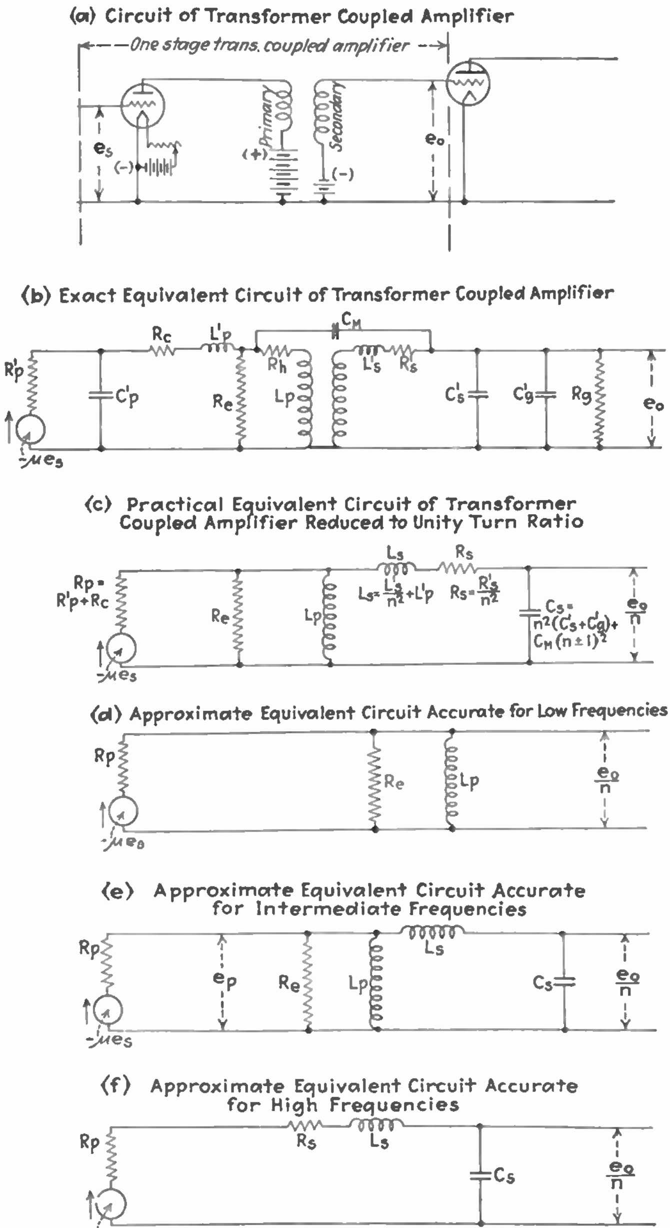
The inductance effective in the equivalent circuit is the inductance which the coupling coil offers to alternating currents and will be greater the larger the alternating-current signal current and the smaller the direct-current plate current, as explained in Sec. 6. The inductance hence tends to be slightly greater at low frequencies where the coil impedance is small (and hence the alternating current great), than at high frequencies where the coil impedance is very large, but for most practical purposes the coil inductance may be considered as constant at the low-frequency value. The coupling inductance must be designed so that the direct-current plate current will not saturate the core, and this requires air gaps at the joints in the magnetic circuit to introduce reluctance that will prevent the normal plate current from producing core saturation. In general it may be said that in every iron-cored inductance coil carrying direct current, there is a particular size of air gap that will give the maximum inductance to superimposed alternating currents, and that this air gap increases in length as the direct current is increased.

The core of the coupling inductance must be assembled from rather thin laminations in order to reduce the eddy-current losses at audio frequencies and should be made of material having high permeability and high electrical resistivity. Silicon steel is the material most commonly used, while some of the high-permeability magnetic alloys, such as permalloy, are resorted to where the design requirements are very severe. The coil winding must be of a type that will have low distributed capacity, for otherwise the response will fall off unduly at high frequencies.

Design of Impedance-coupled Amplifiers.—The design problems of an impedance-coupled amplifier largely center about the coupling inductance. Since the response at low frequencies improves as the coupling inductance is made larger in proportion to plate resistance, it is apparent that with a given inductance the low-frequency response will improve as the plate resistance, and hence the amplification factor, of the tube is decreased. Thus impedance-coupled amplification with high μ tubes gives large amplification but does so at the expense of low-frequency response. The customary procedure is to make the coupling inductance as large as practicable and then use tubes with the highest amplification factor that will meet the low-frequency-response requirements. The use of low-amplification-factor tubes also improves the high-frequency response for the same reasons that were effective with resistance coupling. The grid-leak resistance should have a value at least five, and preferably ten times the plate resistance, in order that the maximum amplification of the impedance-coupled amplifier may approach as nearly as possible the amplification factor of the tube, while the coupling condenser should have a capacity such that the frequency at which the reactance of the coupling condenser equals the grid-leak resistance is somewhat below the frequency at which the reactance of the coupling inductance equals the plate resistance. This latter condition insures that the low-frequency response will not be made appreciably worse by voltage drop in the coupling condenser.

Impedance coupling is superior to resistance coupling in having a somewhat higher maximum amplification per stage and in having an insignificant direct-current voltage drop in the coupling element, permitting the use of a lower plate-supply voltage. These advantages are however largely if not completely offset by the poorer high- and low-frequency-response characteristics and by the comparatively high cost of the coupling inductance. The result is that impedance coupling is not widely used, since resistance coupling is preferred where a uniform amplification over a wide band of frequencies is essential, and transformer coupling is more satisfactory where some frequency distortion can be tolerated.

37. Audio-frequency Voltage Amplifiers—Transformer Coupling.—In the transformer-coupled amplifier the load impedance connected in the plate circuit of the vacuum tube is the primary of a transformer, the secondary voltage of which is applied to the grid of the succeeding tube as shown in Fig. 68. Transformer coupling is superior to resistance and impedance coupling in that the step-up ratio of the transformer permits the amplification to exceed the amplification factor of the tube, and because the output is isolated from the direct-current plate voltage by the transformer secondary without the use of grid leak and coupling condenser. These advantages are, however, gained at the expense of a somewhat less uniform response over a wide band of frequencies, and greater care in the amplifier design and adjustment is required than with resistance and impedance coupling.



70%: $R_p = \omega L_p$

FIG. 68.—Circuit of transformer-coupled amplifier, together with the equivalent circuit and simplifications of the equivalent circuit useful in making amplifier calculations.

The way in which the amplification of a representative transformer-coupled amplifier operated under proper conditions varies with frequency is shown in Fig. 69. The distinguishing characteristics are an amplification that is substantially constant over the major part of the voice-frequency range but which falls off at both high and low frequencies. For purposes of comparison the corresponding characteristics of typical resistance- and impedance-coupled amplifiers are shown dotted in Fig. 69.

Equivalent Circuit of Transformer-coupled Amplifier.—The analysis of the amplification obtained with a transformer-coupled amplifier is

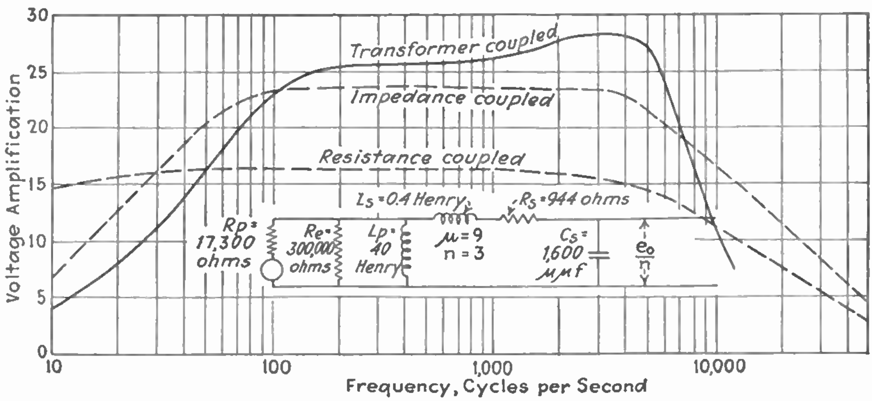


FIG. 69.—Variation of amplification with frequency in typical transformer-, impedance-, and resistance-coupled amplifiers.

somewhat more complicated than with resistance and impedance coupling because the transformer is a more complicated electrical network. The equivalent circuit of the transformer-coupled amplifier taking into account the various factors involved is shown at Fig. 68*b*. The only approximation involved in this circuit is in assuming that the coil capacities can be represented as lumped when they are distributed in fact. The transformer is seen to be a very complicated electrical network, the solution of which involves a great deal of labor, but it is fortunately possible to reduce this exact circuit into the very much simpler network shown at Fig. 68*c* without introducing appreciable errors. The transition from the exact circuit of Fig. 68*b* to the practical circuit shown at Fig. 68*c* consists in neglecting the effects of the distributed capacity of the primary and the plate-cathode capacities of the amplifier tube, in neglecting the resistance resulting from magnetic hysteresis, in reducing the transformer to a unity-turn ratio, in lumping all the leakage inductance on the secondary side of the transformer, and in considering the primary to secondary (*i.e.*, interwinding) capacity C_m as equivalent to a certain other capacity shunted across the secondary of the transformer.[1]

[1] These approximations can be justified as follows: The primary capacity C_p' and the resistance R_h resulting from hysteresis have negligible effect on the behavior of the

When the proper values are assigned to the circuit elements in Fig. 68c this simplified equivalent circuit of the transformer-coupled amplifier will give calculated results that are in error by only a few per cent. If a small additional error can be tolerated, as is generally the case, a further simplification can be made in Fig. 68c by neglecting R_e, *i.e.*, the eddy-current loss in the transformer core. This last approximation makes the calculated amplification slightly high at intermediate frequencies, but inasmuch as R_e is an extremely high resistance in most transformers the error introduced by neglecting it is relatively small.

Calculation of Low-frequency Amplification.—At low frequencies the reactance of the leakage inductance is relatively small while that of the capacity across the transformer secondary is large. This permits the behavior of the transformer-coupled amplifier at low frequencies to be accurately given by the simplified equivalent circuit shown at Fig. 68*d*. A solution of this simplified circuit for the voltage ratio e_o/e_s, which is the amplification at low frequencies, gives the equation

$$\left.\begin{array}{l}\text{Amplification at low}\\ \text{frequencies using}\\ \text{transformer coupling}\end{array}\right\} = \frac{e_o}{e_s} = \mu n \frac{\omega L_p}{\sqrt{\left(\dfrac{R_p R_e}{R_p + R_e}\right)^2 + (\omega L_p)^2}} \qquad (69a)$$

The notation can be understood by reference to Fig. 68. Since the resistance R_e representing eddy currents is normally very high compared with the low-frequency reactance of the transformer primary inductance, it is usually permissible to assume R_e is infinite and if this simplification is made the amplification then becomes

$$\text{Amplification at low frequencies} = \frac{e_o}{e_s} = \mu n \frac{\omega L_p}{\sqrt{(\omega L_p)^2 + R_p^2}} \qquad (69b)$$

An examination of the equivalent circuit of the transformer-coupled amplifier for low frequencies and of Eqs. (69*a*) and (69*b*) shows that the

transformer because of their small size in ordinary circumstances. The reduction to a unity-turn ratio can always be carried out without introducing error provided all impedances on the secondary side of the transformer are divided by the square of the voltage step-up ratio when placed in the equivalent circuit. The leakage inductance can be considered as located entirely on the secondary side because this is where most of it actually is, and because at the high frequencies where leakage inductance becomes important its exact distribution between primary and secondary is not important. The interwinding capacity C_m can be replaced by a suitable capacity located across the secondary as a result of the fact that the voltage difference across this condenser is almost exactly in phase with the voltage developed across the secondary of the transformer up to the frequency at which the secondary capacity is resonant with the leakage inductance, and furthermore this voltage is proportional to the voltage across the secondary but has a magnitude either $(n + 1)$ or $(n - 1)$ (depending on the relative polarity of the two windings) times the voltage across the primary of the transformer.

low-frequency amplification depends primarily upon the magnitude of the transformer primary inductance in relation to the effective plate resistance R_p of the tube and falls off as the frequency is lowered because of insufficient reactance in the primary of the transformer at low frequencies. This low transformer primary reactance results in a large part of the equivalent voltage μe_s, which is acting in the equivalent plate circuit, being consumed by the plate resistance of the tube instead of appearing across the primary of the transformer. In order to obtain good amplification at low frequencies it is therefore necessary that the transformer have a high inductance and that the plate resistance of the amplifier tube used with the transformer be low. The equivalent circuit of the transformer-coupled amplifier at low frequencies is exactly the same as the equivalent circuit of the impedance-coupled amplifier under the same conditions except for the voltage step-up that takes place with transformer coupling. The low-frequency amplification falls off in the same way in both cases and for the same reasons. The extent of the falling off can be estimated from the fact that at the frequency which makes the inductive reactance of the transformer primary equal R_p, the amplification will be almost exactly 70 per cent of the product μn of amplification factor and step-up ratio, and at three times this frequency the amplification will be 95 per cent of μn.

Calculation of Amplification at Intermediate Frequencies.—As the frequency is increased the effective impedance across the primary terminals of the transformer becomes increasingly great because of the higher reactance which the inductance possesses at high frequencies, and because the capacity C_s across the terminals of the transformer is more or less in parallel resonance with the primary inductance. At the frequencies centering about the point at which the capacity C_s is in resonance with the primary inductance L_p the equivalent circuit of the transformer-coupled amplifier approximates that of Fig. 68*e*. While this circuit appears complicated its behavior can be analyzed readily as a result of the fact that when L_p and C_s are in parallel resonance the impedance across the primary terminals of the transformer is substantially equal to the resistance R_e representing the eddy-current loss. The voltage across the primary under these conditions is therefore given by the equation

$$\left.\begin{array}{l}\text{Voltage across transformer primary}\\ \text{when } L_p \text{ and } C_s \text{ are in parallel resonance}\end{array}\right\} = \mu\frac{R_e}{R_p + R_e}$$

The voltage across the secondary of the transformer is not exactly equal to the product of the primary voltage and the step-up ratio of the transformer but is somewhat higher because of resonance taking place inside the transformer between the leakage inductance L_s and the transformer capacity C_s. This causes a resonance rise of voltage which amounts to

$\dfrac{(1/\omega C_s)}{\left(\omega L_s - \dfrac{1}{\omega C_s}\right)}$ times. At intermediate frequencies centering about the parallel resonant frequency of the transformer the amplification is therefore given by the equation

$$\left.\begin{array}{l}\text{Transformer-coupled}\\ \text{amplification at inter-}\\ \text{mediate frequencies}\end{array}\right\} = \frac{e_o}{e_s} = \mu n\left(\frac{R_e}{R_p + R_e}\right)\left(\frac{1/\omega C_s}{\omega L_s - \dfrac{1}{\omega C_s}}\right) \tag{70}$$

In ordinary transformers the amplification at these intermediate frequencies is very nearly equal to μn because, while the factor $R_e/(R_p + R_e)$ has a value that is slightly less than unity, the factor $\dfrac{(1/\omega C_s)}{\left(\omega L_s - \dfrac{1}{\omega C_s}\right)}$ has a value slightly more than unity. With ordinary transformers this parallel resonant frequency where the amplification approximates μn lies between 500 and 1500 cycles.

Calculation of Amplification at High Frequencies.—At frequencies considerably above the point where parallel resonance takes place between the primary inductance and the secondary capacity C_s the equivalent circuit of the transformer-coupled amplifier takes the form shown at Fig. 68*f*, in which the primary inductance and the resistance R_e representing eddy-current losses have been neglected because their impedance is so much greater than that of the shunt formed by the leakage inductance L_s and the capacity C_s. A solution of Fig. 68*f* for e_o/e_s shows that this ratio is given by the equation

$$\left.\begin{array}{l}\text{Transformer-coupled}\\ \text{amplification at high}\\ \text{frequencies}\end{array}\right\} = \frac{e_o}{e_s} = \mu n \frac{1/\omega C_s}{\sqrt{(R_p + R_s)^2 + \left(\omega L_s - \dfrac{1}{\omega C_s}\right)^2}} \tag{71a}$$

Since the equivalent circuit of Fig. 68*f* is a series resonant circuit having a low Q it is convenient to transform Eq. (71*a*) into a form of universal resonance curve in which the amplification at any frequency is expressed in terms of the ratio of this frequency to the frequency at which series resonance takes place, and the value of $\omega L/(R_p + R_s)$, *i.e.*, the Q of the circuit. This transformation results in the following form of Eq. (71*a*):

$$\left.\begin{array}{l}\text{Transformer-coupled amplification}\\ \text{at high frequencies}\end{array}\right\} = \mu n K \tag{71b}$$

where

μ = amplification factor of tube
n = voltage step-up ratio of transformer

$$K = \frac{1}{\sqrt{\left(\frac{m}{\left(\frac{\omega_o L_s}{R_p + R_s}\right)}\right)^2 + (m^2 - 1)^2}}$$

$$m = \frac{\text{actual frequency}}{\text{series-resonance frequency of } L_s \text{ and } C_s}$$

$\omega_o L_s$ = reactance of leakage inductance at series-resonant frequency. The factor K in Eq. (71b) can be plotted as a function of m for any value of $\omega_o L_s/(R_p + R_s)$, with the results shown in Fig. 70. Examination of Fig. 70 shows that the way in which the amplification departs from μn

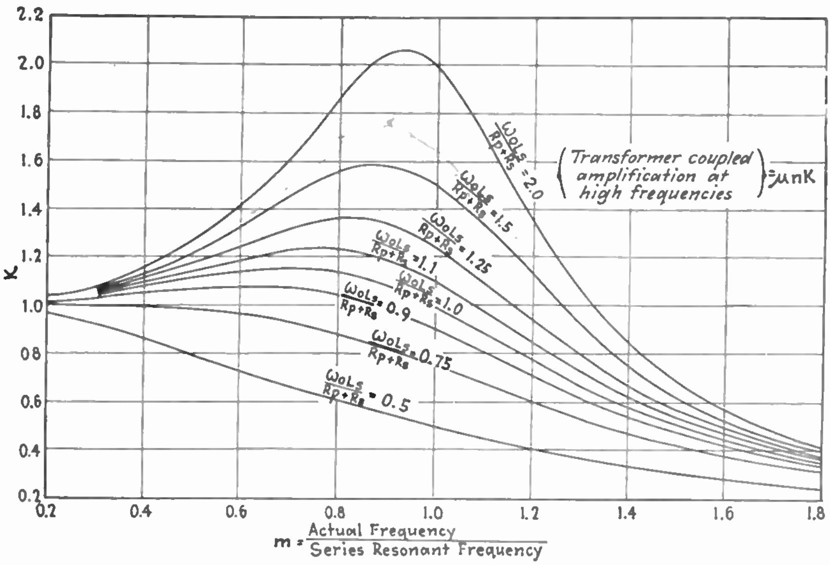


FIG. 70.—Plot of factor K in Eq. (71b) as function of ratio of actual frequency to series-resonant frequency. The transformer-coupled amplification at high frequencies is equal to the product of step-up ratio, amplification factor, and the value of K as read from this curve.

as the series-resonance frequency is approached depends upon the ratio $\omega_o L_s/(R_p + R_s)$ which the circuit possesses at the resonant frequency. If this ratio is appreciably greater than unity the amplification rises to maximum values greater than μn, while if $\omega_o L_s/(R_p + R_s)$ is much less than unity the amplification drops off markedly as the series-resonance frequency is approached. These actions result from the fact that when $\omega_o L_s/(R_p + R_s)$ is greater than unity there is a resonance rise of voltage, while when this quantity is less than unity the resonance rise of voltage becomes a drop because the resistance of the circuit is so great as to more than eliminate all resonance effects. If the amplification is to be substantially constant at the higher frequencies it is necessary that the transformer characteristics be so related to the plate resistance of the tube that the quantity $\omega_o L_s/(R_p + R_s)$ has a value between 0.9 and 1.0.

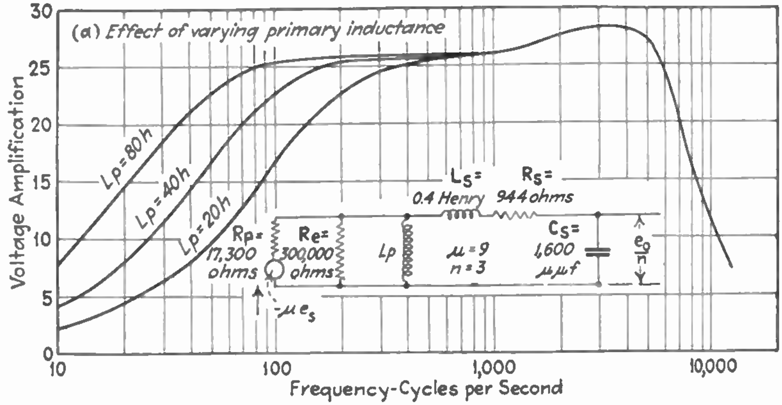


FIG. 71A.

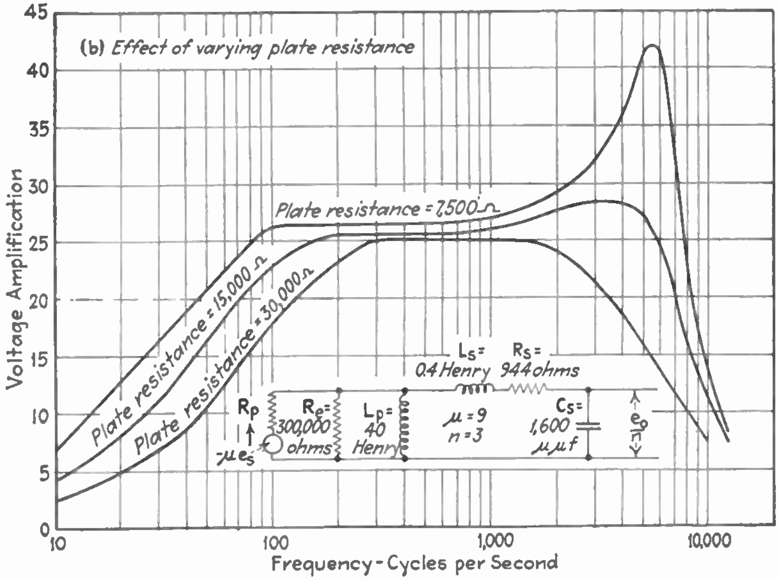


FIG. 71B.

Effect of Circuit Proportions on Amplification Characteristic.—The effect which variations in the circuit constants have upon the performance of the transformer-coupled amplifier can be seen from the series of characteristics given in Fig. 71, and which are typical of the results that can be expected in ordinary cases. Altering the primary inductance of the transformer is shown at Fig. 71*a* to affect only the amplification at low frequencies, and the improvement in the low-frequency response that results from an increase in the primary inductance is clearly evident. At Fig. 71*b* the effect of varying the plate resistance of the amplifier tube is shown. Decreasing the plate resistance improves the low-frequency response in the same way that increasing the primary inductance does

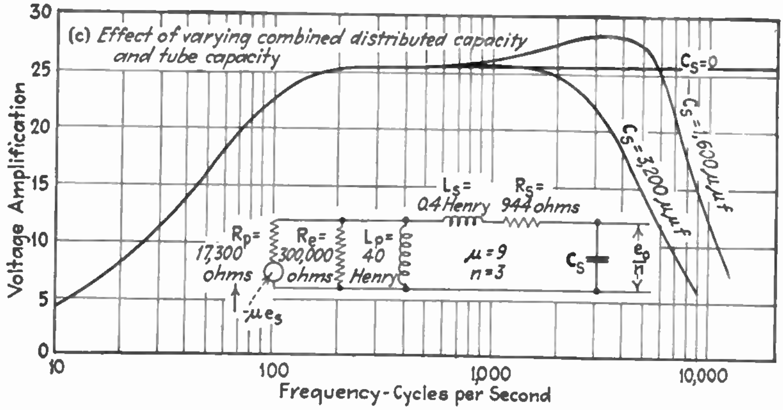


FIG. 71C.

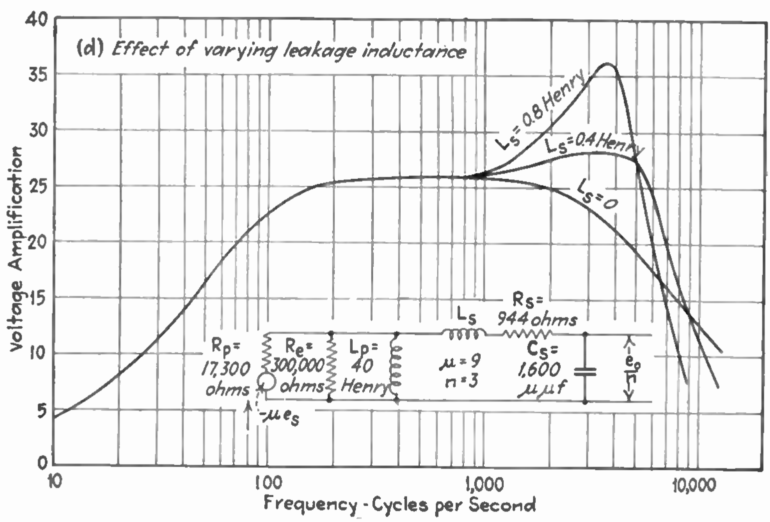


FIG. 71D.

but at the same time causes a resonant rise of voltage at the high frequencies that introduces a large amount of frequency distortion. On the other hand a high plate resistance not only causes poor amplification of the lower frequencies but also reduces the response at high frequencies.

The effect which changes in equivalent capacity C_s and leakage inductance L_s have on the transformer characteristics can be inferred from Figs. 71*c* and 71*d*. Since the highest frequency that the transformer amplifier can amplify satisfactorily even with the most favorable adjustments approximates the frequency at which the leakage inductance L_s is in resonance with the secondary capacity C_s, the high-frequency limit of the transformer decreases as C_s becomes larger, as is apparent from

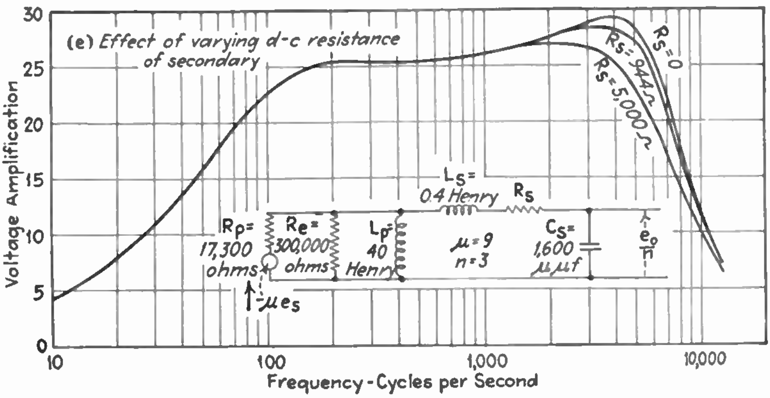


FIG. 71E.

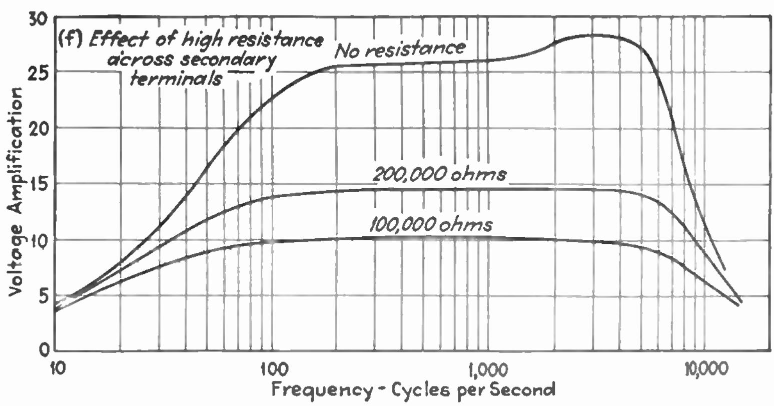


FIG. 71F.

FIG. 71.—Series of curves showing the effects which variations in the circuit constants have upon the way in which amplification varies with frequency in a transformer-coupled amplifier.

Fig. 71*c*. Increasing the leakage inductance increases $\omega_0 L_s/(R_p + R_s)$ and hence increases the resonant rise of voltage at high frequencies, as is evident in Fig. 71*d*.

The resonance effects taking place at high frequencies can be controlled by varying the resistance of the transformer secondary as well as by operating with the proper plate resistance. Figure 71*e* illustrates this effect and shows how increasing the secondary resistance, as might be done by winding the secondary with resistance wire, will reduce the resonance effect at high frequencies. Controlling the resistance in this way can sometimes be used to advantage in the design of transformers which tend to develop a resonant rise of voltage at high frequencies. Another way in which a resonant rise of voltage in the secondary can be eliminated is by placing a high resistance load across the secondary terminals of the transformer. Such a resistance lowers the amplification and at the same time increases the range of frequencies over which the amplification is uniform, as is shown in Fig. 71*f*. The expedient of placing a resistance across a transformer secondary is sometimes employed in poorly designed transformer-coupled amplifiers in order to reduce the frequency distortion, but it is preferable to employ a properly designed transformer in the first place.

Behavior of Transformer-coupled Amplifier Summarized.—The characteristics exhibited by transformer-coupled amplifiers can be summarized by stating that with any ordinary transformer there is some frequency at which the amplification very closely approximates the product of μn of voltage step-up ratio n and tube amplification factor μ, while at low frequencies the amplification is determined by the ratio of primary inductance to tube plate resistance, and at high frequencies the departure of amplification from μn is determined by the relation existing between the tube plate resistance, the leakage inductance, and the equivalent capacity acting across the secondary. The highest frequency that can be amplified without undue frequency distortion approximates very closely the frequency at which the leakage inductance and the secondary capacity are in series resonance, while the amplification begins to fall off rapidly at low frequencies when the frequency is less than the value at which the inductive reactance of the transformer primary equals the plate resistance of the tube. With any given transformer there is always a particular value of plate resistance that must be used if the amplification is to be maintained constant over a reasonable range of frequencies. This is the plate resistance which makes $\omega_0 L_s/(R_p + R_s)$ have a value equal to or slightly less than unity, and with this adjustment the amplification will not be greatly different from μn over a range of frequencies extending roughly from the frequency at which the inductive reactance of the transformer primary equals the tube plate resistance up to the frequency at which there is series resonance in the transformer.

A transformer may be designed to require any value of plate resistance within reason, but in practice high values of plate resistance require low turn ratios, since if more turns are placed on the primary to make the primary inductance keep up with the plate resistance, the secondary will contain fewer turns and the ratio will be correspondingly reduced. Amplifying transformers are generally designed for tubes with amplification factors ranging from 6 to 10, and with plate resistances in the order of 10,000 to 20,000 ohms. There is no advantage gained by the use of higher amplification factors as the lower step-up then obtainable will offset the advantages of a large μ. Moreover tubes with high values of μ have a large input capacity, which reduces the high-frequency response of the amplifier stage preceding the high μ tube.

Determination of Transformer Constants.—The numerical values that must be assigned to the constants in the equivalent circuit of the transformer-coupled amplifier in order to make the computed and measured amplification agree can be determined by making a series of relatively simple measurements upon the transformer.[1] The primary inductance L_p can be measured by an impedance bridge operated at a low frequency and arranged to superimpose a suitable direct-current magnetization on the core. It is essential that a low frequency in the range from 60 to 200 cycles be used in this measurement because at higher frequencies the secondary capacity C_s draws appreciable current and makes the apparent primary inductance as measured somewhat larger than the actual value. The primary inductance will vary with the direct-current magnetization of the core and the alternating current flowing in the transformer windings and will also be affected by the previous magnetic history of the magnetic core. The resistance R_e that represents the effect of eddy currents in the core can be most readily measured by a bridge arranged to measure the impedance across the primary terminals of the transformer at a frequency slightly less than that at which the secondary capacity C_s is in parallel resonance with L_p. Capacity is then added across the primary terminals until parallel resonance is obtained. This makes the primary impedance a pure resistance that approximates R_e very closely.

The step-up ratio of the transformer can be readily obtained by applying a low-frequency voltage directly to the primary terminals and measuring the secondary voltage with a vacuum-tube voltmeter or voltage-ratio bridge. It is necessary to use low frequencies in making this measurement in order to avoid any resonant rise of voltage in the secondary. If the frequency used in this measurement is very low the measured step-up ratio will be too small because an appreciable part of the voltage applied to the primary will be used up by the resistance of the primary winding. This effect can be corrected by multiplying the observed step-up ratio by the correction factor $\dfrac{\sqrt{(\omega L_p)^2 + R_c{}^2}}{\omega L_p}$. The step-up ratio of the transformer will always approximate the turn ratio very closely.[2]

[1] This discussion of the best ways of measuring the characteristics of a transformer is based upon an extensive investigation of this subject made by I. E. Wood while a graduate student at Stanford University.

[2] The ratio of voltage appearing across the secondary to the voltage applied to the primary will be much higher than the ratio of induced voltages in the two windings

The equivalent leakage inductance L_s reduced to a unity turn ratio can be obtained by measuring the impedance at the transformer primary terminals with the secondary terminals short-circuited. This measurement may be made at any convenient frequency and without direct-current magnetization in the core as experiments show that it is substantially independent of the frequency and the state of magnetization of the core. The capacity C_s which can be considered as acting across the secondary terminals of the transformer and which takes into account the distributed capacity of the secondary and the effect of the interwinding capacity is obtainable with fair accuracy by several indirect methods. One way is to measure the voltage ratio of the transformer at one or more moderately high frequencies somewhat less than the resonant frequency of the secondary. The departure of the voltage ratio from the step-up ratio of the transformer at these frequencies is caused by incipient resonance between C_s and L_s and is determined by the following relation, which can be used to determine C_s:

$$\frac{\text{Voltage ratio of transformer}}{\text{Step-up ratio of transformer}} = \frac{(-1)(1/\omega C_s)}{\omega L_s - 1/\omega C_s} \tag{72a}$$

The frequency at which the voltage step-up ratio is maximum is when $1/\omega C_s = \omega L_s$, so that if a curve of step-up ratio as a function of frequency is obtained then C_s can be readily calculated assuming L_s is known. The secondary capacity C_s can also be determined by applying to the transformer primary a frequency that is somewhat less than the series-resonant frequency of the secondary and then adding sufficient capacity C across the secondary to make the voltage ratio reach a maximum (*i.e.*, sufficient capacity to bring the secondary into series resonance at the frequency involved). The secondary C_s can then be determined from the relation

$$\omega L_s = \frac{1}{\omega(C_s + C)} \tag{72b}$$

These methods of measurement will give values of C_s that agree fairly well, and the choice between them is determined primarily by the measuring equipment that is available.

Considerations Involved in the Design of Amplifier Transformers.— In the design of amplifier transformers the objectives to be sought are a large primary inductance, a high turn ratio, and a high series-resonance frequency. The first of these factors gives a good response at low fre-

(*i.e.*, than the step-up ratio) when the frequency is anywhere near the series-resonant frequency of the secondary. The magnitude of the resonance effect may be so great at the secondary resonance frequency as to make the observed ratio as much as 20 times the step-up ratio of the transformer.

quencies, the second a high amplification, and the third a good response at high frequencies. A large primary inductance requires many turns in the primary winding, but there are practical limits to the extent that one may go in this direction. If the winding space is limited the increase in primary turns can be gained only at the expense of secondary turns, with a resulting lower step-up ratio and reduced amplification. On the other hand increasing the number of secondary turns in proportion to the increase in the primary to maintain the step-up ratio constant as the primary inductance increases tends to increase the secondary leakage inductance, thereby lowering the series-resonant frequency and reducing the frequency range of the transformer. It is possible to reduce the leakage inductance by interleaving the primary and secondary windings in alternate sections, or pies, but this increases the interwinding capacity C_m between primary and secondary and so tends to counteract the increase in resonant frequency that would otherwise be obtained. As a consequence of these factors a wide frequency range can be obtained only by employing a low turn ratio (such as 2 or 3) and being content with a low amplification per stage.

It is possible to determine the characteristics of an amplifying transformer with fair accuracy by calculation based upon design data.[1] While such calculations are empirical and so depend to a considerable extent upon the experience and judgment of the computer, they are extremely useful in aiding the designer of audio-frequency amplifier transformers to improve his product. By properly manipulating the factors, such as core dimensions, shape and space arrangement of windings, air gap in the magnetic circuit, etc., it is possible for a skillful designer materially to increase the frequency range of a transformer.

The best transformers obtain a good low-frequency response by using a large number of turns wound on a magnetic core having a large cross section and provided with an air gap sufficient to prevent undue saturation from the direct-current plate current flowing through the transformer primary. The secondary resonant frequency is then made high by using a low turn ratio (2 or 3 to 1) and by properly sectionalizing the secondary winding to give a low distributed capacity and a low leakage inductance. If the secondary capacity C_s is to be kept small it is important that the secondary terminal leading to the grid of the output tube be on the outside of the secondary winding where it is remote from adjacent grounded objects. When well designed, a transformer-coupled audio-frequency amplifier will give a substantially constant voltage amplification of fifteen to thirty times over the frequency range extending from 50 to 7500 cycles. The use of a smaller core or higher turn ratio or both will reduce the frequency range. Examples of the

[1] See Glenn Koehler, The Design of Transformers for Audio-frequency Amplifiers with Preassigned Characteristics, *Proc. I.R.E.*, vol. 16, p. 1742, December, 1928.

amplification-frequency characteristic of a number of typical transformer-coupled amplifiers are given in Fig. 72 and show what may be expected.

The load capacity of a transformer-coupled amplifier (*i.e.*, the maximum voltage that can be handled) is limited by the fact that the grid of the input (*i.e.*, amplifier tube), and of the tube to which the amplified voltage is delivered, must never go positive. If sufficient grid bias is

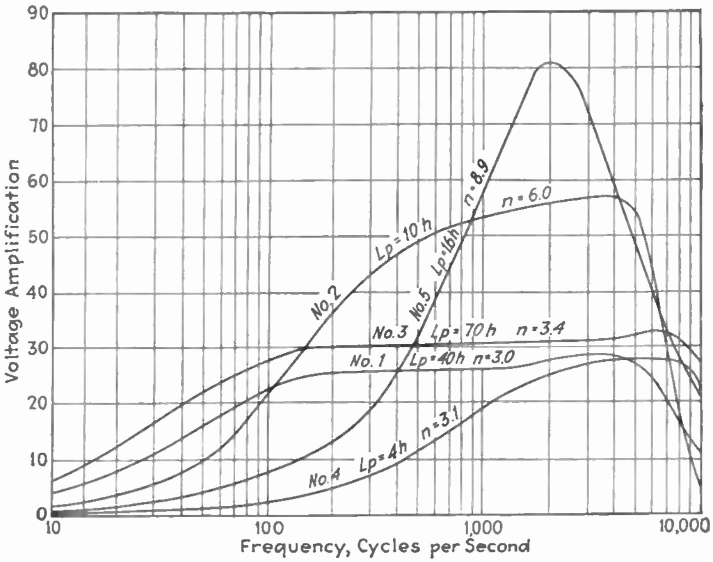


FIG. 72.—Variation of amplification with frequency in a number of typical transformer-coupled amplifiers having different values of primary inductance and different turn ratios.

employed to prevent both grids from becoming positive no trouble is ordinarily experienced from non-linear distortion in transformer-coupled amplifiers. The grid bias that should be used is fixed by the plate resistance that must be present to make the amplification substantially constant up to the series-resonant frequency of the transformer, and will increase as the plate voltage is increased.

Transformer-coupled amplifiers are the most widely used type of audio-frequency amplifiers. They have the advantages of operating satisfactorily on general-purpose tubes having moderate amplification factors, of giving substantially constant amplification over the range of audio frequencies that is most important to speech and music, of requiring no grid leak and grid condenser, and of giving more amplification per stage than other types of amplifiers. Transformer coupling will yield results at least equal to those ordinarily obtained by impedance coupling, and gives way in favor of resistance coupling only where very low frequencies must be amplified.

38. Miscellaneous Types of Audio-frequency Amplifiers.—In addition to the resistance-, impedance-, and transformer-coupled amplifiers that have been described, there are a number of other coupling methods occasionally used, the most important of which are shown in Fig. 73. The arrangement at *a* constitutes an autotransformer and has characteristics similar in all important respects to a transformer-coupled amplifier. The only difference is that no separate primary is employed, which results in a certain economy of materials but requires the use of a grid leak and a grid condenser.

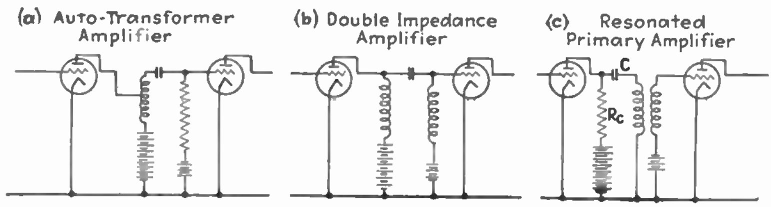


FIG. 73.—Several types of occasionally used audio-frequency amplifiers.

In all amplifiers employing a grid leak it is possible to replace the grid leak by a high inductance, as shown in the impedance-coupled amplifier of Fig. 73*b*. Such an inductance has the advantage of avoiding the problems resulting from a leaky grid condenser but is seldom used because it is more expensive than the grid-leak resistance and at the same time causes the response to fall off at the lower frequencies.

In the amplifier circuit shown at Fig. 73*c* the plate current is prevented from passing through the transformer primary by the condenser *C*, which is of such size as to be in series resonance with the primary inductance of the transformer at an audio frequency that is approximately the frequency at which the inductive reactance of the primary is equal to the plate resistance of the tube. The plate current passes through the high resistance R_c that is in shunt with the transformer and condenser. This arrangement is a combination resistance- and transformer-coupled amplifier, and has the advantage over the transformer-coupled amplifier of giving a slightly more desirable low-frequency response. At the same time, an extra high plate voltage supply is required to make up for the voltage drop of the plate current in the coupling resistance R_c, and as a consequence this arrangement, while having found a certain amount of favor, has not come into general use.

39. Power Amplifiers.—The stage of amplification that delivers the power output of an amplifier is known as the power stage, and the tube employed is called the power tube. There are two different types of requirements that the power stage of amplification can be adjusted to meet. One of these has as its object the production of the maximum possible undistorted power output that can be obtained when a given I

Class A

signal voltage acts on the grid, while the other is the production of the maximum undistorted power output that can be obtained from the tube irrespective of the amount of signal voltage required to develop this output. The differences between the adjustment for maximum power output with a limited signal voltage and the adjustment for maximum undistorted power output without regard to the required signal voltage are in the load impedance and the grid-bias voltage employed.

II

When it is desired to obtain the maximum power output that can be developed with a given signal voltage applied to the grid, the load impedance in the plate circuit of the amplifier should have a magnitude equal to the plate resistance of the tube and should approximate a resistance load as nearly as is practicable. The power output that is obtained from an amplifier operated with a load impedance having a magnitude equal to the plate resistance R_p, and a power factor of cos θ, is given by the equation[1]

I

$$\left.\begin{array}{l}\text{Power output when load}\\ \text{impedance equals plate resistance}\end{array}\right\} = \frac{\mu^2}{R_p}\frac{\cos\theta}{2 + 2\cos\theta}e_s^2 \tag{73a}$$

where

e_s = the signal voltage applied to the grid of the power tube

μ = the tube amplification factor.

When the load is a pure resistance, $\cos\theta = 1$, and Eq. (73*a*) reduces to

I

$$\text{Maximum possible power output from } e_s = \frac{\mu^2}{R_p}\frac{1}{4}e_s^2 \tag{73b}$$

An examination of Eqs. (73*a*) and (73*b*) shows that the maximum possible power output is proportional to the factor μ^2/R_p, which is therefore the

[1] These statements and equations can be readily proved by considering the equivalent circuit of an amplifier having a load impedance Z_L. The amplified signal current produced in the plate circuit by a signal voltage e_s applied to the grid of the tube is

$$\text{Amplified signal current} = \frac{-\mu e_s}{R_p + Z_L}$$

The volt-amperes delivered to the load impedance is the product of the load impedance Z_L and the square of the amplified signal current

$$\text{Volt-amperes delivered to load impedance} = \mu^2 e_s^2 \frac{Z_L}{[(R_p + Z_L)^2]}$$

With a load of constant power factor and varying magnitude the maximum energy will be delivered to the load when its impedance is such as to make the load volt-amperes a maximum. Differentiating the expression for load volt-amperes with respect to Z_L and equating equal to zero shows that the maximum volt-amperes and hence maximum power are obtained when Z_L has a magnitude equal to R_p. With this load impedance the power consumed in load is

$$\text{Load power when } [Z_L] = R_p = \frac{\mu^2 e_s^2}{R_p}\frac{\cos\theta}{[(1 + \cos\theta)^2 + \sin\theta^2]} = \frac{\mu^2 e_s^2}{R_p}\frac{\cos\theta}{2 + 2\cos\theta}$$

figure of merit of a power tube when operated to give maximum possible power output from a given signal voltage. In view of the fact that the power output obtainable is inversely proportional to the plate resistance of the power tube, the plate resistance should be kept low by employing the lowest value of bias voltage that can be used without permitting the grid to become positive.

While the maximum possible output with a given signal voltage is obtained with a resistance load equal to the plate resistance of the tube, the loss in output resulting from a load resistance that does not exactly match the tube resistance is relatively small. This is evident from Fig. 74, which gives the relative power output as a function of the ratio of load to plate resistance and shows that the output is at least 90 per cent of the maximum possible value when the load resistance ranges between approximately one-half and twice the optimum value.

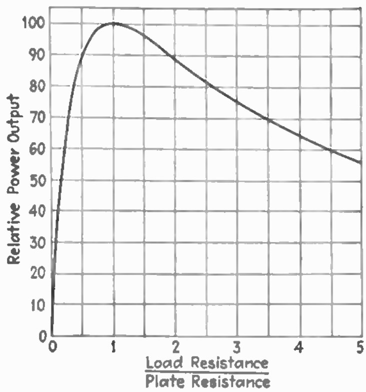


FIG. 74.—Relative power output for a fixed signal voltage shown as a function of the ratio $\frac{\text{load resistance}}{\text{plate resistance}}$. The maximum output is obtained when load and plate resistances are equal.

Adjustment for Maximum Undistorted Output.—In amplifiers delivering large quantities of output power the principal problems are concerned with the avoidance of amplitude distortion and can be understood by studying the voltage and current relations existing in a tube when a large signal voltage is applied to the grid. Consider the case of an amplifier tube having characteristics such as given in Fig. 75, and operated at a grid-bias and plate voltage that place the operating point at the spot marked O. When there is no load impedance in the plate circuit a signal voltage applied to the grid causes variations in plate current that fall along the $E_g - I_p$ characteristic curve of the tube that is marked a in the figure. The effect of a load impedance is to cause the plate current to follow a line such as b, having a slope that is less than that of the tube characteristic a by an amount depending upon the load resistance. The path b is known as the *dynamic characteristic* because it gives the behavior of the tube under actual working conditions, *i.e.*, when an alternating voltage is applied to the grid and there is a load impedance in the plate circuit. The dynamic characteristic does not follow the tube characteristic a because when a load impedance is present the voltage at the plate of the tube differs from the voltage of the plate supply by the drop in the load impedance. Thus when the signal is on the positive half cycle the plate current increases above the average value, causing a voltage drop in the load resistance that reduces the potential actually applied to the plate to something less than the plate voltage at the operating point, while during the negative half of the signal cycle the plate current is less than the average value, and the voltage drop in the resistance is such as to make the instantaneous plate voltage somewhat more than the plate voltage at the operating point. The dynamic characteristic therefore has a slope that is less than the slope of the static characteristic, and in fact has a slope that corresponds to a resistance equal to the plate resistance of the tube plus the load resistance.

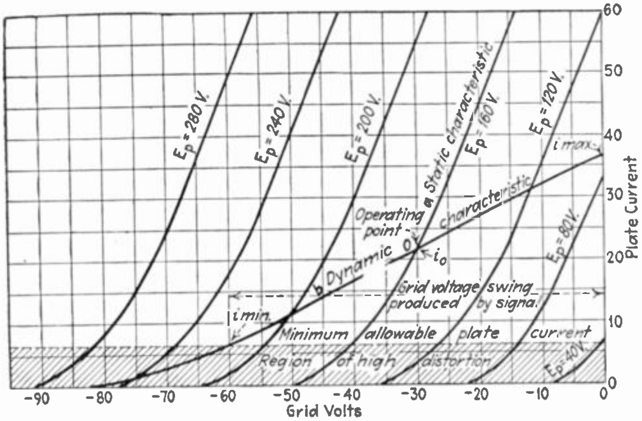


FIG. 75.—Characteristic curves of typical power-amplifier tube, showing region of high distortion, together with dynamic characteristic for a load resistance that gives proper operating conditions.

In order to avoid amplitude distortion, the instantaneous grid potential must never be allowed to become positive, and the dynamic characteristic of the tube must approximate a straight line over the entire operating range. The dynamic characteristic obtained with a resistance load always approximates a straight line very closely except in the region where the plate current is small, and, since this part of the dynamic characteristic is curved, the minimum instantaneous plate current in a power tube should never be allowed to fall below a certain value if amplitude distortion is to be avoided. In Fig. 75 the approximate location of this high-distortion region into which the operating point should not extend is indicated by the shaded area. The allowable crest value of signal voltage is limited on the one hand to a value equal to the grid bias if the grid is not to become positive, and on the other hand to a value that will not extend the operating range into the shaded area of Fig. 75, in which the dynamic characteristic is curved excessively. The maximum power output is obtained from the amplifier when the operating conditions are such that a sine-wave signal voltage having sufficient amplitude to bring the grid voltage to zero during the positive half-cycle just barely reduces the instantaneous plate current to the minimum allowable value at the crest of the negative half-cycle. In order to obtain this condition it is necessary that proper balance exist between plate-supply voltage, grid-bias voltage, load resistance, and plate resistance.

Determination of Proper Grid Bias and Load Resistance of Power Amplifiers.—While an exact mathematical treatment of the relations existing in the power tube is extremely difficult because of the slight curvature in the dynamic characteristic, it is possible to determine the proper operating conditions and the undistorted power output that can be expected from a power tube, with an accuracy that is sufficient for all practical purposes, by considering that the curves showing the relation between grid voltage and plate current are straight lines for values of plate current greater than the shaded region of Fig. 75. Such a characteristic, an example of which is shown in Fig. 76, is subject to mathematical treatment and at the same time does not differ very greatly from the actual characteristic. To the degree of accuracy which this affords the proper grid bias with a resistance load is given by the equation[1]

[1] This equation, together with those that follow, can be derived in the following manner based on the idealized characteristic curves of Fig. 76. The fundamental requirement is that the grid bias E_c be such that when a signal voltage of amplitude E_c is at its negative crest (instantaneous grid bias $= -2E_c$) the instantaneous plate current will just reach the shaded area of Fig. 76. Since this value of plate current is obtained with a plate voltage of E_0 at zero grid bias, it will be obtained at a grid bias of $-2E_c$ when the instantaneous plate voltage E_p exceeds E_0 by an amount $2\mu E_c$. At the same time E_p exceeds E_B by the crest value of voltage drop produced in the load resistance when the signal voltage $-E_c$ acts on the grid, which drop has the value $\mu E_c R_L/(R_L + R_p)$. Incorporating these relations into a single equation gives

$$E_p - E_0 = 2\mu E_c = E_B + \mu E_c \frac{R_L}{R_L + R_p} - E_0$$

When solved for E_c the result is Eq. (74).

Equation (75) is obtained by solving Eq. (74) for R_L, and Eq. (76) is the result of noting that the power output is equal to the square of the crest voltage developed across the load divided by twice the load resistance. Since the crest voltage developed across the load is $\mu E_c \frac{R_L}{R_p + R_L}$, then

$$\text{Undistorted power output with proper grid bias} = (\mu E_c)^2 \frac{R_L}{2(R_L + R_p)^2}$$

When the value of E_c given by Eq. (74) is substituted in this relation the result is Eq. (76). Under ideal conditions in which $E_0 = 0$, the plate current at the operating point is

$$I_p = \frac{(E_B - \mu E_c)}{R_p}$$

$$\text{Proper negative grid bias of power tube} = \left(\frac{E_B - E_0}{\mu}\right)\left(\frac{R_L + R_p}{R_L + 2R_p}\right) \qquad (74)$$

where

E_B = plate voltage at operating point
E_0 = plate voltage required to develop the minimum plate current that will exist during signal cycle when grid-bias voltage is zero
R_p = plate resistance of tube at operating point
R_L = load resistance
μ = amplification factor of tube.

This value of grid bias locates the operating point so that a sine-wave signal voltage reaches the limits of zero instantaneous grid voltage on

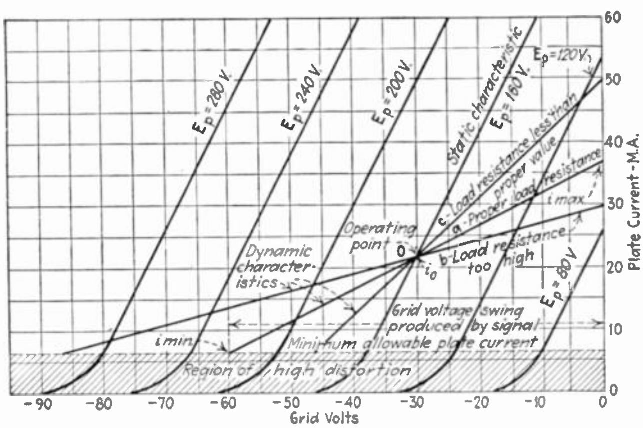


FIG. 76.—Idealized characteristic curves used to determine proper operating conditions for power amplifier, together with idealized dynamic characteristics for the proper load resistance and for load resistances greater and less than the value giving maximum power output.

the positive half-cycle and minimum allowable plate current on the negative half-cycle simultaneously, and places the operating point where the power output is the maximum that can be obtained. The proper value of grid bias increases as the load resistance and the plate-supply voltage become greater.

---

and the power supplied to the plate by the plate voltage source is I_pE_B, or

$$\text{Input power} = \frac{E_B(E_B - \mu E_c)}{R_p}$$

The ratio of undistorted power output given by Eq. (76) to this input power is the plate efficiency and can be reduced to Eq. (78) by substituting for E_c from Eq. (74) and simplifying.

Equation (74) can be rewritten to give the load resistance that should be used with a given grid bias, with the result that

$$\left.\begin{array}{l}\text{Load resistance to be}\\ \text{used with grid bias } E_c\end{array}\right\} = R_L = R_p\left(\frac{E_c}{\left(\frac{E_B - E_0}{\mu}\right) - E_c} - 1\right) \tag{75}$$

In this equation the signs are chosen so that a negative grid bias gives a positive E_c. The effect of varying the load resistance when the grid bias is fixed is shown by dynamic characteristics a, b, and c of Fig. 76. Characteristic a is for the value of resistance called for by Eq. (75) and has a slope such that the operating point O is exactly midway between the point where the dynamic characteristic enters the region of high distortion that exists with low plate currents, and the point where the grid becomes positive. Curve b shows the dynamic characteristic when the load resistance is higher than the value called for by Eq. (75), and it is apparent that in this case the grid will go positive with a signal that does not carry the operating point anywhere near the minimum allowable plate current. Similarly curve c is a dynamic characteristic for a lower load resistance than called for by Eq. (75), and it is seen that a sine-wave signal having an amplitude sufficient to carry the grid to zero potential during one half-cycle will cause the operating point to extend into the high-distortion region during the other half-cycle. The full possibilities of the power tube for the particular grid bias shown are therefore realized only in case a. A grid bias that is either higher or lower than that called for by Eq. (74) has much the same effect as varying the load resistance, since the proper load resistance varies with the bias.

Formulas for Calculating Performance of Properly Adjusted Power Amplifier.—The undistorted power output that can be obtained from a power amplifier adjusted to the proper grid bias for the load resistance being used is given by the equation

Also $P = \frac{(I_{max} - I_{min})(E_{max} - E_{min})}{8}$

$$\left.\begin{array}{l}\text{Undistorted power output}\\ \text{with proper grid bias}\end{array}\right\} = (E_B - E_0)^2\frac{R_L}{2(R_L + 2R_p)^2} \tag{76}$$

The load resistance which gives the maximum possible undistorted power output that can be obtained from the tube at a given plate voltage is obtained by differentiating Eq. (76) with respect to load resistance, equating the result to zero, and solving for load resistance. This operation shows that *the maximum undistorted power output is obtained with a load resistance that equals twice the plate resistance of the tube.* Substituting these relations in Eq. (74) shows that the grid bias for this load resistance should be

$$(\text{Grid bias when } R_L = 2R_p) = \frac{3}{4}\left(\frac{E_B - E_0}{\mu}\right) \tag{77a}$$

Substitution in Eq. (76) shows that when the load resistance is twice the plate resistance the undistorted power output is

$$\text{Maximum possible undistorted power output} = \frac{(E_B - E_0)^2}{16R_p} \qquad (77b)$$

The power that the plate-supply voltage must furnish to the tube is equal to the product of plate voltage and plate current at the operating point, and is unchanged by the presence of a signal. When no signal is applied to the grid of the tube, all of this energy is dissipated at the plate and appears as heat that must be radiated, while in the presence of a signal the plate loss is reduced by an amount equal to the amplifier output. The ratio of the maximum power output that can be obtained from a tube to the power that must be supplied to the plate from the voltage-supply source is called the plate efficiency and cannot exceed 50 per cent even under the most favorable circumstances. Under ideal conditions, when the grid-voltage plate-current characteristic curves are straight lines without a curved region at small values of plate current (shaded area in Fig. 75 reduced to zero) the plate efficiency at maximum output is given by the equation

$$\text{Ideal plate efficiency} = \frac{\text{maximum power output}}{\text{power supplied to plate}} = \frac{R_L}{2(R_L + 2R_p)} \qquad (78)$$

where

R_L = load resistance

R_p = plate resistance of tube.

When the adjustment is for maximum possible power output, the load resistance is twice the plate resistance, and the ideal plate efficiency is 25 per cent. The plate efficiency is less for lower resistance loads and will approach 50 per cent only when the load resistance is very high compared with the plate resistance of the tube. These figures represent ideal cases, and the efficiencies actually realized will be somewhat less because of the unusable high-distortion region at small plate currents.

Adjustment When Plate Dissipation Is the Limiting Factor.—In tubes where the plate voltage is in the order of 1000 volts or more, the operating conditions corresponding to maximum undistorted power output will ordinarily lead to a power dissipation at the plate that exceeds the safe value. Under these conditions the plate current should be reduced to a safe value by increasing the negative grid bias. The load resistance is then chosen in accordance with Eq. (74) to give the best operating conditions with this particular grid bias. This procedure is preferable to limiting the plate loss by lowering the plate voltage, since it permits the use of a high load resistance which increases the plate efficiency above the value that would be obtained with the same plate loss at a lower plate voltage and a lower load resistance.

Variation in Performance with Load Resistance.—The undistorted power output obtainable from a properly adjusted power tube operated at a given plate voltage is not particularly sensitive to load resistance, as is apparent from Fig. 77, which gives relative power output as a function of the ratio of load to plate resistance. The relative signal voltages required to develop maximum power output are also shown in Fig. 77. The gain in output obtained by adjusting for maximum possible power output rather than for a maximum output with a given signal is 9/8 times, while in order to realize this greater output the signal must be 9/8 as large as that required to develop maximum output when the load and plate resistances are equal. These differences are not very great, and it might be wondered why the emphasis in power amplifiers is put on obtaining the conditions for maximum undistorted power output, and why it is a customary practice to use a load resistance that is somewhat greater than the plate resistance. The answer is that, in addition to the advantage of a greater power output from a given tube, the adjustment for maximum possible power requires a high resistance load, which reduces the curvature of the dynamic characteristic and leads to a greater plate efficiency. Thus the adjustment for maximum possible power gives 9/8 times as much undistorted output power as that for maximum power with a given signal and at the same time requires less plate power (only 84 per cent as much in the ideal case). The use of a high load resistance reduces the harmonics because the curvature of the dynamic characteristic is caused by changes in the resistance formed by the plate resistance in series with the load resistance, and a high load resistance makes the variations of plate resistance that occur over the operating range of less relative importance in the combination.

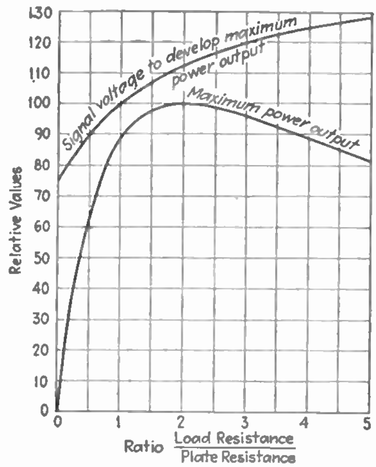


FIG. 77.—Relative value of undistorted power output obtainable with optimum grid bias as a function of the ratio $\frac{\text{load resistance}}{\text{plate resistance}}$, together with relative values of signal voltage required to develop the undistorted output available.

Determination of Exact Dynamic Characteristic and the Calculation of Distortion.—While it is possible to determine satisfactorily the proper grid bias and the undistorted power output that is available, by assuming the characteristic curves over the operating region to be straight lines as shown in Fig. 76, an evaluation of the non-linear distortion requires the use of the actual dynamic characteristic since it is the curvature of this characteristic that gives rise to the distortion. The procedure for deriving the actual dynamic characteristic is as follows: One point on the dynamic characteristic is the operating point (*i.e.*, the point corresponding to the plate voltage and grid bias that exist with no signal), where the plate voltage and plate current can be represented by E_p' and I_p', respectively. Other points can be determined for a load resistance R_L by noting that the plate current I_p'' at which the dynamic characteristic crosses the $E_p - I_p$ static curve for a plate voltage of E_p'' is given by the relation

$$R_L(I_p'' - I_p') = (E_p' - E_p'') \tag{79}$$

This is equivalent to stating that the voltage drop produced in the load resistance by the difference between the plate current at the operating point and the plate current at the point in question is the amount by which the plate voltage at the point in question differs from the plate voltage at the operating point. By use of Eq. (79) the dynamic characteristic can be calculated point by point for a resistance load, with a result such as is shown in Fig. 75.

The dynamic characteristic gives the relationship between the actual plate current and the voltage applied to the grid, and so can be used to determine the wave shape of the amplifier output. When the dynamic characteristic is a straight line over the operating range, the wave shape of the output is exactly the same as that of the signal and no distortion is produced provided the grid is not allowed to go positive, but if the characteristic is curved the output is not proportional to the signal, and amplitude (*i.e.*, non-linear) distortion results. If the operating range does not extend appreciably into the region of small plate current which is shown shaded in Fig. 75, the amount of distortion is small, and with a sine-wave signal voltage is made up almost solely of second harmonic. When this is true the percentage of second harmonic voltage in the amplified power output is readily calculated from the dynamic characteristic in terms of the plate current at the operating point, and at the positive and negative crests of the signal voltage, by the relation[1]

$$\left.\begin{array}{l}\text{Percentage of second harmonic}\\ \text{voltage in output}\end{array}\right\} = \frac{\frac{1}{2}(I_{\text{max.}} + I_{\text{min.}}) - I_p'}{I_{\text{max.}} - I_{\text{min.}}}100 \tag{80}$$

where

I_p' = plate current at operating point

I_{max} = maximum instantaneous plate current during cycle of signal voltage

I_{min} = minimum instantaneous plate current during cycle of signal voltage.

[1] This equation is discussed by E. W. Kellogg, Design of Non-distorting Power Amplifiers, *Trans. A. I. E. E.*, vol. 44, p. 302, 1925. Also see J. C. Warner and A. V. Loughren, The Output Characteristics of Amplifier Tubes, *Proc. I.R.E.*, vol. 14, p. 735, December, 1926.

When the distortion is small, *i.e.*, less than 15 to 20 per cent, the percentage of second harmonic is almost proportional to the amplitude of signal voltage. The amount of distortion when maximum power output is being delivered increases as the adjustment causes the minimum plate current (as represented by the upper edge of the shaded area of Fig. 75) to be decreased, and the minimum allowable plate current should accordingly be chosen in accordance with the amount of distortion that can be tolerated, which is usually a second harmonic voltage that does not exceed 5 to 10 per cent of the fundamental. For distortions of this order of magnitude the percentage of third, fourth, fifth, and higher harmonics is almost zero, but as the distortion is increased the third harmonic component becomes relatively larger and may even exceed the second harmonic.

The non-linearity that results in the production of harmonics also causes a rectification of the signal to take place in the plate circuit. *This rectified current is a measure of the amount of distortion, which can therefore be estimated by noting the change in the direct-current plate current that is produced when a signal voltage is applied to the tube.* This is a method frequently used to check the conditions under which an amplifier is operating and is a very satisfactory means of detecting the presence of distortion.

Use of Output Transformer to Match Load Impedance to Tube.—In many cases it is either not convenient or not possible for the load impedance to which the power is delivered to have the magnitude called for by the plate resistance of the power tube. In such cases the load is connected across the secondary of a transformer, the primary of which is in the plate current of the tube, as shown in Fig. 78*a*. Such a transformer is known as an output transformer and has the effect of transforming impedances in proportion to the square of its turn ratio. If the ratio of primary to secondary is N (step-down ratio $= N$), then an impedance Z across the secondary is equivalent, as far as the primary terminals are concerned, to an impedance N^2Z connected across the primary. When the load impedance is Z_L and an impedance Z_p is desired in the plate circuit of the tube, then the transformer should have a step-down ratio N of $\sqrt{Z_P/Z_L}$. The principal design requirements of the output transformer are that the core must be supplied with sufficient air gap to prevent direct-current saturation with the relatively large plate currents drawn by power tubes, and that there must be sufficient primary turns to give the primary inductance L_p a value that will have an inductive reactance at least equal to N^2Z_L for the lowest frequency at which power is to be transferred to the load. These requirements usually call for a transformer with somewhat more iron and more copper than necessary in a transformer used for audio-frequency voltage amplification.

Shunt (or Parallel) Feed and Series Feed.—In all the amplifiers that have been described, except the transformer arrangement of Fig. 78*a*, the load impedance is in series with the source of plate power so that the d-c plate current flows through the load impedance, and the a-c signal currents flow through the plate power source. This arrangement is called series feed and has the disadvantage of placing the load impedance at a high direct-current potential with respect to the cathode and of requiring the load to carry the d-c plate current. These disadvantages can be avoided by the arrangement shown in Fig. 78*b*, in which the alternating- and direct-current components of the plate current are separated by inserting in series with the plate-voltage supply the inductance L that

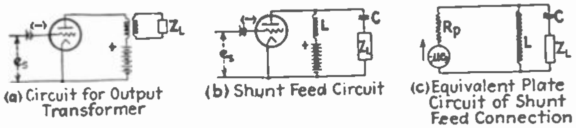


FIG. 78.—Circuits showing: (*a*) load impedance coupled to plate circuit by a transformer, and (*b*) load impedance connected into plate circuit with shunt feed.

has such a high impedance compared with the load as to prevent the signal currents from flowing through the plate-voltage source, while allowing easy passage of direct current, and by placing a condenser C of low reactance to the signal-frequency currents in series with the load in order to prevent the passage of direct current through the load. The coil L is sometimes called a retard coil, while C is spoken of as the blocking condenser.

The characteristics of a tube operated with shunt feed are the same as when series feed is used provided the voltage actually on the plate of the tube when no signal is present is the same in both cases, and further provided that the inductance L offers an extremely high impedance and the condenser C an extremely low impedance to the flow of signal currents. The actual characteristics of a shunt feed arrangement under any conditions can be analyzed in terms of the equivalent plate circuit given at Fig. 78*c*, in which it is seen that as far as the signal-frequency currents are concerned, the actual load impedance effective in the plate circuit of the tube consists of the series combination of the actual load and condenser C shunted by the inductance L.

Effect of Reactance Loads.—The discussion of power amplifiers has so far been confined to resistance loads. When the load has a reactive component the situation is in general about the same as with resistance except that the dynamic characteristic is an ellipse instead of a straight line because the amplified signal current in the plate circuit is not in phase with the signal voltage applied to the grid. The analysis of the case where the load has a reactive component becomes exceedingly complex unless the characteristic curves of the tube are considered as straight lines, and even then the results are too involved to be worth the considerable difficulty of deriving them.[1] Examples of dynamic characteristics for loads having the same magnitudes but different power factors are given in Fig. 79, and it is seen that as the power factor becomes lower the dynamic characteristic tends to extend more and more deeply into the region of high distortion. Since the same effect is obtained with resistance loads that are too small in magnitude, it is apparent that if the high-distortion region is to be avoided with reactive loads, the magnitude

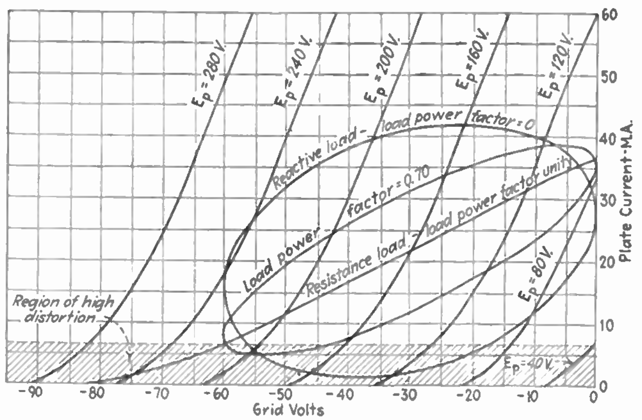


FIG. 79.—Dynamic characteristics for load impedances of varying power factors but constant magnitude, showing how the reactive loads cause the characteristic to open up into ellipses that extend to a lower value of plate current than are reached with the same signal when the load is a resistance.

of the load impedance must exceed the load resistance that would be used with the same grid bias, by an amount that increases as the load power factor becomes less. Stated in another way this means that, as the load power factor is reduced without changing the magnitude of its impedance, the grid bias must be decreased. When the vector impedance of the load changes considerably with frequency the amplifier is adjusted on the basis of the minimum impedance that will be encountered, as this is the condition most apt to produce distortion.

Tubes for Power Amplification.—The type of tube best suited for power amplification differs in its characteristics from the tube best adapted to voltage amplification. Since the plate efficiency of an ideal tube, as given in Eq. (78), depends only upon the ratio of load to plate resistance and not upon the plate voltage or current, it is apparent that the undistorted output available from a power tube in the ideal case is exactly proportional to the plate power and is approximately so in the actual case. This means that a tube having a high power capacity must be operated at a high plate voltage, or at a high plate current, or both.

[1] For an analysis of power amplifiers with reactive loads see Manfred von Ardenne, On the Theory of Power Amplification, *Proc. I.R.E.*, vol. 16, p. 193, February, 1928.

TABLE VI.—CHARACTERISTICS OF REPRESENTATIVE POWER TUBES

Type	Filament				Recommended operating conditions							Plate efficiency, per cent	Amplification factor	Allowable plate loss, watts	Uses
	Volts	Amperes	Watts	Type	Plate voltage, volts	Grid bias, volts	Plate current, milliamperes	g_m, micromhos	R_p, ohms	Load resistance for maximum undistorted output, ohms	Undistorted output, milliwatts				
231	2.0	0.13	0.26	Oxide coated	135	22.5	8.0	875	4,000	......	170	15.7	3.5	.....	Where filament power must be as low as possible
171-A	5.0	0.25	1.25	Oxide coated	90 180	16.5 40.5	12 20	1,330 1,620	2,250 1,850	3,200 5,350	125 700	11.6 19.5	3.0	About 5.0	Power tube in radio receivers
245	2.5	1.5	3.75	Oxide coated	180 250	33.0 48.5	25 34	1,850 2,000	1,900 1,750	3,500 3,900	780 1,600	17.3 18.8	3.5	About 15	Most commonly used power tube in radio receivers
250	7.5	1.25	9.2	Oxide coated	250 450	41.0 80.0	28 55	1,800 2,100	2,100 1,800	4,300 4,350	1,000 4,600	14.3 18.6	3.8	About 30	In radio receivers delivering large outputs and in public-address systems
211	10	3.25	32.5	Thoriated tungsten	1,000	55	72	3,530	3,400	6,000	10,000	13.9	12	75	Public-address systems and modulation of small oscillators
849	11	5.0	55	Thoriated tungsten	3,000	132	100	6,000	3,200	17,500	100,000	33.3	19	300	Modulation of oscillators
848	22	52	1,144	Tungsten	10,000	1,000	750	3,300	2,400	......	2,150,000	28.6	8	7,500	Water-cooled tube for modulation of large oscillators

NOTE: All values of grid bias except those for the three largest tubes are for direct current on the filament.

In the smaller power tubes, such as those employed in radio receivers, the plate voltage available is limited by practical requirements, and a large power output can be obtained only by increasing the plate current at this fixed plate voltage. This calls for a cathode with high electron emission and an amplification factor that is low enough to enable the limited plate voltage to produce the relatively strong electrostatic field in the vicinity of the cathode that is needed to draw the large space current. A tube with a low amplification factor requires a large grid bias, as indicated by Eq. (74), and this means that a large signal voltage must be applied to the grid to develop the full output of the tube. This disadvantage represents the price paid to obtain a large power output with a limited plate voltage. Where the plate voltage is not limited it is possible to obtain a large undistorted output by using a high plate voltage, and in such cases tubes with moderately high amplification factors can be employed.

The characteristics of a number of power tubes of different sizes are shown in Table VI. It is to be noted that the small tubes have a low amplification factor and require relatively large filament power, while the larger tubes, designed for operation at high voltages, have a higher amplification factor and a somewhat smaller filament (and hence less electron emission) in proportion to the output. The operating conditions given in this table are those published by the tube makers and will be found to agree very satisfactorily with Eqs. (74) to (77) even though these equations are based on an idealized straight-line tube characteristic.

CHAPTER VI

TRIODE AMPLIFIERS—*Continued*

40. Multistage Audio-frequency Amplifiers.—The wiring diagram of a typical multistage audio-frequency amplifier is given in Fig. 80, which shows two stages of transformer-coupled voltage amplification followed by a power tube, making three stages in all. It will be noted that common filament, grid bias, and plate batteries (or other energy sources) are employed, with the exception that an added grid bias and a somewhat higher plate voltage are provided for the power stage by booster batteries.

In designing a multistage amplifier the first step is to select a power tube that is capable of giving the amount of undistorted power that

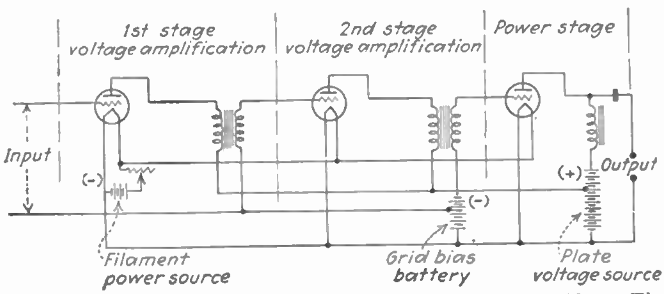


FIG. 80.—Diagram of typical multistage audio-frequency amplifier. The amplifier illustrated consists of two stages of transformer-coupled audio-frequency voltage amplification followed by a power tube with shunt feed in the plate circuit.

it is desired to obtain from the amplifier, and to determine the alternating voltage that must be applied to the power tube to give this output. The ratio of this voltage to the signal voltage available gives the amount of voltage amplification that must be provided and determines the design of the voltage amplifiers.

It is generally found that the total amplification of a multistage amplifier is not exactly equal to the product of the amplifications developed by the individual stages when acting as separate single-stage amplifiers. This is because of what is called regeneration, which is the name given to the effects produced by the transfer of energy between stages. This transfer can take place through stray electrostatic and magnetic couplings between stages, through the electrostatic capacity existing between the grid and plate electrodes of the amplifier tubes, or as a result of impedances common to two or more stages of amplification.

Stray couplings, either magnetic or electrostatic, between amplifier stages can alter the characteristics of a multistage amplifier greatly. There is enormously more power in the amplifier output circuits than at the input, and it requires only a small coupling between the output and input to transfer back to the input an amount of energy that is comparable with the signal energy to be amplified. Electrostatic coupling is usually the result of improper circuit arrangements and can be eliminated either by shielding the parts causing trouble or by rearranging the wiring in such a way that portions of the amplifier at widely different power levels are physically as far apart as is possible. The only magnetic coupling of importance is that which may exist between transformers or coupling inductances as a result of leakage flux, and this can be reduced to a negligible amount by separating the affected parts and by properly orientating the windings. The regenerative effects produced by electrostatic and magnetic couplings are directly proportional to the frequency, but in properly designed audio-frequency amplifiers are of negligible importance at even the highest audio frequencies.

Energy is transferred between the output (plate) and input (grid) circuits of an amplifier tube through the grid-plate interelectrode tube capacity as a result of the voltage difference between the grid and plate electrodes. This action is considered in detail in Sec. 43 and has the effect of giving the amplifier an equivalent grid-cathode (or input) impedance that depends upon the frequency and the load impedance in the plate circuit. The result is that the input impedances of the various tubes in the complete amplifier will differ from the input impedances of the stages considered separately, and this affects the amplification characteristic of the multistage combination by an amount which can be accurately computed by the methods discussed in Sec. 43 and which begins to be of importance at the higher audio frequencies.

Regeneration Caused by a Common Plate Impedance.—In amplifiers where the different stages receive their plate voltage from a common source, as is the case in Fig. 80, the internal impedance of this voltage source is common to the different stages and hence provides a coupling common to the plate circuits of the amplifiers involved. This results in regeneration that may either increase or decrease the amplification, the exact effect depending on the phase relations. A common plate impedance representing the internal impedance of a source of plate voltage is usually the most important cause of regeneration in a carefully built amplifier, and produces effects that are of such importance as to warrant detailed consideration.

When the plate current of more than one amplifier tube flows through the same common impedance, the voltage drop which the currents from one stage develop across the common impedance will transfer energy to all of the other stages. The part of this action that is important in determining the characteristics of a multistage amplifier is the transfer of energy from the plate circuit of the last stage of amplification to the plate circuit of the first stage. The energy transferred between plate circuits of other than the last and first tubes is relatively unimportant because the difference in energy level between the end tubes is much greater than between any other pair. The result is that the effect which a common plate-circuit impedance has on the behavior of a multistage amplifier can be analyzed with a high degree of accuracy by considering that the impedance of the voltage source is common to only the first and last tubes fed from this source.

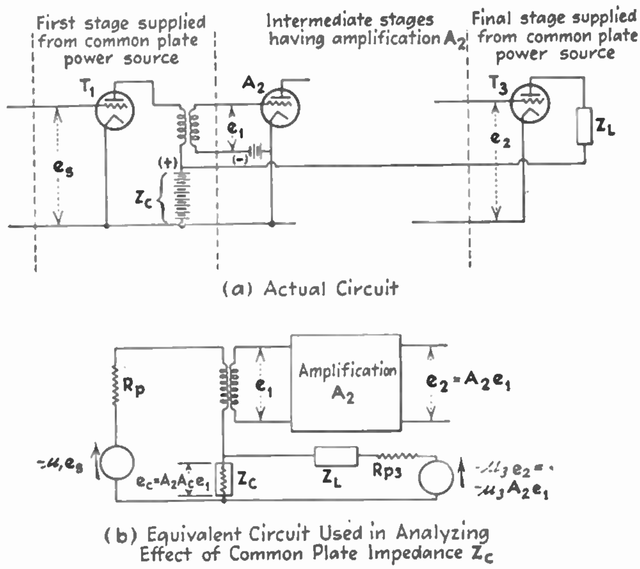


FIG. 81.—Circuit of multistage transformer-coupled amplifier having a common plate impedance resulting from the use of a common source of plate power, together with equivalent circuit that can be utilized to analyze the effect of the common plate impedance.

Analysis of Regeneration Produced by a Common Plate Impedance.— The analysis of the effect of the common impedance is carried out with the aid of equivalent circuits of the type shown in Figs. 81 and 82, which are for the cases where the first stage is transformer coupled, and resistance coupled, respectively.[1] In both cases the amplified signal currents in the plate circuit of the last tube T_3, flowing through the common plate impedance Z_c, produce a voltage drop e_c that acts in the plate circuit of the

[1] Much of the material in this discussion of the quantitative effects produced by a common plate impedance represents the results of a research carried out at Stanford University under the supervision of the author by D. H. Ring. Among other things this study showed that the effect of a common plate impedance could be accurately calculated by Eqs. (81) and (82).

first tube in addition to the equivalent voltage $-\mu_1 e_s$, and so affects the amplification of the first stage. The common impedance Z_c is always so small (seldom in excess of 100 ohms) that it acts as a generator of negligible internal impedance which inserts a voltage e_c into the plate circuit of the first stage. The voltage e_c represents the voltage drop developed across the common coupling impedance Z_c by the amplified signal e_2 acting on the grid of tube T_3. The magnitude of the effect produced by the common plate impedance is seen to depend upon the size of e_c relative to the equivalent voltage $-\mu_1 e_s$ acting in the plate circuit of tube T_1, and will be greater the more the amplification A_2 between the output e_1 of the first stage and the input e_2 of the last stage, and the larger the common impedance Z_c. The introduction of the feed-back voltage e_c into the plate circuit of the first tube has the effect of changing the ratio e_1/e_s, *i.e.*, regeneration resulting from a common plate impedance changes the effective amplification of the first stage. The amount by which the actual amplification of the first stage differs from the amplification that would be obtained if there were no common impedance is the effect which regeneration from a common plate impedance has on the over-all characteristics of the multistage amplifier.

Formulas for Regeneration with Transformer Coupling.—The action of a common plate impedance is analyzed quantitatively by the aid of equivalent circuits of the type shown in Figs. 81 and 82. In the former case, that of the transformer-coupled amplifier, the feed-back voltage e_c has the value

$$e_c = -\mu_3 e_2 \frac{Z_c}{R_{p3} + Z_L + Z_c}$$

$$e_c = A_2 A_c e_1$$

where

e_c = voltage developed across common plate impedance by amplified signal currents in plate circuit of last tube

μ_3 = amplification factor of last tube

A_2 = amplification between output of first stage and input to last stage

e_1 = amplified signal voltage developed by first stage

$e_2 = A_2 e_1$ = amplified signal voltage applied to grid of last tube

$A_c = \dfrac{e_c}{e_2} = -\mu_3 \dfrac{Z_c}{R_{p3} + Z_L + Z_c}$ = fraction of signal voltage applied to grid of last stage which appears across common plate-circuit impedance.

The voltage e_c is effectively in series opposition to the equivalent voltage $-\mu e_s$ acting in the plate circuit of the first stage, and so is equivalent in its effect to superimposing a voltage $e_c/\mu_1 = A_2 A_c e_1/\mu_1$ upon the grid of the first tube T_1, where μ_1 is the amplification factor of this tube. If A_1 is taken to represent the amplification of stage T_1 when no common plate impedance exists, then

$$\left(e_s + \frac{A_2A_ce_1}{\mu_1}\right)A_1 = e_1$$

Solving this for the ratio e_1/e_s gives the following expression for the actual amplification of the first tube with the effect of the common plate impedance taken into account:

$$\left.\begin{array}{c}\text{Actual amplification of transformer-coupled}\\ \text{first stage taking into account regeneration}\\ \text{from common plate impedance}\end{array}\right\} = A_1\frac{1}{1 - \dfrac{A_1A_2A_c}{\mu_1}} \quad (81)$$

This amplification differs from the amplification A_1 that would be obtained

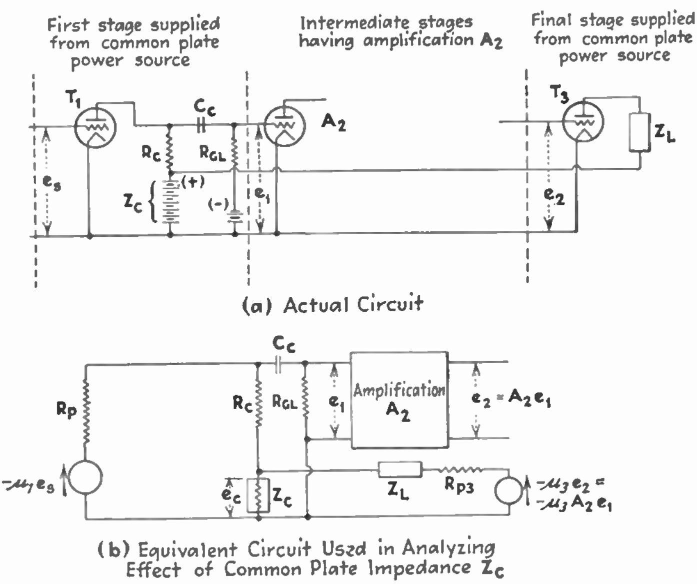


Fig. 82.—Circuit of multistage resistance-coupled amplifier having a common source of plate power, and equivalent circuit for analyzing the effect which the common plate impedance has upon the amplification.

if no feed back were present by the factor $\dfrac{1}{\left(1 - \dfrac{A_1A_2A_c}{\mu_1}\right)}$ which will be small only when $A_1A_2A_c/\mu_1$ is small compared with unity. Since $A_1A_2A_c$ represents the ratio of voltage that would be produced across the impedance Z_c in the plate circuit of the last tube T_3, to the signal voltage e_s applied to the grid of the first tube T_1, if there were no common plate impedance, *it can be concluded that for a common plate impedance to have negligible effect on the amplification of a multistage amplifier when the first stage is transformer coupled, it is necessary that the voltage developed across the common impedance by the amplified signal currents flowing in the plate circuit of the last tube be small compared with* μ_1e_s*, where* e_s *is the* *signal voltage applied to the first tube, and μ_1 is the amplification factor of this tube.* The maximum value of common impedance that can be tolerated is inversely proportional to the total amplification A_1A_2 obtained when the common impedance is zero, and is extremely small when more than two stages of amplification are involved.

Formula for Regeneration with Resistance Coupling.—The effect which a common plate-circuit impedance has when the first stage is resistance coupled can be analyzed by solving the equivalent circuit of Fig. 82 for the ratio e_1/e_s (*i.e.*, for the effective amplification of the first stage when the effect of regeneration is taken into account). The detailed steps in this solution differ slightly from those in the case of transformer coupling because the equivalent circuits are different, and lead to the result:

$$\left.\begin{array}{l}\text{Effective amplification of first}\\ \text{stage having resistance coupling}\end{array}\right\} = A_1 \frac{1}{1 - \dfrac{A_1A_2A_c}{\mu_1}\dfrac{R_p}{R_c}} \qquad (82)$$

where

A_1 = amplification of first stage when there is no common impedance
A_2 = amplification from output of first tube to grid of last tube
A_c = ratio of voltage developed across common plate impedance by amplified signal currents in plate circuit of last tube T_3, to voltage acting on grid of this tube
μ_1 = amplification factor of first tube T_1
R_p = plate resistance of first tube T_1
R_c = coupling resistance employed in plate circuit of first tube T_1.

If impedance coupling is used the results are identical with Eq. (82) except that R_c is replaced by the vector impedance of the coupling inductance. The similarity of Eqs. (81) and (82) means that the effects of the regeneration are of the same character with either resistance or transformer coupling, the only difference being that in the former case the regeneration decreases as the coupling resistance in the plate circuit of the first tube is made large compared with the plate resistance of the tube.

Discussion of Analysis of Regeneration Caused by a Common Plate Impedance.—The factors A_1, A_2, and A_c that appear in Eqs. (81) and (82) are all vector quantities, having magnitudes that represent the magnitudes of amplifications involved, and angles representing the phase shifts of the amplified voltages. The result is that when regeneration is present the denominators of Eqs. (81) and (82) may be either larger or smaller than unity depending upon the phase angle of the factor $A_1A_2A_c$, so that regeneration arising from a common impedance in the plate circuit may either increase or decrease the amplification. In an actual amplifier the magnitude and phase angle of the quantity $A_1A_2A_c$ will be different for different frequencies, with the result that the regeneration will depend upon the frequency, and it is quite possible to have the phase relations such that regeneration increases the amplification at

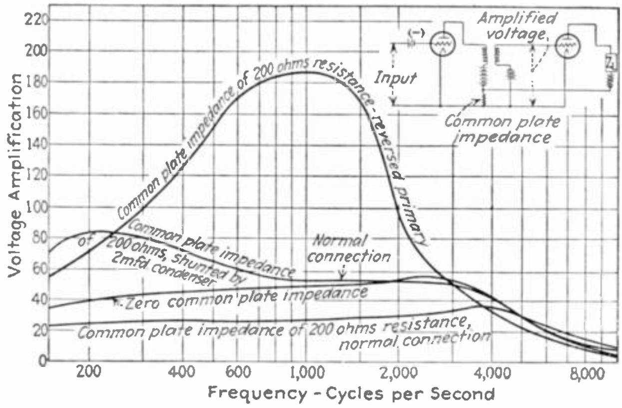


FIG. 83.—Experimental curves showing the effect of a common impedance in the plate circuit on the amplification characteristic of a transformer-coupled amplifier. The effect of the regeneration is to either increase or decrease the amplification according to the polarity of the transformer secondary. Shunting the common impedance by a large condenser eliminates the regeneration at high frequencies while allowing large regeneration to exist at low frequencies.

some frequencies while decreasing it at others. Typical examples of regeneration produced as the result of a common plate-circuit impedance

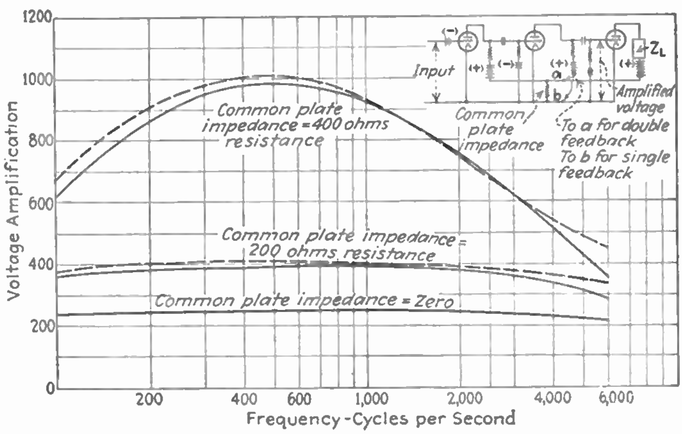
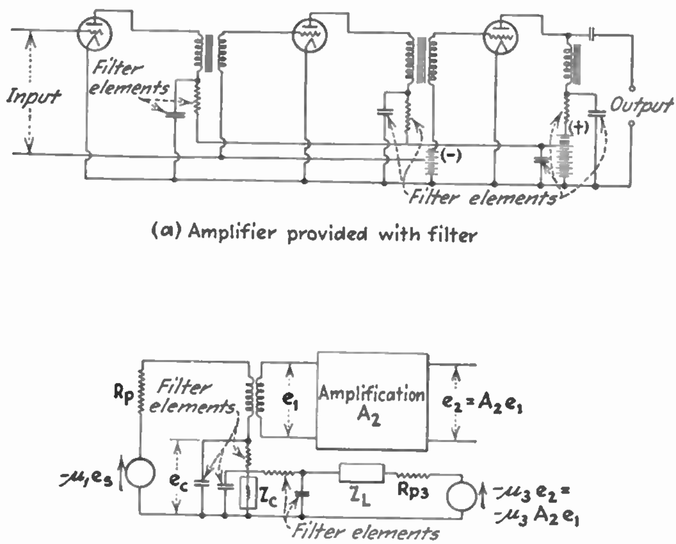


FIG. 84.—Experimental curves showing the effect of a common impedance in the plate circuit on the amplification characteristic of a three-stage resistance-coupled amplifier. The dotted curves are for the case where the impedance is common to only the first and last tubes while the solid curves are for the case where the impedance is common to all three plate circuits. The close agreement between the two is the practical justification for neglecting all energy transfer except that between first and last tubes when making calculations.

are shown in Figs. 83 and 84. The solid curves of Fig. 84 are measured results obtained when the common plate impedance was in the plate circuits of all three tubes, while the dotted curves give the observed behavior when the impedance was common to only the first and last tubes. The close agreement shows that the approximation involved by neglecting all energy transfer except that between the last and first plate circuits is justified in practical work.

When the common impedance is a resistance, the principal effect in resistance-coupled amplifiers is to cause a general increase or decrease



(a) Amplifier provided with filter

(b) Equivalent circuit for analyzing effect of common plate impedance in amplifier with filter

FIG. 85.—Diagram of the amplifier of Fig. 80 after filters have been placed in each plate lead to minimize the effects of common plate impedance, together with equivalent circuit for analyzing the effect of energy transfer from the last stage to first stage through the common plate impedance when filters are present.

of amplification over the entire frequency range. The phase relations are such that with an even number of stages, such as two or four, the total amplification is reduced by the regeneration through the common plate resistance, while with an odd number of stages, as is the case in Fig. 84, the amplification is increased by the action of the common resistance in the plate circuits. With impedance and transformer coupling the effect of a common resistance in the plate circuit of the several stages is to increase the frequency distortion because the phases change much more rapidly with frequency in these types of amplifiers than with resistance coupling. The actual behavior in any case depends upon the number of amplifier stages, transformer polarities, and circuit constants, and can easily lead to excessive frequency distortion, as is made apparent by the results of Fig. 83.

Elimination of Regeneration Caused by a Common Plate Impedance.—The regeneration produced when the plate circuits of several stages of amplification are supplied from the same power source can be eliminated only by making the voltage e_c introduced into the plate circuit of the first tube by the amplified signal currents in the plate circuit of the last tube negligible compared with μe_s, and this requires that the internal impedance of the common power source be kept low. Where the amplification is large, however, even an extremely low impedance will have considerable effect, and in such cases it is necessary to provide series impedances (which may be either resistances or inductances) and shunt condensers in the plate leads of each tube supplied from the common power source, as illustrated in Fig. 85. These series- and shunt-impedance combinations in each plate lead are frequently called filters and act to reduce the voltage that is introduced in a plate circuit of one tube by the amplified signal currents flowing in the plate circuit of another tube, as an analysis makes clear. Thus the greater part of the amplified signal current in the plate circuit of the last tube of Fig. 85 flows to the cathode through the low-impedance path formed by the filter condenser rather than through the higher impedance offered by the filter resistance, with the result that relatively little signal current flows through the voltage source. Furthermore only a small part of this small current reaches the plate circuits of the other tubes because the plate lead to each of these is also equipped with a filter.

It is apparent from these considerations that the filter elements should be so chosen that the reactance of the shunting condensers is low compared with the impedance of the series resistance (or inductance). When the plate leads in a multistage amplifier are equipped with filters, the voltage e_c, which is introduced in the plate circuit of the first tube by the amplified signal currents of the last tube, can be calculated by using a suitable equivalent circuit, such as that of Fig. 85, that takes into account the filters.

It is frequently thought that regeneration resulting from the internal impedance of a common power source can be eliminated by shunting the output terminals of the power source with a condenser having a capacity of several microfarads. This arrangement accomplishes the desired results for the higher audio frequencies where the condenser reactance is low, but fails to do any good at frequencies in the neighborhood of 100 cycles or less where the condenser has a high reactance. The result is that when the common impedance is shunted by a condenser the amplification is normal at the higher audio frequencies but is far from normal for the lower frequencies, thereby giving rise to considerable frequency distortion. A good example of this effect is shown in Fig. 83.

The magnitude of common plate impedance that can be expected depends largely upon the character and condition of the source of plate voltage. Where this potential is supplied from batteries the internal impedance of the power source will be an extremely low resistance when the batteries are fresh, but will increase as the batteries are used, and will be extremely high as the exhaustion point is approached. The result is that oscillations in amplifiers can often be cured by replacing a worn-out plate battery. In plate power sources of the B eliminator type (*i.e.*, where alternating current is rectified and filtered) the internal impedance is usually in the nature of a capacity reactance, and reaches such magnitudes, especially at low frequencies, that special filters in the separate plate leads are practically always necessary and are usually built in the eliminator.

It is usual amplifier practice to employ a common grid-bias battery which introduces another possibility of energy transfer between stages, but since the signal currents flowing in the grid circuit of a properly adjusted amplifier are negligibly small, such a common impedance gives rise to little or no regeneration.

41. Miscellaneous Characteristics of Audio-frequency Amplifiers. *Push-pull Amplifiers.*—When it is desired to obtain a somewhat greater undistorted output than that given by a single power tube, it is possible to put two or more tubes in parallel, or, as an alternative, to connect them in the push-pull circuit shown in Fig. 86. In the push-pull circuit the outputs of the tubes are combined through the output transformer, while the inputs of the two tubes are effectively in series but of opposite phase. This arrangement is preferable to placing the tubes in parallel because it gives smaller distortion, slightly greater output, and no trouble with direct-current saturation in the output transformer. The lower distortion results from the fact that the phase relations in the push-pull amplifier are such that if identical tubes are used the positive and negative half-cycles of the wave of output voltage (or current) will be the same except for sign, *i.e.*, the negative half-cycle of output wave follows through the same sequences of values in a negative way that the positive half-cycle goes through positively. This is shown in Fig. 86 and is true because the two tubes operate so that whatever distortion one tube introduces in the positive half-cycle the other introduces in the negative half-cycle, and *vice versa*.

The result is that in the push-pull connection no even harmonics can appear in the output because even harmonics make the positive and negative halves of the wave have different shapes, and the push-pull connection always balances out such dissymmetries. This elimination of the even harmonics by the balancing action of the two tubes permits the operation of the tubes at a somewhat higher power level than would otherwise be possible, without at the same time introducing excessive distortion. Another feature of push-pull amplifiers is that the d-c plate currents flowing through the two primaries of the output transformer magnetize the core in opposite directions and so do not cause direct-current saturation even when the tubes draw large plate currents. It will also be noted that in the push-pull connection there are no signal currents flowing through the plate power source, which avoids the possibility of energy being transferred from the power stage to other stages through a common plate impedance.

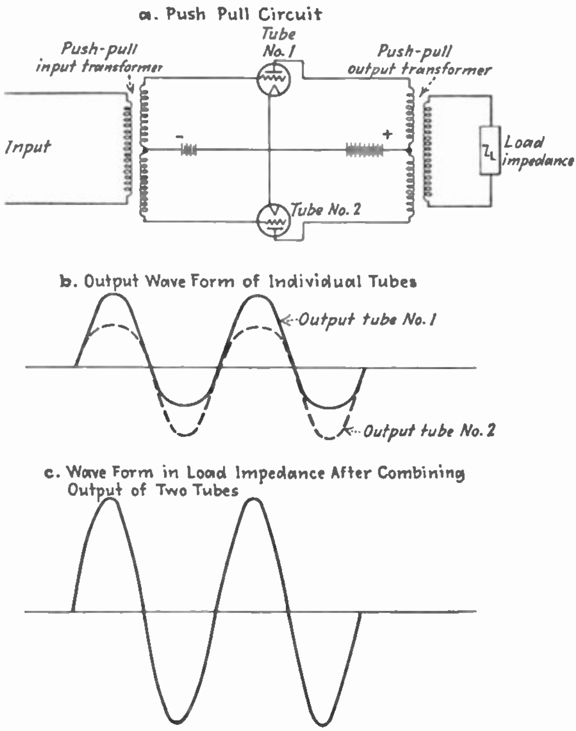


FIG. 86.—Circuit diagram of push-pull amplifier, together with wave shapes produced, showing how the push-pull connection makes the positive and negative halves of the output wave have the same shape, even though this is not true of the outputs of the individual tubes. The result is that the output wave contains no even harmonics and hence suffers less distortion than do the outputs of the individual tubes.

These advantages of the push-pull amplifier ordinarily outweigh the disadvantage of the special input and output transformers that are required, with the result that the push-pull connection is usually employed in preference to the parallel arrangement when two output tubes are required. Push-pull amplifiers are also used where the amplitude distortion must be kept very small, and are capable of giving an almost ideally distortion-free output because of the elimination of the second harmonic, which is the principal and almost sole distortion frequency produced in ordinary amplifiers when not overloaded. The push-pull amplifier will not increase the undistorted output per tube very greatly because a tube operated beyond its normal output produces odd harmonics which the push-pull connection does not eliminate.

Another feature of push-pull amplifiers is that alternating-current components in the plate voltage, such as are always present to some extent in plate power derived from rectified and filtered alternating current, are in phase in the plate circuits of the two tubes, and hence cancel out in the secondary of the output transformer. The result is that rectified plate power need not be so thoroughly filtered when supplied to a push-pull amplifier as when a single tube, or two tubes in parallel, are employed. This makes possible a considerable saving in designing filter systems.

The input transformer used to couple the grids of a push-pull amplifier to the plate circuit of the preceding tube is essentially an ordinary transformer used in voltage amplifiers except for the center-tapped secondary, and the fact that the secondary winding should be arranged in two symmetrical halves. The turn ratio between the full secondary and the primary should be approximately the same as in an ordinary voltage amplifier. The turn ratio between the full primary and the secondary of the output transformer is chosen so that the actual load impedance will be a proper match for an effective plate resistance corresponding to the sum of the plate resistances of the two tubes in push pull. This is because these tubes are effectively in series as far as the a-c currents in the transformer are concerned.

Voltage Surges in Audio-frequency Apparatus.—When the direct current flowing through the primary of an audio-frequency transformer is suddenly interrupted there are high surge voltages developed across the transformer primary and secondary terminals, which are caused by the same action that develops a high voltage across the field of an electrical machine when the field current is suddenly interrupted. This voltage may reach surprisingly large values in the case of ordinary audio-frequency transformers because, although the current interrupted is small, the inductance through which it flows is very great. The potentials obtained under ordinary operating conditions are in the neighborhood of several thousand volts and are often sufficient to produce short sparks in a gap across the transformer secondary. Surges of this sort also occur whenever a lighted tube is removed from its socket, or when stages are switched in and out of the amplifier. The voltage developed by these surges frequently causes the breakdown of insulation in the transformer, with resultant microphonic effects that appear as noises in the amplifier output.[1]

[1] For an extended discussion of surges in audio-frequency amplifiers, see E. H. Fisher, Voltage Surges in Audio-frequency Apparatus, *Proc. I.R.E.*, vol. 17, p. 841, May, 1929. It is shown in this paper that if i_p is the transformer primary current that is suddenly broken, L_p the inductance of the transformer primary, and C_s the

Volume Control.—It is often necessary to control the output power of an audio-frequency amplifier independently of the input voltage. Coarse adjustment of volume, *i.e.*, output power, can be obtained by varying the number of stages of amplification, but finer adjustments

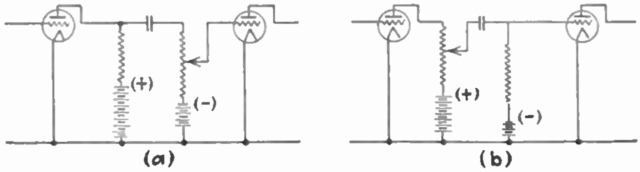


FIG. 87.—Satisfactory methods of controlling volume in resistance-coupled amplifiers without introducing frequency distortion.

require some control on the individual stage. With resistance-coupled amplifiers this control can be obtained satisfactorily by one of the methods illustrated in Fig. 87, while Fig. 88 gives arrangements that are satisfactory with impedance coupling. The amplification of a transformer-

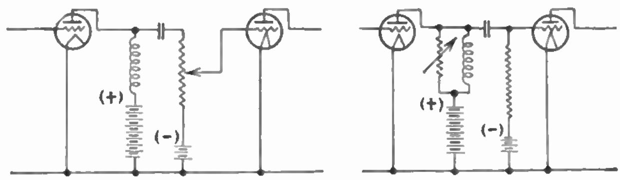


FIG. 88.—Satisfactory methods of controlling volume in impedance-coupled amplifiers without introducing frequency distortion.

coupled amplifier may be controlled in a number of ways, some of which are good while others are very bad. The satisfactory methods are shown in Fig. 89, and will give no more frequency distortion at low power levels

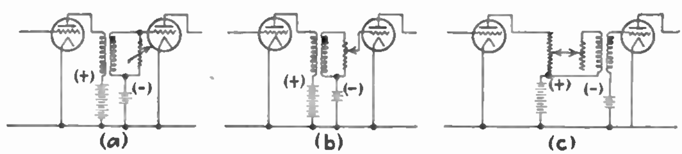


FIG. 89.—Satisfactory methods of controlling volume in transformer-coupled amplifiers without introducing frequency distortion.

than when the volume is maximum. In the case of transformer-coupled amplifiers it is necessary that the resistance across the primary terminals be constant as the volume is changed, for otherwise the frequency distortion will be excessive. Thus a variable resistance across the trans-

---

equivalent capacity that can be considered as acting across the secondary terminals, then the maximum voltage that will be developed across the secondary terminals is given with good accuracy by the relation

$$\text{Voltage across secondary} = i_p\sqrt{\frac{L_p}{C_s}} \tag{83}$$

former primary is not satisfactory, since at low volumes, when this resistance would be small, the amplifier acts as though it were operating with a plate resistance that was too low (*i.e.*, there is an excessive resonant rise of voltage at the high frequencies).

Phase Distortion.—While phase distortion is of little importance with audio-frequency amplifiers, it is interesting to note the type of phase shifts obtained with different types of coupling. The phase shifts for several representative amplifiers are shown in Fig. 90, where it is seen that all of the amplifiers have some phase distortion, since the curves of phase shift are not straight lines passing through zero or some integral multiple of π radians at zero frequency. The phase distortion is ordinarily least for those amplifiers having the lowest frequency distortion, and with any particular amplifier the phase distortion becomes large at frequencies where frequency distortion is pronounced.

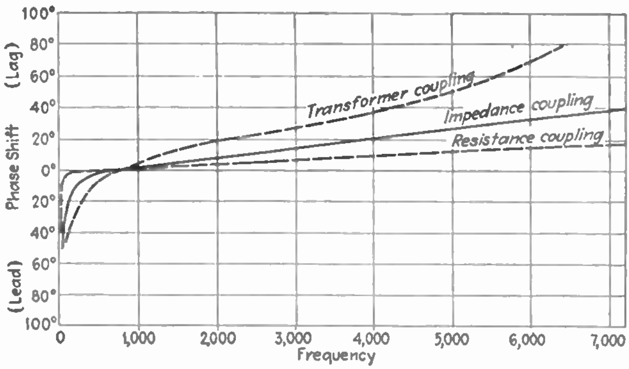


FIG. 90.—Phase shift as a function of frequency for typical amplifiers. Since absence of phase distortion requires that the phase-lag curve be a straight line passing through zero or some integral multiple of π radians at zero frequency, it is apparent that all of these curves show phase distortion.

42. Radio-frequency Voltage Amplification.—Voltage amplification of radio frequencies is carried out using the same general principles that are employed in audio-frequency amplifiers, but with the difference that it is impossible to use ordinary resistance, impedance, or transformer coupling at frequencies in excess of 100 to 200 kc unless special low-capacity tubes are employed. This is because the tube capacities from grid to cathode and plate to cathode have such a low reactance at higher frequencies as practically to short-circuit the high impedance load. It is therefore necessary that the load impedance of a radio-frequency amplifier consist of a resonant circuit in which the various tube and stray capacities assist in tuning to resonance. In this way it is possible to obtain a load impedance that is great enough to give efficient amplification.

An amplifier which utilizes resonance to develop the load impedance is called a tuned amplifier and may exist in several forms, of which the most common are shown in Fig. 91. The circuit shown at Fig. 91*a* is known as the transformer-coupled radio-frequency amplifier since it utilizes a tuned secondary circuit coupled to a primary inductance that is in the plate circuit of the tube. The load impedance in this case is supplied by the impedance which the secondary couples into the plate circuit of the tube. A somewhat different arrangement is shown at Fig. 91*b*, which is known as direct coupling, and in which the load impedance is formed by the parallel impedance of the tuned circuit. This arrangement is employed when the plate resistance of the amplifier tube is comparable with the parallel-resonant impedance of the tuned circuit, and requires a grid leak and grid blocking condenser to isolate the grid of the succeeding tube from the plate voltage. A modification of the direct-coupled circuit is shown at Fig. 91*c*, where the plate connection goes to an intermediate point on the inductance of the tuned circuit. This arrangement is equivalent to transformer coupling but has the disadvantage of requiring a grid leak and grid condenser to isolate the grid of the succeeding tube from the plate-supply voltage. In the circuit shown at Fig. 91*d* the tuned circuit is replaced by a suitable inductance coil, giving what is commonly called an untuned radio-frequency amplifier, but which is actually brought into parallel resonance by the various stray capacities shunting the terminals of the inductance.

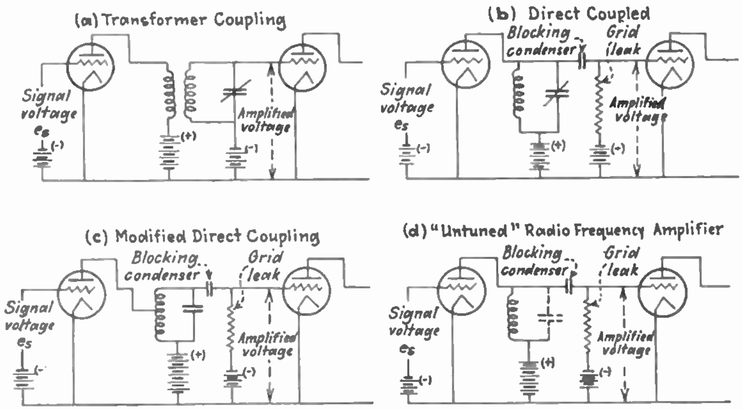


FIG. 91.—The most common types of tuned radio-frequency amplifiers.

Equivalent Circuit of Radio-frequency Amplifier.—The various circuit arrangements for radio-frequency amplification shown in Fig. 91 can be analyzed by equivalent amplifier circuits analogous to those used in connection with audio-frequency amplifiers. Since the transformer-coupled tuned radio-frequency amplifier is by far the most important circuit for radio-frequency amplification, its performance will be analyzed in detail. The exact equivalent circuit of this form of amplifier is shown at Fig. 92*b*. Since the capacity C_p across the primary inductance helps tune the resonant circuit and for all practical purposes acts as though it were in the secondary, the exact equivalent circuit can be simplified to the arrangement shown at Fig. 92*c* with negligible error. A still further simplification can be obtained by considering that the only effect exerted by the mutual capacity C_m between the primary and secondary windings is to help tune the secondary, and results in Fig. 92*d*. This final circuit introduces only relatively small errors and, because it simplifies the analysis greatly, is always used unless very precise results are required.[1]

This final simplified equivalent circuit of the tuned radio-frequency amplifier is the same coupled circuit that was discussed to some extent

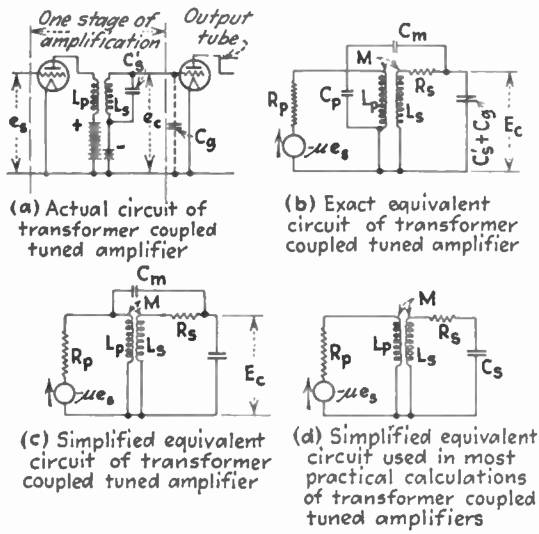


FIG. 92.—Exact equivalent circuit of transformer-coupled tuned radio-frequency amplifier of Fig. 91*a*, together with simplifications of this circuit. The final form shown at *d* gives results that are sufficiently accurate for all ordinary purposes.

in Sec. 17, and can be analyzed by the principles given in Sec. 16. It is shown in Sec. 17 that when a voltage of constant amplitude is acting in the equivalent circuit (*i.e.*, when a constant signal voltage is applied to the amplifier grid) the secondary current will vary with frequency in very much the same way as does the current in a series-resonant circuit, being maximum at the resonant frequency of the secondary and falling off at both higher and lower frequencies. The amplification of the transformer-coupled radio-frequency amplifier (*i.e.*, the ratio of voltage developed across the secondary condenser C_s to signal voltage e_s applied to the amplifier grid) will vary in almost exactly the same way as the secondary current, and so will be a maximum at the resonant frequency

[1] For an exact analysis of a tuned amplifier taking into account the mutual capacity C_m see H. Diamond and E. Z. Stowell, Note on Radio-frequency Transformer Theory, *Proc. I.R.E.*, vol. 16, p. 1194, September, 1928.

of the secondary and will become less as the frequency varies from the secondary resonant frequency. The amplification characteristic of a typical example is shown in Fig. 93.

Formulas Giving Performance of Radio-frequency Amplifier.—The amplification obtained from the transformer-coupled tuned radio-frequency amplifier at the resonant frequency of the secondary is given to a high degree of accuracy by the equation[1]

$$\text{Amplification at resonance} = \frac{\mu(\omega M)Q_s}{R_p + \frac{(\omega M)^2}{R_s}} \tag{84}$$

where

μ = amplification factor of tube
R_p = plate resistance of tube
Q_s = resonant rise of voltage in secondary
$= \omega L_s/R_s$
M = mutual inductance between primary and secondary
R_s = series resistance of secondary
ω = 2π times frequency. re..)

[1] The amplification obtained with a tuned transformer coupling can be calculated using the equivalent circuit of Fig. 92*d* and the principles outlined in Sec. 16. The equivalent impedance Z_p of the primary circuit taking into account the coupled impedance $(\omega M)^2/Z_s$, where Z_s is the series impedance of the secondary circuit, is

$$Z_p = R_p + \frac{(\omega M)^2}{Z_s} + j\omega L_p$$

The primary current I_p produced by the voltage $-\mu e_s$ is then

$$I_p = \frac{-\mu e_s}{R_p + \frac{(\omega M)^2}{Z_s} + j\omega L_p}$$

The voltage which this current induces in the secondary is

$$\text{Induced secondary voltage} = -\mu e_s \frac{j\omega M}{R_p + \frac{(\omega M)^2}{Z_s} + j\omega L_p}$$

The voltage appearing across the secondary condenser is $(1/\omega C_s)/Z_s$ times the induced voltage, or

$$\text{Amplified voltage} = -\mu e_s \frac{j\omega M(1/\omega C_s)/Z_s}{R_p + \frac{(\omega M)^2}{Z_s} + j\omega L_p}$$

The ratio of amplified voltage to signal voltage, *i.e.*, the voltage amplification, is then

$$\text{Voltage amplification} = \mu \frac{\omega M(1/\omega C_s)/Z_s}{\left|R_p + \frac{(\omega M)^2}{Z_s} + j\omega L_p\right|} \tag{85}$$

This expression is an exact solution of the equivalent circuit of Fig. 92*d*. For all practical purposes it can be simplified by neglecting the term $j\omega L_p$ in the denominator, since $j\omega L_p$ is ordinarily negligible in comparison with R_p. Equation (84) is a special case of Eq. (85), obtained by substituting R_s for Z_s and neglecting $j\omega L_p$.

The amplification at resonance depends upon the mutual inductance M and will be maximum when[1]

$$(\omega M) = \sqrt{R_pR_s} \tag{86}$$

This optimum mutual inductance is the value that makes the coupled resistance $(\omega M)^2/R_s$ equal to the plate resistance, and hence is the mutual inductance that satisfies the condition for maximum possible energy delivered to the tuned circuit. With this value of mutual inductance the amplification has the maximum possible value that can be obtained with the tube and resonant circuit being employed, which upon substituting Eq. (86) into Eq. (84) is found to be

$$\text{Amplification with optimum coupling} = \sqrt{\mu G_m}\sqrt{\omega_o L_s Q_s/2} \tag{87}$$

where

μ = amplification factor of tube

$G_m = \mu/R_p$ = mutual conductance of tube

$\omega_o L_s$ = reactance of secondary inductance at resonant frequency, which is also the reactance of secondary condenser at resonance.

$Q_s = \omega L_s/R_s$ for secondary circuit

Effect of Circuit Proportions and Tube Constants upon Amplification.—While the maximum amplification is obtained when the mutual inductance M with which the tuned circuit is coupled into the plate circuit of the tube satisfies the relation given in Eq. (86), this optimum condition is not highly critical, and the loss in amplification resulting from failure to realize exactly the optimum value of M is not great. The ratio of actual amplification at a given value of M to the amplification with optimum M can be readily expressed in terms of the ratio $\dfrac{(\omega_o M)^2/R_s}{R_p}$ and is[2]

$$\frac{\text{Actual amplification}}{\text{Amplification with optimum coupling}} = \frac{2\left[\dfrac{(\omega_o M)^2/R_s}{R_p}\right]^{1/2}}{1 + \dfrac{(\omega_o M)^2/R_s}{R_p}} \tag{88}$$

where

$(\omega_o M)^2/R_s$ = impedance which the tuned secondary couples into plate circuit at resonance

R_p = plate resistance of tube.

[1] Equation (86) can be derived by differentiating Eq. (84) with respect to ωM, equating the result equal to zero and solving for ωM.

[2] Equation (88) can be derived by taking the ratio of actual amplification as given by Eq. (84) to the amplification obtained when Eq. (86) is satisfied. That is

$$\frac{\text{Actual amplification}}{\text{Amplification with optimum } M} = \frac{\mu Q_s \dfrac{(\omega_o M)}{R_p + (\omega_o M)^2/R_s}}{\mu Q_s \sqrt{R_pR_s}/2R_p}$$

This reduces to Eq. (88) upon simplification.

The results of Eq. (88) have been plotted in Fig. 94 and show that as long as the actual mutual inductance does not depart from the optimum

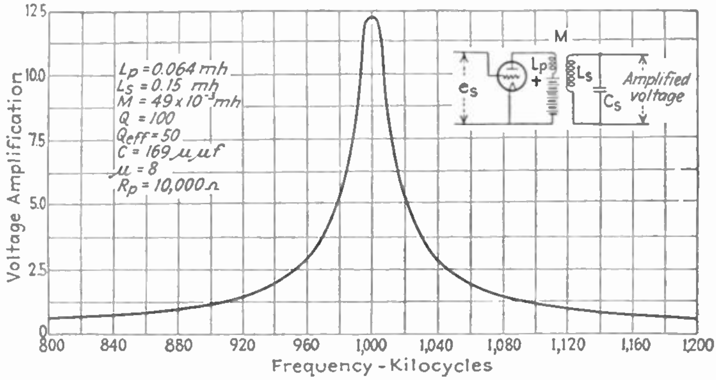


FIG. 93.—Amplification as a function of frequency of a typical tuned radio-frequency amplifier. The shape of the amplification curve is almost exactly the same as that of a resonance curve of a tuned circuit having a Q somewhat lower than the Q of the actual tuned circuit in the amplifier.

value to an extent that makes the coupled impedance $(\omega M)^2/R_s$ more than twice or less than half the plate resistance, the amplification will exceed 94 per cent of the maximum possible value.

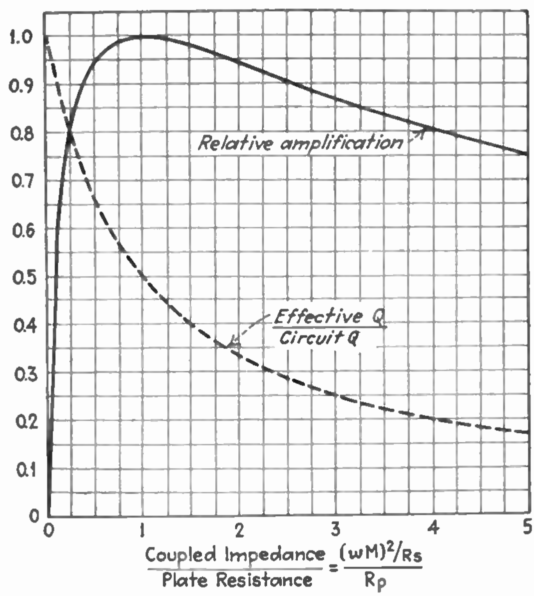


FIG. 94.—Curves showing effect which variations in the ratio of impedance coupled into plate circuit at resonance to tube plate resistance have upon the effective Q and the amplification. These curves are perfectly general and apply to all transformer-coupled tuned radio-frequency amplifiers irrespective of the circuit proportions.

The effect which the tube characteristics and the constants of the tuned secondary circuit have on the amplification obtainable can be found by inspecting Eq. (87). The influence of the tube characteristic is taken into account by the term $\sqrt{\mu G_m}$, which indicates that tubes with high mutual conductance and amplification factors are preferred. There are however certain practical considerations that limit the highest amplification factor that can be used to advantage in triodes to values in the neighborhood of 8 to 10. The effect of the constants of the tuned circuit is incorporated in the term $\sqrt{\omega L_s Q_s}$, from which it is seen that the amplification can be increased by employing a secondary with the highest possible inductance (a high L/C ratio) and the lowest possible resistance.

Selectivity of Tuned Amplifiers.—The tuned radio-frequency amplifier is selective, that is, it amplifies some frequencies very much better than others, as is apparent from Fig. 93. *The curve of amplification as a function of frequency in a tuned amplifier has exactly the same shape as the curve of current in a series-resonant circuit (or of impedance in a parallel circuit), but the effective Q of the amplification curve is always less than the actual Q of the tuned circuit of the amplifier.* The ratio of the effective Q of the amplification curve to the actual Q of the tuned circuit depends upon the ratio of the resistance $(\omega M)^2/R_s$ coupled into the plate circuit by the tuned circuit at resonance to the plate resistance R_p of the amplifier tube, according to the relation[1]

$$\frac{\text{Effective } Q \text{ of amplification curve}}{\text{Actual } Q \text{ of tuned circuit}} = \frac{R_s}{R_s + \dfrac{(\omega M)^2}{R_p}} \tag{89a}$$

$$= \frac{R_p}{R_p + \dfrac{(\omega M)^2}{R_s}} = \frac{1}{1 + \left[\dfrac{(\omega M)^2/R_s}{R_p}\right]} \tag{89b}$$

[1] The derivation of this equation follows: The exact solution for the amplification of the transformer-coupled amplifier having the equivalent circuit of Fig. 92*d* is given by Eq. (85), which for all practical purposes can be simplified by neglecting the term $j\omega L_p$ in the denominator, with the result

$$\text{Amplification} = \frac{(\omega M)(1/\omega C_s)/Z_s}{R_p + \dfrac{(\omega M)^2}{Z_s}}$$

This can then be rearranged to give

$$\text{Amplification} = \frac{(M/C_s R_p)}{Z_s + \dfrac{(\omega M)^2}{R_p}} = \frac{(M/C_s R_p)}{jX_s + \left[R_s + \dfrac{(\omega M)^2}{R_p}\right]} \tag{90}$$

In the last equation the series impedance Z_s of the tuned secondary circuit has been divided into its resistance and reactance components R_s and X_s respectively. The numerator of Eq. (90) is a constant that is independent of frequency, while the denominator is exactly equal to the series impedance of a resonant circuit having the same inductance and capacity, *i.e.*, the same X, as the actual tuned circuit, but with a resistance of $(R_s + (\omega M)^2/R_p)$ instead of the actual circuit resistance of R_s. Since the current in a series-resonant circuit is inversely proportional to the series impedance

The results of Eq. (89*b*) have been plotted in Fig. 94 to give the ratio of effective to actual Q for any ratio of coupled resistance to plate resistance of the tube. This curve applies to all tuned transformer-coupled amplifiers irrespective of the circuit proportions or the resonant frequency of the tuned circuit, and shows that the effective Q decreases as the mutual inductance is increased. At the value of mutual inductance giving maximum amplification, *i.e.*, when $(\omega M)^2/R_s = R_p$, the effective Q of the amplification curve is one-half the actual Q of the circuit.

The fact that the amplification of a tuned radio-frequency amplifier varies with frequency in exactly the same way as does the current in a series-resonant circuit having an effective Q given by Eq. (89) can be used to advantage in determining the amplification as a function of frequency. The procedure is first to calculate the amplification at the resonant frequency of the tuned circuit using Eq. (84). The effective Q of the amplification curve is next determined from the actual Q of the tuned circuit and the ratio of the coupled resistance $(\omega M)^2/R_s$ at resonance to the plate resistance, by the aid of Fig. 94 or Eq. (89). The ratio of actual amplification at any desired frequency to amplification at resonance is then determined by applying the universal resonance curve of Fig. 29 for the case of the $Q_{\text{eff.}}$ just calculated.

The same working rules that were used to estimate the shape of the resonance curve of a circuit are also of value in estimating the shape of the curve of amplification in a tuned amplifier. Thus at a frequency which differs from the resonant frequency by $1/2Q_{\text{eff.}}$ times the resonant frequency, where $Q_{\text{eff.}}$ is the effective Q of the amplification characteristic, the amplification is almost exactly 70 per cent of the amplification at resonance, while at a frequency which differs from resonance by $1/Q_{\text{eff.}}$ of the resonant frequency the amplification is approximately 45 per cent of the value at resonance.

Regeneration in Multistage Radio-frequency Amplifiers.—A single stage of tuned radio-frequency amplification will give a voltage amplification ranging from 5 to 50 times with normal circuit and tube constants. When greater amplification is required a number of stages can be employed. Thus if the output of a single stage having an amplification of 20 is applied to another similar stage, the total amplification will be 400, while a third stage would increase the amplification to 8000. These over-all amplifications as computed on the basis of Eq. (84) represent the results under ideal conditions and are never actually obtained in practice because of regeneration resulting from energy which is trans-

---

of the circuit, the amplification of a tuned radio-frequency amplifier varies with frequency in exactly the same way as does the current in a series circuit having a resistance $(R_s + (\omega M)^2/R_p)$. Since the Q of a resonance curve is inversely proportional to circuit resistance, Eq. (89) follows at once.

ferred from the amplifier output to the input. Energy can be fed back from output to input in a number of ways, the most important of which are the following:

1. Through the grid-plate tube capacity.
2. Stray inductive couplings.
3. Stray capacitive couplings.
4. Through impedances common to different stages.

Energy transferred from one part of an amplifier to a preceding part may cause the amplification to be either greater or less than the ideal value, and the exact effect is equivalent to changing the resistance and the reactance of the tuned circuits in the amplifier. Such variations in reactance alter the resonant frequencies of the tuned circuits slightly and so are relatively unimportant, but the resistance changes that accompany regeneration have a profound effect. This is because the actual resistance of tuned circuits is so low that only a few added ohms will largely destroy all resonance effects, while a few ohms subtracted from the actual circuit resistance can bring the effective resistance to zero, with the result that the amplification becomes infinite, *i.e.*, oscillations are produced. The character of the action in any particular case depends upon the phase and magnitude of the energy fed back to the amplifier input. Regenerative action is always greater as the amplification is increased, since then the extreme difference in the energy levels of different parts of the amplifier is larger. Regeneration also becomes greater the higher the frequency being amplified because capacities pass current and mutual inductances induce voltages in proportion to the frequency.

Methods Used to Minimize Regeneration in Radio-frequency Amplifiers. The magnitude of regeneration in ordinary radio-frequency amplifiers is usually sufficient to alter the characteristics completely unless some special precautions are taken to eliminate the causes of energy feed-back or to compensate for its effects. Regeneration resulting from stray capacitive and inductive couplings between different parts of the amplifier can be reduced by properly positioning these parts of the circuit with respect to each other, and can be eliminated completely by shielding with non-magnetic shields of the type described in Sec. 11. In orientating the various parts of the circuit it is particularly important that the inductance coils of the different stages be as far apart as possible and be arranged with their axes at right angles, in order to minimize mutual inductance. It is also desirable to use coils with small dimensions so that the spacing between coils will be large compared with the coil size, as this reduces the inductive coupling between coils. Electrostatic coupling between parts of the amplifier is not so difficult to control as inductive coupling between coils, and can be kept small by proper arrangement of the wiring. It is impossible, however, to eliminate all electrostatic and magnetic coupling unless each stage of amplification is completely enclosed in a container made of copper, brass, aluminum, or some other good electrical conductor.

Impedances common to two or more stages are particularly troublesome causes of regeneration in radio-frequency amplifiers because at the high frequencies involved a wire only a few inches long will often have sufficient reactance to provide an effective means of transferring energy. The common impedances most frequently encountered in radio-frequency amplifiers are those arising from the use of the same grid-bias and plate-voltage sources for more than one stage. The resulting regeneration can be kept low by shunting the sources of voltage with condensers so large (usually 0.1 to 1.0 μf) as to be substantially a short circuit to the frequencies involved, and can be completely eliminated by equipping the plate leads to the different tubes with filters while at the same time by-passing the grid-bias voltage with a condenser. These filters consist of shunt capacities and series resistances or series inductances, exactly as in the case of the multistage audio-frequency amplifier filters illustrated in Fig. 85, and are for the purpose of preventing radio-frequency currents of different stages from flowing through the same circuit elements.

Regeneration resulting from energy transferred through the grid-plate tube capacity can be eliminated in three-electrode tubes by the use of one of the neutralizing systems described in Sec. 43, or can be avoided by using special vacuum tubes containing an electrostatic shield between grid and plate electrodes. Such tubes are known as screen-grid tubes and are discussed at length in Chap. IX. It is also possible to minimize the worst effects of regeneration resulting from the grid-plate tube capacity by the use of one of the compensating arrangements described in Sec. 43.

When regeneration occurs in a radio-frequency amplifier the observed effect is that by suitable adjustment the amplification will be very much greater than would be obtained if all regeneration was absent, and furthermore the amplifier becomes very much more selective. These two effects are the result of a common cause, namely, neutralization of a portion of the actual circuit resistance, which increases the selectivity by giving the circuit a higher effective Q and increases the amplification because greater resonance effects are obtained with this higher effective Q. A certain amount of regeneration is therefore not a disadvantage because it will increase the amplification and the selectivity, both of which may be very desirable results. The difficulty of using regeneration for this purpose is, however, that the amount of regeneration obtained depends upon the frequency to which the amplifier is adjusted and is also very critical with regard to the exact circuit adjustments, so that for most practical purposes it is preferable to obtain the required amplification without the aid of regeneration wherever this is possible. When regeneration is used, it should be introduced in controllable amounts in a definite way, usually by employing a regenerative detector as described in Sec. 66.

When all feed-back effects have been eliminated in a multistage radio-frequency amplifier, the over-all amplification is then the product of the amplification of the individual stages. The amplification characteristic that results when a number of identical stages perfectly shielded are connected in cascade is a curve having sides considerably steeper than those of an ordinary resonance curve, as is shown in Fig. 95. These results represent an ideal condition that is practically never met with in practice, however, because it is almost impossible to eliminate all regen-

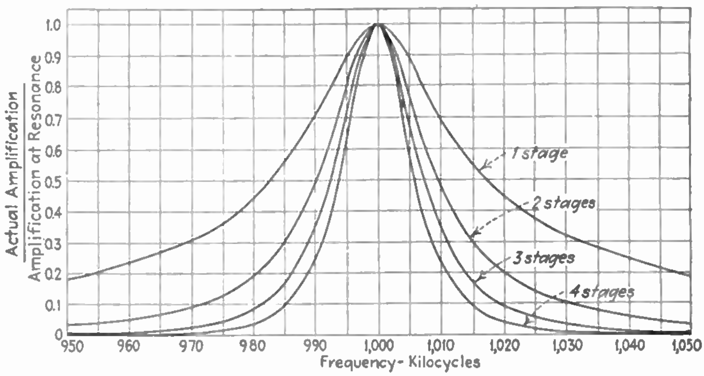


FIG. 95.—Relative amplification as a function of frequency for an amplifier containing different numbers of identical stages and with no regeneration present. The response at frequencies somewhat off resonance decreases greatly as the number of stages is increased, while at the same time the response at frequencies near resonance is relatively much less affected by the number of stages.

eration, even in laboratory equipment designed and built with extreme care, but are of value in that they indicate the general trend of results that can be expected.

Factors Affecting the Design of Radio-frequency Amplifiers.—In the design of tuned radio-frequency amplifiers the arrangement and size of parts, location of wiring, shielding, etc., are determined by considerations relating to the elimination or control of regeneration, as is also the amplification factor of the tube, while the amplification and selectivity obtained are fixed by the characteristics of the tuned load circuit, its coupling into the plate circuit of the tube, and the mutual conductance and plate resistance of the tube. The circuit design is complicated by the fact that the resonant frequency of the usual tuned radio-frequency amplifier is adjustable over a considerable range of frequencies by varying the capacity of the tuning condenser. This introduces problems because the impedance which the tuned secondary introduces into the plate circuit of the tube will vary with the resonant frequency of the secondary and will prevent optimum circuit conditions being realized under all conditions. Another feature involved in the design of tuned radio-frequency amplifiers is that the ideal to be striven for is a large and equal amplification of all frequencies contained in the signal being amplified, and very poor amplification of all other frequencies.

The width of the band of frequencies that is amplified with substantial equality is inversely proportional to the effective Q of the amplification curve, and so can be widened either by using a less efficient tuned circuit (*i.e.*, one with a lower Q) or by increasing the mutual inductance between the tuned secondary and the plate circuit of the amplifier tube. The former method permits the use of a less expensive and smaller coil and is usually considered preferable even though it gives slightly less amplification. It is usually found that adjustments and circuit proportions which give a high amplification per stage also give a very selective amplifier, while adjustments and circuit proportions that will amplify a wide band of frequencies about equally well will give a low amplification per stage, and as a general rule the amplification obtainable is roughly inversely proportional to the width of the frequency band that is amplified. The width of the band of frequencies amplified about equally well is also proportional to the resonant frequency of the amplifier when the effective Q is constant. The result is that in broadcast receivers the selectivity offered to low-frequency broadcast signals tends to be so great as to suppress the higher frequency components of the side bands of such signals, while at the same time the selectivity offered to high-frequency broadcast signals is so low as to permit undesired signals to be received along with the desired frequency.

When the band of radio frequencies to be amplified is extremely wide the best amplifier performance can be obtained by tuning both the primary and secondary inductances of the amplifier separately to the same frequency, as shown in Fig. 96. This arrangement gives a band-pass effect of the type described in Sec. 20, and with proper circuit constants will give substantially constant amplification over a frequency band having a width determined by the mutual inductance between the two tuned circuits. The amplification characteristic obtained from such an arrangement is similar to the resonance curve of the same pair of circuits considered as a band-pass filter. Thus the width of the pass band is determined by the mutual inductance between the two circuits in accordance with Eq. (47*b*), and will be greater the larger the coefficient of coupling. The uniformity of amplification within the pass band is determined by the resistance of the two circuits in relation to the width of the pass band, and can be made substantially constant. The amount of amplification obtained with such a band-pass circuit adjusted to give uniform amplification within the pass band is almost inversely proportional to the width of the band, and will be small if this width is at all large.

Design Procedure.—In designing a transformer-coupled tuned radio-frequency amplifier the first step is to select a secondary circuit that will have the highest inductance that can be used while still allowing the circuit to be tuned over the necessary frequency range. The Q of this secondary circuit should be as high as is possible in view of the maximum selectivity that can be allowed and of the practical limitations imposed by coil design. The mutual inductance between primary and secondary is then given a value that will realize the conditions for maximum possible amplification somewhere near the center of the frequency band to be amplified. In radio-frequency amplifiers intended for code signals the highest selectivity and the greatest possible amplification are desired, so that the tuned circuit should have the highest possible Q and should be coupled into the plate circuit of the tube with a mutual inductance that approximates the optimum value. On the other hand in the amplification of telephone signals, which contain appreciable side bands, the effective Q should vary inversely with frequency in order that the width of the frequency band being equally well amplified will be constant. In broadcast receivers it is furthermore considered desirable to have approximately the same amplification irrespective of the frequency to which the amplifier is tuned.

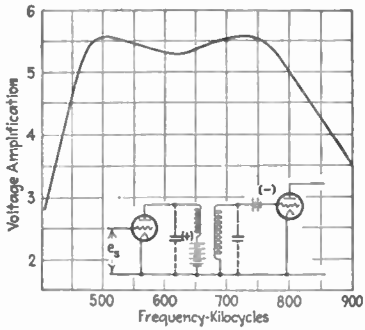


FIG. 96.—Circuit diagram and amplification characteristic of a band-pass amplifier. The amount of amplification per stage is low but is substantially constant over a wide frequency band.

These ideal characteristics for an amplifier of telephone signals are mutually conflicting, so that the actual design is a compromise between them. In practical amplifiers the selectivity is usually allowed to decrease and the amplification to increase somewhat as the resonant frequency is increased. It is theoretically possible to build an amplifier that will have constant amplification and will respond to a frequency band of constant width without using a mutual inductance that is varied with the tuning adjustment, but this requires that the actual Q of the tuned circuit vary with frequency in a way that is virtually impossible to obtain, and furthermore represents an adjustment that does not give the maximum possible amplification.

Miscellaneous Features of Radio-frequency Amplifiers.—The considerations which govern the choice of grid-bias and plate voltage of the amplifier tube are the same for tuned radio-frequency amplifiers as for audio-frequency voltage amplifiers; that is, the negative grid bias should be at least as great as the crest value of the largest voltage to be amplified, but should be no more than necessary, while the plate voltage should be as high as practicable, although there is no great advantage gained by going to extremely high plate voltages. These adjustments insure that the grid of the tube will be maintained negative at all times, and that the plate resistance will be as low as possible.

While the analysis that has been given of the tuned radio-frequency amplifier is for the particular case of transformer coupling, the results obtained with the other forms of coupling shown in Fig. 91 are qualitatively the same. The factors controlling the amount of amplification, the circuit proportions, and the regeneration are fundamentally the same for all types of coupling, but since the transformer-coupled arrangement is most generally employed it has been taken up in detail to show the method of analysis and the outstanding characteristics. When these have been mastered for the case of the transformer-coupled tuned radio-frequency amplifier it is a relatively simpler matter to carry out the analogous procedure with any other type of coupling that may have to be treated.

A radio-frequency amplifier produces phase shifts in the radio-frequency voltages which are similar to the phase shifts that exist in a simple resonant circuit having a Q equal to the effective amplifier Q as defined in Eq. (89). These phase shifts are symmetrical about the resonant frequency of the tuned circuit and follow a path that passes through zero shift at the resonant frequency and which is very nearly a straight line in the vicinity of resonance. Thus referred to the resonant frequency there is little phase distortion for the range of frequencies over which the amplification is relatively large.

43. Input Admittance of Triode Amplifiers.—The admittance between the grid and cathode electrodes of a vacuum tube is called the input admittance of the tube, and takes into account the grid current that flows into the grid-cathode electrode capacity as a result of the signal voltage acting across this capacity, and also the current that flows into the grid-plate interelectrode capacity as a result of the voltage difference between the plate and grid electrodes. This latter component of the grid current depends upon the load impedance in the plate circuit because the alternating-current voltage between plate and grid is the difference between the signal voltage applied between grid and cathode and the amplified voltage developed across the plate load impedance, and the latter obviously depends upon the load. When the plate load impedance is great enough to produce appreciable amplification the potential difference between grid and plate electrodes will be considerably greater than the signal voltage, with the result that a relatively large grid current flows from grid to plate, causing this part of the input admittance of the tube to be very important. If the load impedance in the plate circuit is a resistance the input admittance of the tube is a pure capacity, but if the plate load impedance has a reactive component, the input admittance of the tube will have a resistance component even though the grid is at a negative potential with respect to the cathode and attracts no electrons.

The input admittance of a vacuum tube can be represented by a resistance in parallel with a condenser, as shown in Fig. 97. If the

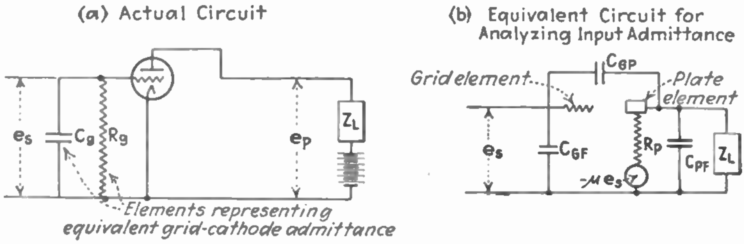


FIG. 97.—Equivalent circuit for analyzing amplifier input admittance. The input admittance is the ratio of current entering the grid electrode to the signal voltage applied to the grid, and depends upon the load impedance in the plate circuit.

ratio of the amplified voltage developed across the load impedance in the plate circuit to the signal voltage being amplified is called A, and θ is used to represent the difference in phase between the voltage across the load impedance and the equivalent voltage acting in the plate circuit, then the resistance and capacity components of the input admittance of an amplifier tube are given by the following formulas.[1]

$$\text{Input resistance} = R_g = -\frac{(1/\omega C_{gp})}{A \sin \theta} \tag{91}$$

$$\text{Input capacity} = C_g = C_{gf} + C_{gp}(1 + A \cos \theta) \tag{92}$$

[1] The derivation of Eqs. (91) and (92) is as follows: Referring to Fig. 97, E_s will be used to represent the signal voltage applied to the grid, and E_p the magnitude of the amplified voltage developed across the load impedance between anode and cathode. The voltage E_p leads $-\mu E_s$ by an angle θ and hence leads E_s by an angle $(\theta + 180°)$. With these definitions the voltage across the grid-plate tube capacity C_{gp} is $(E_s - E_p\underline{/\theta + 180°})$, and the current flowing from grid to plate as a result of this voltage across C_{gp} is $j\omega C_{gp}(E_s - E_p\underline{/\theta + 180°})$. The current flowing from grid to cathode through the grid capacity C_{gf} is $j\omega C_{gf}E_s$, so that the total grid current, which is the sum of these, is

$$\begin{aligned}\text{Total grid current} &= j\omega C_{gf}E_s + j\omega C_{gp}(E_s - E_p\underline{/\theta + 180°}) \\ &= \omega C_{gf}E_s\underline{/90°} + \omega C_{gp}(E_s\underline{/90°} - E_p\underline{/\theta + 270°})\end{aligned}$$

This total current divided by the voltage E_s gives the admittance of the grid, which is therefore

$$\text{Admittance of grid} = \omega C_{gf}\underline{/90°} + \omega C_{gp}\left(1\underline{/90°} - \frac{E_p}{E_s}\underline{/\theta + 270°}\right)$$

The real part of this admittance represents the input conductance (*i.e.*, the reciprocal of the input resistance), while the quadrature part is the input reactance, which when divided by ω gives the input capacity. Equations (91) and (92) are merely these two components of the input admittance with E_p/E_s denoted by the symbol A.

where

C_{gp} = grid-plate tube capacity
C_{gf} = grid-cathode tube capacity
A = ratio of voltage developed across load impedance in plate circuit to applied signal (*i.e.*, A is the amplification of tube alone not taking into account any step-up of voltage)
θ = angle by which voltage across load impedance leads equivalent voltage acting in plate circuit (θ positive for inductive load impedance).

Examination of Eq. (92) shows that with zero plate load impedance the input capacity has the value $(C_{gf} + C_{gp})$ and reaches a value of $(C_{gf} + AC_{gp})$ at very high load impedances (provided no negative resistances are present in the plate circuit). It is to be noted that the input capacity of the tube is dependent only upon the amplification A, the phase shift θ, and the tube capacities, and is independent of frequency, although at high frequencies the same input capacity is more important because it draws a greater current.

The input resistance of the vacuum tube may be either positive or negative, as seen from Eq. (91). A positive input resistance results when the load impedance in the plate circuit is a capacitive reactance, while a negative resistance is obtained with an inductive load in the plate circuit. *A positive input resistance means that energy is transferred from the grid to the plate through the grid-plate capacity, while a negative input resistance indicates that the phase relations are such that energy is transferred from the output or plate circuit of the tube to the grid circuit.* The value of input resistance for the same amplification A and phase shift θ varies inversely as the frequency and may be very low at high frequencies. The minimum value of input resistance that can be obtained without the use of a negative resistance in the plate load occurs when the load impedance is a pure reactance having a magnitude equal to the plate resistance. Under these conditions the input resistance is

$$\text{Minimum possible input resistance} = \pm\frac{(1/\omega C_{gp})}{(\mu/2)} \tag{93}$$

where μ is the amplification factor of the tube, and $(1/\omega C_{gp})$ is the reactance of the grid-plate capacity. When the plate load impedance is a tuned circuit the minimum input resistance is obtained at a frequency that is off resonance by $1/2Q_{\text{eff.}}$ of the resonant frequency, where $Q_{\text{eff.}}$ is the effective Q of the amplification curve. The value of the input resistance at this frequency is

$$\left.\begin{array}{l}\text{Minimum possible input resistance}\\ \text{with tuned load impedance}\end{array}\right\} = \pm\frac{(1/\omega C_{gp})}{(A/2)} \tag{93a}$$

where A is the ratio of voltage developed across the plate load impedance at resonance to the signal voltage applied to the tube (A is the amplification of the tube at resonance). If oscillations are to be avoided in a tuned radio-frequency amplifier it is necessary that the resistance developed between grid and cathode by the parallel-resonant impedance of the tuned circuit supplying the amplifier input be less than the minimum input resistance, as given by Eq. (93*a*).

The input admittance $\frac{1}{R_g} + j\omega C_g$ of a vacuum-tube amplifier is seen to be proportional to the frequency, provided the magnitude and phase shift of amplification are the same for the different frequencies, and as a consequence becomes of greater importance the higher the frequencies which the amplifier is designed to handle. The input admittance is also almost (but not exactly) proportional to the tube amplification A, and since tubes with high amplification factors have a high tube amplification, the input admittance in amplifiers will tend to be proportional to the amplification factor of the tube. This sets the limit to the highest amplification factor that can be used to advantage with different types of amplification.

Effect of Input Admittance on Audio-frequency Amplifiers.—In audio-frequency amplifiers the input admittance is of considerable importance at the higher audio frequencies because it sets an upper limit to the load impedance that can be obtained in the plate circuit of the preceding tube at these frequencies. The effect of the input admittance (especially the capacity component) is therefore to lower the response at high frequencies, and because this is particularly the case with tubes having a high amplification factor, the highest amplification factor that it is practicable to use with resistance- and impedance-coupled amplifiers is limited. The same situation makes it preferable to use tubes with moderate amplification factors with transformer-coupled amplifiers and to rely upon the turn ratio to obtain the desired amplification.

The effect which the input capacity of amplifier tubes has in audio-frequency amplifiers can be accurately taken into account by computing the amplification backward, stage by stage, beginning with the output tube. The amplification of the last tube can be determined without difficulty as the load impedance in the plate circuit of the last stage does not include the input admittance of a succeeding vacuum tube. If the amplification of the last stage is known the input admittance of the last tube can be determined by Eqs. (91) and (92). This input admittance is a part of the load impedance in the plate circuit of the next to last tube, the amplification of which can now be computed. From this amplification the input admittance of the next-to-last tube can be obtained and combined with the remaining load impedance in the plate circuit of the second from the last tube to make possible the computation of the amplification of the second from the last stage, and so on, back to the input.

Effect of Input Admittance with Tuned Amplifiers.—The magnitude A and phase shift θ of the amplification in a tuned radio-frequency amplifier varies greatly with changes in frequency, and hence the input resistance and capacity will go through corresponding changes. The values

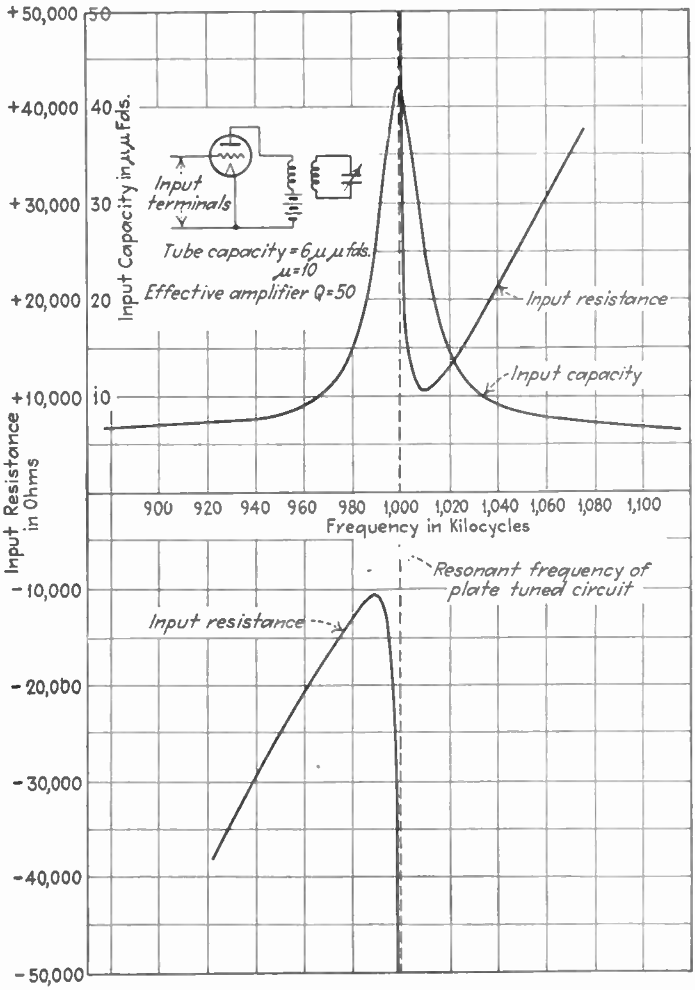


FIG. 98.—Input capacity and resistance of typical tuned radio-frequency amplifier as a function of frequency, showing the large variations that take place in these quantities near the resonant frequency of the tuned plate load impedance.

of input resistance and capacity in a typical radio-frequency amplifier operating at a broadcast frequency are shown in Fig. 98. The input capacity varies with frequency in very much the same way as does the amplification, while the input resistance goes through wide variations, being positive for frequencies higher than resonance (plate load impedance a capacitive reactance), negative for frequencies below resonance (plate load impedance an inductive reactance), and infinite at resonance.

When a tuned circuit is connected across the grid-cathode terminals of a tube having an input impedance such as shown in Fig. 98, the resonance characteristics of this tuned circuit are completely altered. Thus when the tuned circuit connected between grid and cathode has a higher resonant frequency than does the tuned circuit forming the plate load impedance, the negative input resistance of the tube neutralizes at least some of the resistance of the resonant circuit associated with the grid.

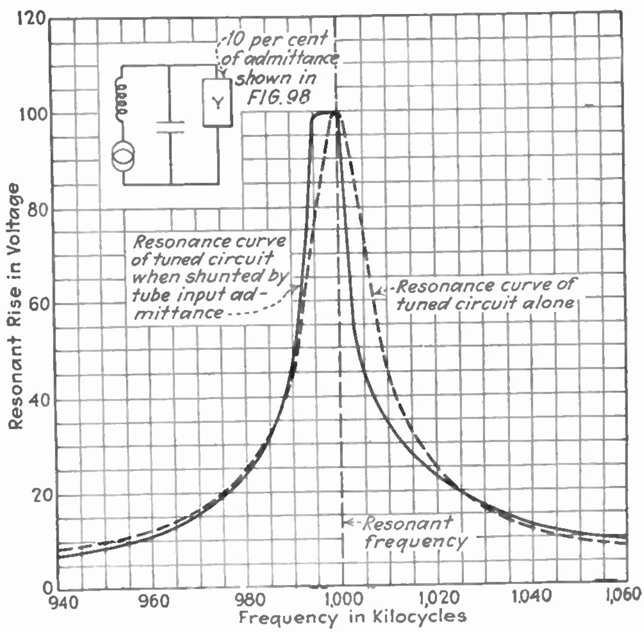


FIG. 99.—Unsymmetrical resonance curve obtained when a resonant circuit is shunted by an admittance similar to that of Fig. 98, but only one-tenth as great. This is one of the effects that can be produced by the input admittance of an amplifier tube operating with a plate load impedance consisting of a resonant circuit.

If the negative resistance is less than the parallel impedance which this grid tuned circuit develops at its resonant frequency, oscillations will be set up in the amplifier. Other peculiarities, such as double- or triple-resonant frequencies, unsymmetrical resonance curves, etc., can also occur as a result of input-admittance characteristics. These effects are so very detrimental as to make it necessary in practical radio-frequency amplifiers to employ special devices for minimizing the effect of the input admittance of the amplifier tube.

Even when the input admittance is relatively small, as is the case at the lower radio frequencies or with screen-grid tubes, or when the neutralization of input admittance is not quite complete, the energy transferred between grid and plate circuits of the amplifier through the grid-plate capacity, although small in amount, gives rise to peculiar effects, such as illustrated in Fig. 99, which shows the actual resonance curve obtained when a tuned circuit is shunted by an admittance that is 10 per cent of that shown in Fig. 98. The extra high response for frequencies slightly below resonance is a result of the negative input resistance for such frequencies, while the low response at frequencies slightly above resonance is caused by the low positive input resistance of the tube for this situation.

44. Neutralization of Input Admittance of Vacuum-tube Amplifiers.—The effects which are produced by the transfer of energy between

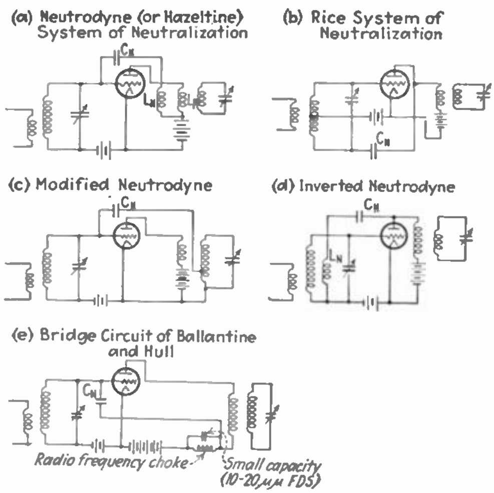


FIG. 100.—Typical circuit arrangements for neutralizing the effect of energy transfer between grid and plate circuits through the grid-plate tube capacity by an equal and opposite energy transfer through a neutralizing condenser C_N.

the grid and plate circuits of a vacuum-tube amplifier through the grid-plate tube capacity can be neutralized by an electrical network that transfers an equal amount of energy in the opposite direction. There are a number of ways in which this operation can be carried out, the most common of which are shown in Fig. 100. All of these arrangements employ a neutralizing condenser C_N connecting the input (*i.e.*, grid) circuit and the output (*i.e.*, plate) circuit in such a way that the current passing through the neutralizing condenser is of the proper amplitude and phase to neutralize exactly the effect of the current flowing between plate and grid circuits of the amplifier *via* the grid-plate tube capacity. Thus consider the circuit of Fig. 100*a*, which consists of an ordinary transformer-coupled radio-frequency amplifier to which there has been added a neutralizing inductance L_N connected in such a way that the voltage at the end of this coil connected to the neutralizing condenser C_N is in phase opposition to the voltage at the plate end of the primary inductance L_p. The voltage across L_N is then applied to the neutralizing condenser and causes the grid to receive a current that with proper size of C_N is equal in magnitude and opposite in phase to the current flowing through the grid-plate tube capacity, and so completely neutralizes the energy transfer through the tube capacity. The input capacity of such an amplifier with perfect neutralization is $C_{gp} + C_N$, and the input resistance is infinite. If the coupling between L_N and L_p is very close, the neutralization is substantially independent of frequency.

A neutralizing circuit of a slightly different type is shown at Fig. 100*b*. In this arrangement the neutralizing condenser C_N is given such a capacity that the current through it as a result of the voltage developed in the plate circuit is equal in magnitude to the current passing through the grid-plate tube capacity, but since these two currents produce effects in the tuned circuit that are in phase opposition, they neutralize each other. This neutralization is theoretically independent of the frequency at which the tuned circuit is resonant and under practical conditions can be made approximately so. Several additional types of neutralizing circuits are shown in Fig. 100, and still other arrangements have been devised, but since all of these make use of the same general principles that have been applied to the two specific cases discussed above, they need not be given special consideration.

A different and sometimes very profitable viewpoint is to consider the neutralizing arrangement as a bridge in which the output and input circuits are connected across the opposite diagonals. When the bridge is balanced the input circuit receives no energy from the output circuit because the two are in electrically neutral locations with respect to each other.

Effect of Imperfect Neutralization.—Perfect neutralization cannot be maintained in practice over a wide band of frequencies because leakage inductances and stray capacities prevent the neutralizing current from being exactly proportional to, and exactly out of phase with, the current through the grid-plate tube capacity at all frequencies. Imperfect neutralization gives rise to a certain amount of regeneration, but it is possible to maintain the balance sufficiently well under actual operating conditions to give an entirely satisfactory performance of the neutralized radio-frequency amplifier up to frequencies exceeding several million cycles per second. Figure 101 shows the effect which various degrees of unbalance in the neutralization system produce upon the input capacity and resistance of a typical tuned radio-frequency amplifier. With perfect balance the input resistance is infinite, and the input capacity is constant. With insufficient neutralization the input resistance and capacity curves are similar to those for no neutralization but vary through a smaller range, while overneutralization causes the input capacity to be negative near resonance and makes the input resistance negative at high frequencies and positive at low frequencies. The practical effects of overneutralization are thus similar to those of underneutralization except that the distortion in the input resonant circuit is of opposite symmetry from that with insufficient neutralization. In either case

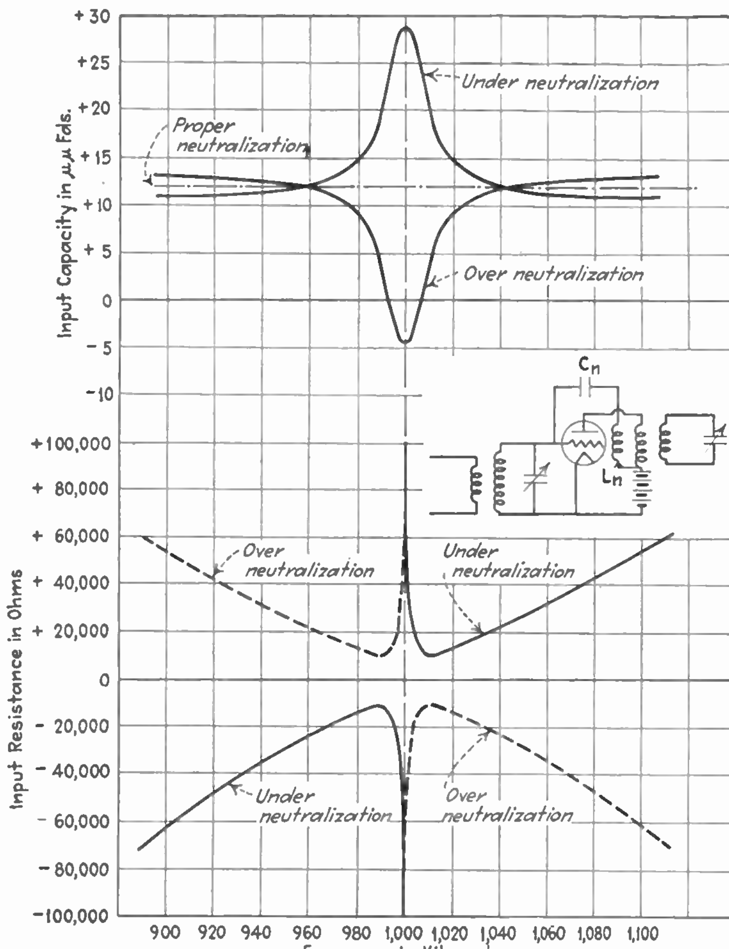


FIG. 101.—Curves showing effect of improper neutralizing capacity C_n on input resistance and capacity of tuned radio-frequency amplifier.

the effect is to distort the resonance curve of the input circuit in a manner analogous to that shown in Fig. 99.

Loss Methods of Minimizing the Effects of Regeneration through Grid-plate Capacity.—Methods other than those of neutralization are sometimes employed to control the detrimental effects of energy transfer through the grid-plate tube capacity. Since it is the negative resistance of the input admittance that causes the largest part of the trouble in radio-frequency amplifiers, it is possible to compensate for this effect by employing positive resistance to neutralize the negative input resistance. The only arrangement of this type that is at all satisfactory in practice is that illustrated in Fig. 102, in which the compensation for the negative input resistance is supplied by the losses introduced in the input circuit by a resistance of about 1000 ohms in series with the grid of the tube. Since the input admittance is greater, and hence the current through the resistance larger, as the frequency is increased, the losses in the grid resistor increase with frequency and automatically tend to compensate for the increased energy transfer at higher frequencies. Actually, since the current in the grid resistor is proportional to the frequency, the losses which it introduces are proportional to the square of the frequency, so that if an amplifier is correctly compensated at one frequency when using a particular value of grid resistance, it is undercompensated at lower frequencies, and overcompensated at higher ones. The grid resistor has a greater effect when the amplification is increased, which comes about because a greater amplification means a greater input admittance and more grid current. This tends to equalize the amplification for different adjustments and to prevent oscillations that might otherwise occur.

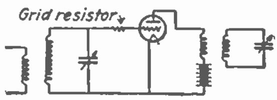


FIG. 102.—Tuned radio-frequency amplifier in which a series grid resistor is used to compensate for the worst effects of the tube input admittance.

45. Noise Level of Amplifiers.—All amplifiers give some output voltage even when the input voltage is reduced to zero. In audio-frequency amplifiers the output currents obtained in this way in the absence of a signal voltage produce what is commonly referred to as "noise" when flowing through a telephone receiver or loud-speaker, and it is also common practice to apply the term "noise" to the corresponding radio-frequency currents obtained in the output of a radio-frequency amplifier, although these lie above the range of audible frequencies.

A portion of the noise in amplifiers is the result of such obvious and easily remedied causes as induction from power circuits, exhausted batteries, poor contacts, faulty resistors, mechanical vibrations, and the like, but even when all of these effects are eliminated there is always a residual noise that arises from effects within the tubes themselves and cannot be eliminated. This residual noise is caused by the "shot effect," by thermal agitation of the electrons in conductors associated with the amplifier, and by miscellaneous causes, such as ionization and secondary electron emission.[1]

[1] For a somewhat more detailed discussion of noise in amplifiers see F. B. Llewellyn, A Study of Noise in Vacuum Tubes and Attached Circuits, *Proc. I.R.E.*, vol. 18, p. 243, February, 1930.

Shot Effect.—The shot effect results from the fact that the stream of electrons flowing from cathode to plate is made up of a series of particles rather than a continuous fluid. As a result, the electron flow to the plate is somewhat irregular, resembling hailstones striking a metal surface, and this gives rise to slight irregularities in the plate current of the vacuum tube and hence to noises in the amplifier. The presence of a space charge in the vacuum tube tends to smooth out the irregularities in the arrival of electrons at the plate, and this smoothing effect is so great as practically to eliminate the shot effect when complete temperature saturation exists. *It is therefore very important that the electron emission from the cathode be sufficient to produce an adequate space charge if the noise level is to be kept low.* The irregularities produced by the shot effect represent a distribution of alternating-current energy that is substantially uniform throughout the frequency spectrum, so that the magnitude of the noise voltage produced on frequencies lying between 1000 and 2000 cycles is just the same as the magnitude of the noise voltages between 1,000,000 and 1,001,000 cycles.

Thermal Agitation.—The most important source of noise in properly operated amplifiers arises from thermal agitation of the electrons of the conductors in the input circuit of the amplifier. It is well known that the conductivity of metals is a result of the presence of free electrons, and that these electrons are continuously moving about in the conductor at a velocity which depends upon the temperature. At any one instant there will ordinarily be more electrons moving in one direction than in the other, causing a voltage to develop across the terminals of the conductor. This voltage will vary from instant to instant in an irregular manner in accordance with the predominant motion of the electrons in the conductor. The square of the voltage developed in this way across the terminals of an impedance is directly proportional to the resistance component of the impedance and also directly proportional to the absolute temperature of the impedance, while the energy is uniformly distributed over the entire frequency spectrum from zero frequency up to frequencies much higher than those used in radio communication. The magnitude of the voltage produced by thermal agitation can be calculated by the formula

$$\left.\begin{array}{l}\text{Square of effective value of voltage com-}\\ \text{ponents lying between frequencies } f_1 \text{ and } f_2\end{array}\right\} = E^2 = 4kT\int_{f_1}^{f_2} R\,df \qquad (94)$$

where

k = Boltzmann's constant = 1.374×10^{-23} joule per degree
T = absolute temperature
R = resistance component of impedance producing voltages of thermal agitation (a function of frequency)
f = frequency.

In the special case where the resistance component of the impedance is constant over the range of frequencies from f_1 to f_2, Eq. (94) reduces to the much simpler form

$$E^2 = 4kTR(f_2 - f_1) \tag{95}$$

The noise voltage produced in an amplifier by thermal agitation can be calculated on the assumption that all of this noise arises from the impedance across the input terminals of the amplifier. On this basis the square of the effective value of the noise voltage developed in the amplifier output is

$$\left.\begin{matrix}\text{Square of effective noise voltage}\\ \text{in amplifier output}\end{matrix}\right\} = E^2 = 4kT\int_0^\infty R|A|^2 df \tag{96}$$

where $|A|$ is the magnitude of voltage amplification between the point where the thermal voltage is produced and where E^2 is measured. Equation (96) shows that the square of the noise ~~power~~ voltage in the output is proportional to the width of the frequency band being amplified, to the amplification, and to the resistance component of the impedance connected between the input terminals of the amplifier.

Miscellaneous Aspects of Noise in Amplifiers.—In addition to the noise resulting from shot effect and thermal agitation, there is ordinarily a small residual noise caused by irregularities resulting from the emission of secondary electrons at the plate of the tube, and by disturbances in the space charge around the cathode resulting from ionization. These effects are not entirely understood, but are fortunately relatively small compared with the thermal-agitation noise and so are not of major importance.

The noise produced in an amplifier by the combined action of shot effect, thermal agitation, secondary emission, and so on, determines the minimum voltage that can be amplified without being lost in a background of noise. The square of this minimum voltage is inversely proportional to the width of the frequency band to which the amplifier responds and to the resistance component of the input impedance. An idea of the magnitude of this minimum voltage can be gained by noting that substitution in Eq. (95) shows that the effective voltage produced by thermal agitation (which usually accounts for the major part of noise in amplifiers) across a ½ megohm resistance at a temperature of 300°K. for a frequency band 5000 cycles wide is 6.4 μv.

When a tube is jarred the electrodes tend to vibrate mechanically at their own natural periods, which gives rise to irregularities in the output of the tube referred to as microphonic noises. The vibrations that produce these microphonic effects may be received by the tube either through the base or from sound waves striking the glass bulb. Vibrations are especially troublesome in radio equipment used on airplanes, and in radio receivers having a loud-speaker in the same cabinet as the receiving tubes. Microphonic effects can be reduced by employing a spring support for the tube base, which prevents the transmission of mechanical vibrations to the tube through the base, and by covering the tube with a sound-absorbing cap to prevent acoustic waves from reaching the glass bulb. Special non-microphonic tubes having very rigid elements (and especially rigid filaments) have been developed for airplane and other uses where intense vibrations are present.[1] The most important microphonic effects are usually those introduced by the first tube of the audio-frequency amplifying system, since any noises introduced at this point are amplified by all of the stages of audio amplification. Microphonic effects can also be introduced in radio-frequency amplifiers, since vibrations will produce variations that modulate the radio frequency being amplified, and ultimately appear as noise after rectification.

46. Class B (Linear) Power Amplifiers.—The Class B amplifier is a power amplifier operated in such a way as to develop a voltage output proportional to the input voltage applied to the grid (*i.e.*, power output proportional to the square of the grid-exciting voltage) at a higher plate efficiency than that obtained with the distortionless power amplifier of Sec. 39. This is accomplished by operating with a negative grid bias that approximates cut-off (*i.e.*, the bias that makes the plate current approach zero with full plate-supply voltage and no grid excitation) and using a load impedance consisting of a resonant circuit tuned to the frequency being amplified. Under these conditions the application of an alternating exciting voltage to the grid causes the plate current to flow in impulses that are essentially half sine waves which flow during the positive half-cycle of the exciting voltage, as shown in Fig. 103. The fundamental frequency component of these half sine waves is picked out by a suitable tuned load impedance, and, if the plate-current impulses are true half sine waves having an amplitude proportional to the grid-exciting voltage, the voltage developed by the tuned load impedance will be exactly proportional to the grid-exciting voltage, and the amplifier will be linear. The Class B amplifier is not distortionless because the plate-current impulses do not reproduce the wave form of the grid-exciting voltage, but it can be used in the amplification of a modulated sine wave because the harmonic (*i.e.*, distortion) frequencies contained in the plate-current impulses are prevented from appearing in the output as a result of the resonance action of the tuned load impedance. The plate efficiency of the Class B amplifier is higher than the plate efficiency of the distortionless amplifier because no d-c plate current flows when the grid exciting voltage is removed, and for this reason the linear, *i.e.*, Class B, amplifier is always used in the amplification of modulated waves when plate efficiency is of importance.

[1] See Alan C. Rockwood and Warren R. Ferris, Microphonic Improvements in Vacuum Tubes, *Proc. I.R.E.* vol. 17, p. 1621, September, 1929.

Voltage and Current Relations in Linear Amplifiers.—The behavior of a linear amplifier can be determined by a study of the voltage and

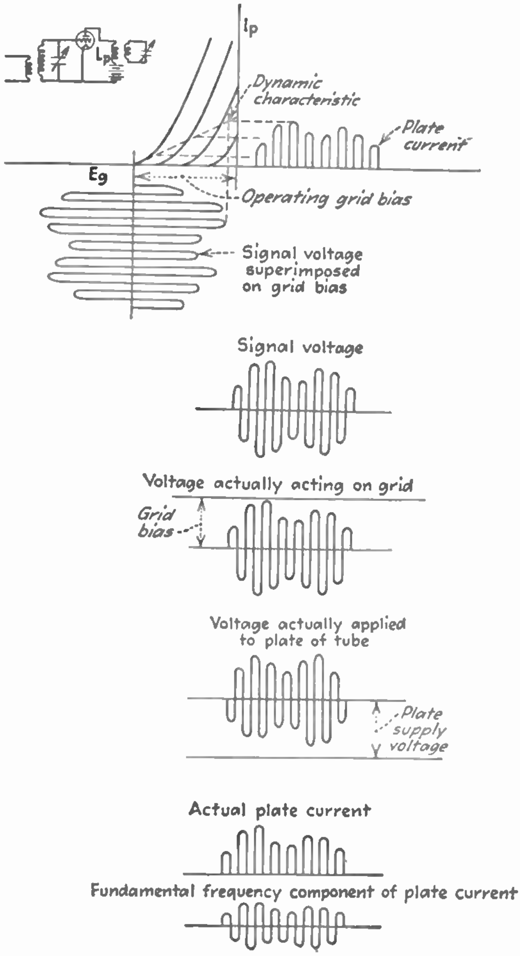


FIG. 103.—Figure illustrating operation of linear amplifier, giving circuit diagram, static and dynamic tube characteristics, and oscillograms of instantaneous grid and plate voltages, and plate current.

current relations that exist during operation. In making this analysis the following notation will be used:

E_s = vector value of alternating signal voltage applied to grid
E_B = voltage of plate-supply source
E_c = grid bias (E_c negative with negative bias)
μ = amplification factor of tube
I_m = crest value of plate-current impulses
I_{ac} = vector value of signal-frequency component contained in plate-current impulses

Z_L = impedance which tuned load circuit inserts in series with plate

r = direct-current resistance of plate circuit (*i.e.*, ratio of direct-current plate voltage to total direct-current plate current) when the grid bias is zero

$\alpha = I_{ac}/I_m$ = ratio of signal-frequency component of plate-current impulses to maximum value of impulses.

The voltage applied to the plate of the tube is the plate-supply voltage less the voltage drop developed across the tuned load circuit, and so is $E_B - I_{ac}Z_L$. This voltage across the load circuit contains only the signal-frequency component of the plate-current impulses because the plate load impedance is a resonant circuit and so offers negligible impedance to the harmonic frequencies of the signal frequency. The voltage applied to the grid of the tube is $E_c + E_s$, and the plate current actually flowing is the current that is produced by the combined action of these plate and grid voltages. These waves are shown in Fig. 103 for a typical case and, when worked out from the characteristic curves of the tube, will be found to lead always to plate-current impulses that approximate half sine waves.

The exact shape of the plate-current impulses can be determined by replacing the voltage acting on the grid by a voltage μ times as great acting in the plate circuit, which makes the effective plate voltage have the value $E_B - I_{ac}Z_L + \mu E_c + \mu E_s$. When the grid bias E_c is adjusted to cut-off, then $E_B + \mu E_c = 0$, and the effective plate voltage is $\mu E_s - I_{ac}Z_L$. To the extent that the amplification factor near cut-off is the same as at zero grid bias, *the plate current that actually flows in the linear amplifier is the same as the plate current that is produced when an alternating voltage $\mu E_s - I_{ac}Z_L$ is applied to the plate of the tube with zero grid bias and zero direct-current plate voltage.* The plate current flows only during the half-cycle when this equivalent plate voltage is positive, and is in the form of impulses that resemble half sine waves and which would be exactly so if the $I_p - E_p$ characteristic curves of the tube near cut-off grid bias were straight lines. There is always some curvature present, however, and this has the effect of making the plate-current impulses slightly peaked.

The amplitude I_m of the plate-current impulses produced by the effective plate voltage $(\mu E_s - I_{ac}Z_L)$ that can be considered as acting with zero grid bias is

$$I_m = \frac{\mu E_s - I_{ac}Z_L}{r} \tag{97a}$$

where r is the ratio of plate voltage to plate current for the point on the tube characteristic corresponding to the current I_m, when taken for zero grid bias. It is to be noted that r is the direct-current (not dynamic) resistance of the plate circuit at the current I_m and represents an effective plate resistance which approximates the average value of the dynamic plate resistance as the plate current varies from zero to I_m.[1] It is possible to replace I_{ac} by its equivalent αI_m, with the result

$$I_m = \frac{\mu E_s - \alpha I_m Z_L}{r}$$

or

$$I_m = \frac{\mu E_s}{r + \alpha Z_L} \tag{97b}$$

The amplitude of the plate-current impulse accordingly has a value that corresponds to an equivalent plate voltage μE_s acting in a circuit of impedance $(r + \alpha Z_L)$, and the load impedance effective in the linear amplifier is α times the actual load impedance physically present.

Formulas for Designing and Calculating Performance of Linear Amplifiers.—The proper load impedance, the power output, the plate efficiency, and the power dissipation at the plate can be determined for the linear amplifier with an accuracy sufficient for all ordinary purposes by an analysis based on Eq. (97*b*). In carrying out this analysis the load circuit will be considered as having been adjusted to resonance with the voltage being amplified, so that Z_L becomes a pure resistance R_L. The plate-current impulses will also be considered as approximating half sine waves, in which case $\alpha = 0.5$. To the degree of accuracy that is afforded by these approximations the maximum power output that can be delivered by the linear amplifier without allowing the grid to become positive is[2]

$$\text{Power delivered to load} = \frac{E_B^2}{2r} \frac{R_L/r}{\left(2 + \dfrac{R_L}{r}\right)^2} \tag{98}$$

Differentiating Eq. (98) with respect to load resistance R_L and equating the result to zero shows that *the maximum power output is obtained with a load resistance R_L equal to twice the direct-current plate resistance r of the tube*, which gives

[1] The ratio r/R_p of direct-current plate resistance with zero grid bias to dynamic plate resistance when taken for the same current I_m, is equal to the exponent in Eq. (59) and so is very close to $3/2$ in practical cases.

[2] This equation can be derived from Eq. (97) as follows: The maximum signal voltage E_s that will not make the grid go positive is $E_s = E_c$, and since E_c is the cut-off grid bias, then $\mu E_s = E_B$. Hence

$$I_m = \frac{E_B}{r + \alpha R_L}$$

The power delivered to the load resistance R_L by the signal-frequency component I_{ac} of I_m is $(I_{ac})^2 R_L/2$, but $I_{ac} = I_m/2$, so that

$$\text{Maximum power} = \frac{(I_{ac})^2 R_L}{2} = \frac{E_B^2 R_L}{8(r + \alpha R_L)^2}$$

This readily reduces to Eq. (98).

$$\text{Maximum possible power output} = \frac{E_B{}^2}{16r} \tag{99}$$

The power input to the amplifier must be supplied by the plate-voltage source and is equal to the product of the plate-supply voltage and the average value of the plate current. Since the average value of a half sine wave is $1/\pi$ of the maximum, then the power when $E_s = E_c$ is

$$\text{Power input to amplifier} = \frac{E_B I_m}{\pi} = \frac{E_B{}^2}{\frac{\pi}{2}r\left(2 + \frac{R_L}{r}\right)} \tag{100}$$

The efficiency with which the vacuum tube transforms the direct-current power supplied by the plate-voltage source into amplified signal is the ratio of Eq. (98) to Eq. (100) and therefore is

$$\text{Plate efficiency} = \frac{\pi}{4}\frac{R_L/r}{\left(2 + \frac{R_L}{r}\right)} \tag{101}$$

This represents the plate efficiency with the largest possible applied signal (*i.e.*, $E_s = E_c$). For smaller applied signals the plate efficiency will be proportional to the ratio of actual signal to maximum possible signal. The efficiency is seen to increase as the ratio of load resistance to effective plate resistance is increased, and reaches a maximum of $\pi/4$, or 78.3 per cent, when the load resistance is very large compared with the plate resistance. The power that must be dissipated by the plate of the tube is the difference between the power supplied by the plate-voltage source and that delivered to the load. The plate loss with maximum possible signal voltage applied to the grid is therefore the difference between Eqs. (100) and (98) and is

$$\text{Plate loss} = \frac{E_B{}^2}{r}\frac{4}{\pi}\frac{\left(1 + 0.1073\frac{R_L}{r}\right)}{\left(2 + \frac{R_L}{r}\right)^2} \tag{102}$$

With a signal voltage having a crest amplitude less than the grid bias, the plate loss will be directly proportional to the signal amplitude.

When the radio-frequency voltage applied to the grid of a linear amplifier is a modulated wave, Eqs. (98) to (102) apply only at the crest of the modulation cycle. Since the power output of a linear amplifier is proportional to the square of the alternating-current input voltage, while the d-c plate current (and hence the power input to the amplifier) is proportional to the amplitude of the plate-current impulses, Eqs. (98) to (101) take the following forms when a wave modulated to a degree m is being amplified:

$$\left.\begin{array}{r}\text{Carrier power that can}\\ \text{be delivered to load}\end{array}\right\} = \left(\frac{1}{1+m}\right)^2 \frac{E_B^2}{2r} \frac{R_L/r}{\left(2+\frac{R_L}{r}\right)^2} \qquad (98a)$$

$$\text{Maximum possible carrier output} = \left(\frac{1}{1+m}\right)^2 \frac{E_B^2}{16r} \qquad (99a)$$

$$\text{Average power input to the amplifier} = \left(\frac{1}{1+m}\right) \frac{E_B^2}{\frac{\pi}{2} r \left(2+\frac{R_L}{r}\right)} \qquad (100a)$$

$$\left.\begin{array}{r}\text{Plate efficiency when}\\ \text{carrier is unmodulated}\end{array}\right\} = \left(\frac{1}{1+m}\right) \frac{\pi}{4} \frac{R_L/r}{\left(2+\frac{R_L}{r}\right)} \qquad (101a)$$

Examination of these equations shows that when provision is made for amplifying a completely modulated wave the carrier power is only one-fourth of the output obtainable at the same direct-current plate voltage when the maximum possible alternating-current input is applied, and the plate efficiency obtained in amplifying the carrier wave is only one-half as great.

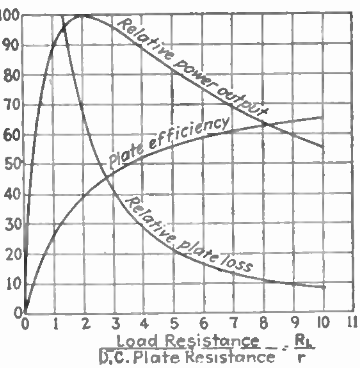


FIG. 104.—Variation of power output, plate efficiency, and plate loss, as a function of R_L/r, while maintaining the plate-supply voltage and direct-current plate resistance r constant.

The way in which the power output, the plate efficiency, and the plate loss vary as the ratio R_L/r changes is shown in Fig. 104. It is seen from these curves that the load resistance should be at least twice the effective plate resistance of the tube, and that increasing the load resistance raises the efficiency and reduces the plate loss while lowering the maximum power output that can be obtained. The exact value of load resistance is not highly critical, however, and satisfactory results can be obtained over a relatively large range.

Distortion in Linear Amplifiers.—The linear amplifier is intended to develop an output voltage proportional to the signal applied to the grid, but will do so only when the dynamic characteristic of the amplifier is a straight line. The dynamic characteristic of a linear amplifier represents the path along which the plate current actually varies when a signal voltage is applied to the grid and there is a load impedance in the plate circuit. When the tuned load circuit is resonant at the frequency being amplified, the dynamic characteristic has a slope that corresponds to an effective plate resistance approximating $r + R_L/2$, where R_L is the resistance offered by the tuned load circuit at parallel resonance. The dynamic characteristic can be determined exactly only by a trial-and-error process, since the amplitude of the alternating voltage developed across the load impedance depends upon the dynamic characteristic, and the dynamic characteristic in turn depends upon the voltage developed across the load impedance. A typical dynamic characteristic is to be found in Fig. 103. The ideal straight-line characteristic is more nearly approximated the larger the ratio R_L/r and the larger the signal voltage. The use of a load resistance R_L that is large compared with the direct-current tube plate resistance r has the effect of straightening out the characteristic as described in Sec. 39, while the larger the signal the greater will be the proportion of the operating range that approximates a straight line, because, as is apparent in Fig. 103, the dynamic characteristic has most of its curvature in the immediate vicinity of cut-off. It is also always desirable to use a plate-supply voltage that is as high as possible, because this calls for a high grid bias and so permits the signal to be large, all of which helps make the amplifier more nearly linear. The degree of linearity actually obtained in practical Class B amplifiers is sufficiently high to meet all ordinary requirements provided the signal voltage is not too small and provided the plate load impedance is reasonably high. If the amplifier is truly linear the d-c plate current, and hence the power input to the plate, does not change as the carrier wave is modulated.

Tubes Suitable for Use in Linear Amplification.—Tubes intended for use in a linear amplifier should have a low amplification factor and should be operated with a load resistance R_L that is high in comparison with the effective plate resistance r of the tube. The use of a high ratio of load resistance to effective plate resistance increases the linearity of the amplifier and also the plate efficiency, while the low amplification factor makes it possible to obtain a large power output with a given plate-supply voltage and fixed ratio of R_L/r. An examination of Eqs. (98) and (101) shows that the plate voltage used with the amplifier should be as large as possible, for with a given ratio R_L/r and a given value of effective plate resistance r, the power output will vary as the square of the plate-supply voltage, while when the tube losses are the limiting factor, increasing the plate voltage permits the use of a higher ratio of load resistance R_L to effective plate resistance r, which increases the plate efficiency and hence gives more power output with a fixed plate loss.

Design and Adjustment of Linear Amplifiers.—In designing linear amplifiers and proportioning the associated circuits there are two cases to consider. When the power output is not limited by tube losses the load resistance (*i.e.*, the resistance which the tuned circuit couples into the plate circuit of the tube at resonance) should be twice the effective plate resistance of the tube, and the grid bias should approximate the plate-supply voltage divided by the tube amplification factor. In the case where the power output is limited by the plate loss of the tube, the highest possible plate voltage should be selected, together with a grid bias approximating cut-off. The ratio of load resistance R_L to effective plate resistance r that should be used is the lowest value that will not make the tube loss exceed the allowable value, and can be evaluated with the aid of Eq. (102). The tuned load circuit is then coupled into the plate circuit in such a way as to give the value of load resistance R_L corresponding to this ratio.

The procedure for placing a linear amplifier in operation is as follows: First, the grid bias is adjusted to cut-off for the plate-supply potential that is to be employed. This bias is the plate-supply potential divided by the amplification factor of the tube, and can be estimated either from known tube characteristics or by observing the grid bias at which the plate current approaches zero. The next step is to obtain the proper alternating-current grid excitation. This adjustment, which can be made with any convenient load impedance in the plate circuit, consists in varying the grid excitation until a milliammeter in the grid circuit indicates that there is a small d-c grid current when the exciting voltage is modulated as completely as the apparatus will allow. Finally the plate load impedance is adjusted to give the highest plate efficiency consistent with the desired output. This adjustment should be carried out with the exciting voltage unmodulated and having the value obtained above. The load impedance should be at least twice the direct-current plate resistance of the tube since this is the condition that gives the maximum output that can be obtained at a given plate-supply potential. Lower values of load impedance give less output than the maximum and at a lower plate efficiency, and should never be used. A load impedance higher than twice the direct-current plate resistance increases the linearity of the amplifier and raises the plate efficiency but does not give as much output. The most convenient and safest way to adjust the load impedance is to start with the greatest load impedance available. This is then reduced until the plate losses become excessive, or until the maximum output is obtained. If the plate losses are excessive at the highest plate impedance that can be obtained, it is necessary to reduce the plate-supply potential and readjust the grid bias and alternating-current-exciting voltage accordingly. If on the other hand the tube is not operated up to its full capacity when the plate load impedance is the value giving maximum output, it is necessary to increase the plate voltage with corresponding increases in the grid bias and grid excitation.

The reason for making the adjustment of load impedance when the exciting voltage is unmodulated is that the plate losses under this condition are greater than when the same carrier wave is modulated. This is because the plate current drawn by a linear amplifier is proportional to the amplitude of the exciting voltage, and since the average amplitude of a modulated wave is the same as the carrier amplitude, the average or d-c plate current of the linear amplifier, and hence the direct-current power supplied to the anode, is not altered by modulation of the carrier. On the other hand the output of the linear amplifier is greater when the wave is being modulated because of the side-band power. The result is that the average plate efficiency is increased by modulation, causing the plate losses to be greatest when the exciting voltage is the unmodulated carrier.

The justification for adjusting the excitation so that grid current will flow at the modulation crest is that this introduces relatively little distortion if the current is not excessive, and increases the power-handling capacity of the tube as well as giving a slightly greater plate efficiency than would otherwise be obtained. For these reasons linear amplifiers handling modulated waves are nearly always operated so that the grid goes positive by a moderate amount during full modulation.

The fact that the d-c plate current of a truly linear amplifier depends upon the average amplitude of the exciting voltage can be used as a check on the linearity of the linear amplifier. If a milliammeter in series with the plate does not vary as the modulation changes, it is reasonably certain that the linear amplifier is not introducing non-linear distortion. If the plate milliammeter indicates that the plate current does change with variations in the degree of modulation it is highly probable that the amplifier is distorting and needs attention.

Linear amplifiers employed in the amplification of modulated waves differ from other types of radio-frequency amplifiers only in the grid bias which is employed. Linear amplifiers therefore have the same tendency to regenerate and oscillate as do all other types of radio-frequency amplifiers and, if of the three-electrode type, must be neutralized. An example of a neutralized push-pull linear amplifier is shown in Fig. 105.

In the analysis that has been given of linear amplification it has been assumed that the load impedance was a sharply resonant circuit that offered negligible impedance to the direct-current component of the plate current and the various harmonics of the signal voltage being amplified. This condition is approximated to a satisfactory degree by a resonant load circuit having a reasonably high value of Q and coupled into the plate circuit with a high coefficient of coupling. If the load consists of an actual resistance instead of a resonant circuit, the maximum plate efficiency will be exactly half of that obtainable with a tuned load circuit. This is because of the direct current and harmonic power dissipated in the resistance. If the plate load impedance offers appreciable resistance or reactance to the d-c plate current or the harmonics of the signal voltage being amplified, the maximum plate efficiency that can be obtained will lie somewhere between the value given by Eq. (101) and one-half this value.

Push-pull linear amplifiers are frequently employed, particularly where more output is desired than can be obtained from a single tube.

The circuit for a push-pull linear amplifier is shown in Fig. 105 and has the advantage of balancing out the effect produced on the resonant load circuit by the even harmonics in the impulses of plate current, as discussed in Sec. 41. The result is to cause the percentage of harmonics contained in the resonant load circuit to be much less than when the usual connection is employed. In a linear push-pull amplifier the tuned load circuit is adjusted so that the effective load resistance existing between the two plates has the proper value to match a plate resistance which is twice the value of r for a single tube. The performance of each tube is then calculated by taking R/r in Eqs. (98) to (102) as the ratio of impedance connected between the two plates divided by $2r$.

Class B Audio-frequency Amplifiers.—Linear push-pull amplifiers can be arranged to produce distortionless audio-frequency amplification with

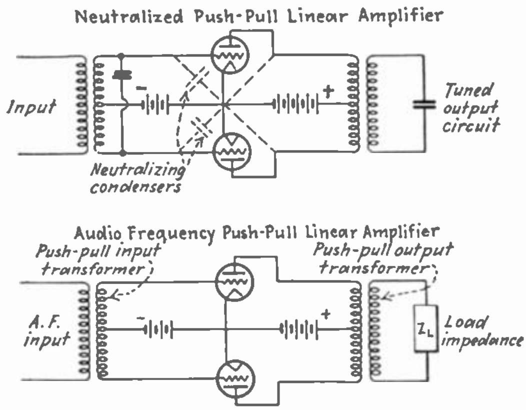


FIG. 105.—Push-pull linear amplifiers.

a plate efficiency considerably greater than obtainable with a Class A amplifier. The circuit for accomplishing this is shown in Fig. 105 and differs from the ordinary push-pull amplifier only in that the grid bias is adjusted to cut-off. The idea is that one of the tubes in push pull acts as a linear amplifier of the positive half-cycles of the input voltage, while the other tube in push pull acts as a linear amplifier for the negative half-cycles. The output transformer combines these separate outputs and therefore delivers a distortionless reproduction of the input signal provided the individual tubes are truly linear. There is little practical difficulty in obtaining in this way a substantially distortionless output at a much higher plate efficiency than can be realized with a Class A audio-frequency amplifier. If ordinary tubes are employed it is found, however, that at the rated plate voltage the output obtainable is small unless the grids are run positive, even though this small output is obtained at a high anode efficiency. This disadvantage of limited output can be overcome by making the input voltage so large that the grids of the linear push-pull amplifier swing positive by a very considerable amount. With such an arrangement large outputs can be obtained, and if the input circuits are so proportioned that the resulting grid current does not introduce excessive distortion in the input voltage, it is possible to obtain a practical

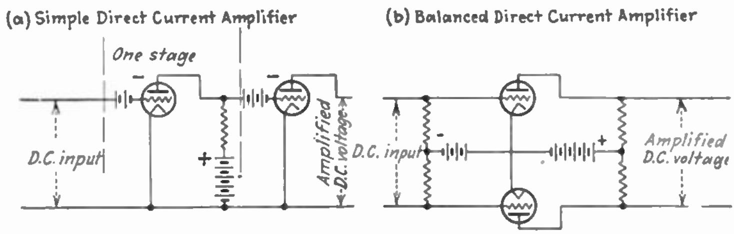


FIG. 106.—Simple and balanced types of direct-current amplifiers.

audio-frequency amplifier having a large power output and a plate efficiency in the order of 50 per cent.[1]

47. Direct-coupled Amplifiers.—The amplification of direct-current voltages (or voltages of extremely low frequency) requires a direct resistance coupling between the output of one stage and the input of the following amplifier tube without the use of grid-blocking condensers.

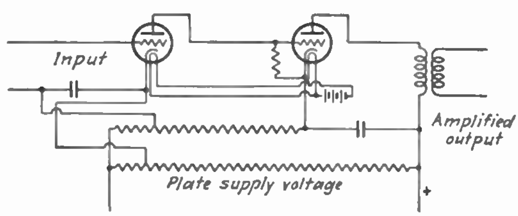


FIG. 107.—Direct-coupled amplifier of Loftin-White type, in which all of the several stages are connected in series across a single voltage source that supplies all plate- and grid-bias potentials.

One possible arrangement of this type is shown at Fig. 106*a*, in which the voltage drop across the coupling resistance is impressed directly on the grid of the succeeding tube, with a grid-bias potential sufficient to balance out the voltage drop across the coupling resistance resulting from the normal plate current that passes through this resistance and to maintain the grid slightly negative.

Where the cathodes of different tubes in the amplifier can be operated at different potentials to ground, as is the case with heater-type tubes (where the cathode does not need to be at the same potential as the heater), it is possible to place the plate circuits of the various amplifier stages in series across a single source of voltage, with the result that the bias batteries required in the arrangement shown in Fig. 106 can be

[1] See Loy E. Barton, High Audio Power from Relatively Small Tubes, *Proc. I.R.E.*, vol. 19, p. 1131, July, 1931.

eliminated.[1] Figure 107 gives such a circuit in simplified form and illustrates the manner of obtaining the various grid-bias and plate potentials. Amplifiers of this type ordinarily require some form of stabilizing arrangement (not shown in Fig. 107) to prevent small variations in operating conditions from shifting the operating point to undesirable locations. In the circuit of Fig. 107 the plate circuits of the two tubes are effectively in series and are also shunted by voltage-dividing resistances which help divide the potentials properly and provide the necessary grid-bias potentials.

Direct-current amplifiers of all sorts experience trouble from slow drifts in the plate currents of the tubes, since any slow change in plate current is amplified and masks the amplified direct-current voltage, particularly if this voltage is small. These defects can be largely eliminated by the balanced direct-current amplifier shown at Fig. 106*b*, in which the two similar tubes are arranged so that changes in plate or filament potentials, or in grid bias, will be balanced out as far the amplifier output is concerned, while a direct-current voltage applied to the input of the amplifier will unbalance the two tubes and produce an output voltage. An arrangement of this sort can be used to measure voltages as small as 50 μv without encountering troubles from drift in the direct-current plate current.[2]

Direct-current amplifiers are essentially resistance-coupled audio-frequency amplifiers in which the low frequency range has been extended down to zero frequency. Such amplifiers have the advantage of zero-phase shifts at low frequencies and are also very useful in the amplification of small voltages having a direct-current component that is of importance.

48. Vacuum-tube Harmonic Generators.—The vacuum-tube harmonic generator is essentially an amplifier operated under conditions which lead to high distortion and hence cause the amplified output to contain harmonics of the voltage applied to the grid. The distortion for the production of these harmonics may be obtained either by utilizing the non-linear relation that exists between the grid voltage and plate current, or by the non-linear relation between grid voltage and grid current. Harmonic generators of the former kind are said to be of the plate distortion, or C bias type, while the latter kind are called grid-distortion harmonic generators.

[1] Amplifiers of this type have been developed by Loftin and White. See their paper, Cascade Direct-coupled Tube Systems Operated from Alternating Current, *Proc. I.R.E.*, vol. 18, p. 669, April, 1930.

[2] For more information on amplifiers of this type see J. M. Eglin, A Direct-current Amplifier for Measuring Small Currents, *Jour. Optical Soc. Amer.*, vol. 18, p. 393, May, 1929. This article shows how small differences in tube characteristics can be compensated so completely as to give extremely great stability.

In the plate-distortion harmonic generator the tube is operated at a grid bias that exceeds the cut-off value (*i.e.*, $E_c > E_p/\mu$), and a fundamental frequency voltage having a crest amplitude approximating the grid bias is applied to the tube. Under these conditions the plate current flows in impulses somewhat similar to the way in which the plate current of the Class B amplifier does, but with the exception that with the usual adjustment these impulses last for less than a half cycle,

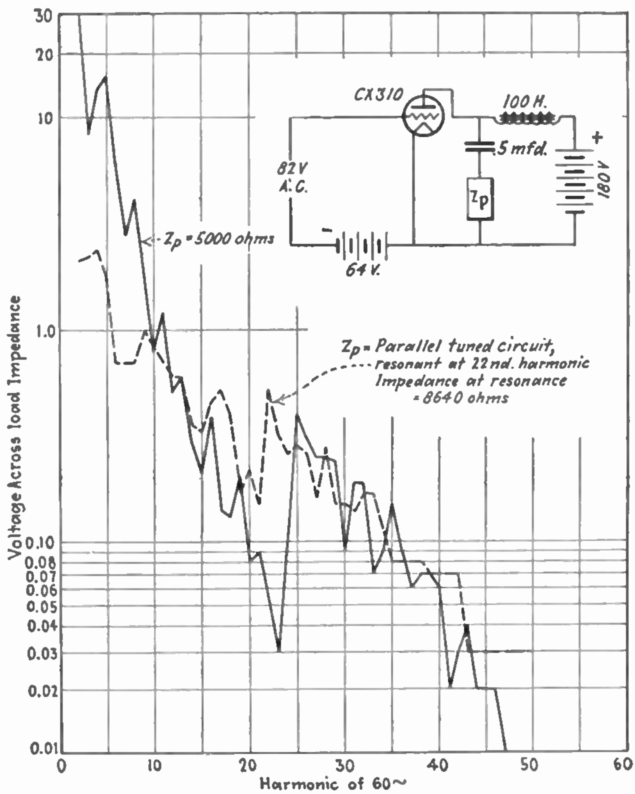


FIG. 108.—Harmonic spectrum produced by a plate-distortion type of harmonic generator with resistance load (solid lines) and load impedance formed by a parallel-resonant circuit tuned to the twenty-second harmonic (dotted lines).

causing the wave shape of plate current to depart very greatly from a sinusoidal form. The amount of power developed on the different harmonic frequencies of the input voltage depends to a large extent upon the circuit and tube characteristics, but in general is quite large for the second harmonic and decreases rapidly as the order of the harmonic increases. This is brought out by Fig. 108 where it is apparent that the amount of power on the higher harmonics is very small.

Since the plate-distortion harmonic generator operates under circuit conditions similar to those existing in the Class B amplifier, the same type of tubes is suited to both applications. In the generation of harmonics by plate distortion a high plate voltage should be employed, together with a negative grid bias somewhat greater than the cut-off value, and the signal voltage should be as large as possible without making the grid positive. The best grid bias will be different for different order harmonics and in general can be determined only by experiment. The load impedance employed in the plate circuit should be of a character that will select the desired harmonic or harmonics and reject the unwanted frequency components. Where a single harmonic is to be obtained the load impedance should be a resonant circuit tuned to the desired harmonic, while if a number of harmonics are to be simultaneously generated the load impedance should consist of either a resistance or an inductance.

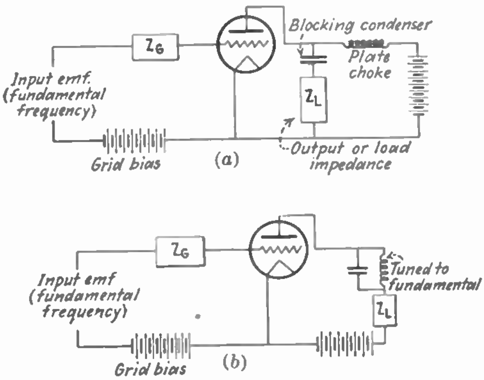


Fig. 109.—Basic circuit for grid-distortion type of harmonic generator, showing shunt-feed (at *a*) and series-feed (at *b*).

Grid-distortion Type of Harmonic Generator.—The grid-distortion type of harmonic generator makes use of the non-linear relation between grid voltage and grid current.[1] This is accomplished by operating the tube in such a way that the instantaneous grid potential is positive for a small part of each cycle of the signal voltage, and then forcing the resulting impulses of grid current to flow through an impedance that is in series with the grid and which offers a high impedance to the desired harmonic or harmonics. The circuit arrangements of such a harmonic generator are shown in Fig. 109. The grid current, by flowing in impulses lasting only a short portion of the cycle, is rich in harmonics of the applied voltage and will produce an appreciable voltage drop in the series grid impedance at the harmonic frequencies for which the impedance is large. This harmonic voltage developed across the series grid impedance is applied to the grid of the tube, as can be seen from Fig. 109, and is amplified in

[1] For additional information on harmonic generators of the grid-distortion type see F. E. Terman, D. E. Chambers, and E. H. Fisher, Harmonic Generation by Means of Grid Circuit Distortion, *Trans. A.I.E.E.*, vol. 50, p. 811, June, 1931.

the plate circuit, where the harmonic output is obtained. The operation of the grid-distortion harmonic generator can thus be summarized by stating that the grid circuit is used to produce the harmonics, which are then amplified in the plate circuit.

The way in which the actual details of the grid-distortion harmonic generator work out can be seen by examining the oscillograms in Fig. 110, which are for an example in which the series grid impedance is supplied by a parallel-resonant circuit tuned to the twenty-second harmonic of

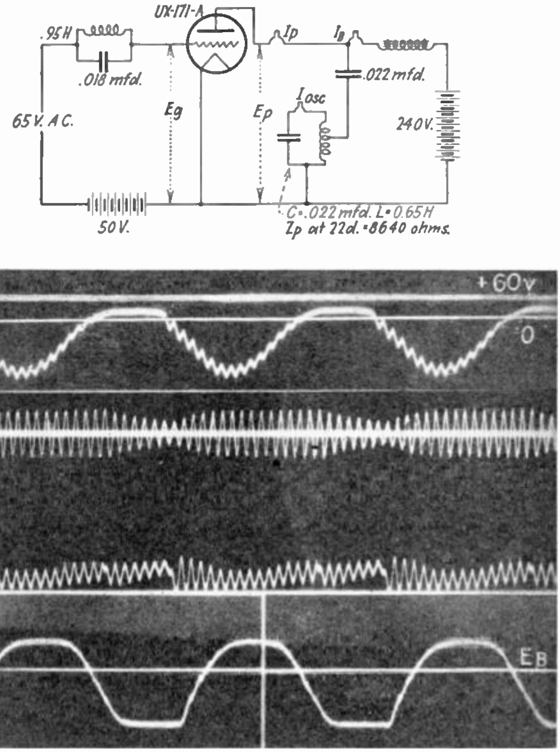


FIG. 110.—Circuit, test conditions, and series of oscillograms, showing performance of grid-distortion type of harmonic generator when using a distorting grid impedance tuned to the twenty-second harmonic. The twenty-second harmonic component of I_{osc} amounts to 25 milliamp.

the applied voltage. The tube is operated at a grid bias that approximates the normal value that would be used for distortionless power amplification, but the applied voltage is made so large that the grid is positive for approximately one-fourth of each cycle. The resulting grid current flows in a series of impulses that produce a voltage drop across the series grid impedance consisting primarily of the twenty-second harmonic of the applied voltage. The voltage actually applied to the tube is then the grid bias plus the sum of the applied voltage and the drop across the series grid impedance, which as seen in Fig. 110 contains a pronounced twenty-second harmonic ripple. This grid voltage is amplified in the plate circuit and the twenty-second harmonic component is picked out by a tuned load circuit. The actual amount of harmonic power is seen to be rather great in spite of the high order of the harmonic and is many thousands of times that obtainable on the same harmonic with a plate-distortion type of generator.

In order to obtain the best results with the grid-distortion type of harmonic generator it is necessary that the grid current be in impulses of short duration but large amplitude, that the series grid impedance to the desired harmonics be large, and that conditions be favorable for the amplification in the plate circuit of the harmonic voltage developed across the series grid impedance. The highly distorted grid impulses can be obtained by operating the tube with a fairly large grid bias combined with a signal voltage which causes the instantaneous grid potential to be positive for not more than about one-fourth of each cycle. The character of the series grid impedance to be used depends upon the object of the harmonic generator. Where a large amount of power is desired on a single frequency the impedance should be formed by a parallel-resonant circuit tuned to the desired harmonic frequency and having a large ratio of inductance to capacity. Where it is desired to obtain power on a number of higher harmonics simultaneously the series grid impedance should consist of a large inductance, since in this way the series grid impedance will be increasingly great to the high-order harmonics. Examples of these two types of series grid impedances are found in Figs. 110 and 111. The problem of obtaining good amplification of the harmonic voltage developed across the series grid impedance is somewhat difficult, because the large grid bias and the large amplitude of applied voltage that are required to produce highly distortioned impulses of grid current make the instantaneous plate current tend to reach cut-off at the negative crest of the signal voltage. It is shown in Sec. 39 that good amplification can be obtained with a large grid bias only when the tube has a low amplification factor and is operated with a very high load impedance in the plate circuit. This high load impedance to the fundamental frequency voltage can be obtained either by using a shunt-feed arrangement with a blocking condenser, as shown in Fig. 109*a*, or by a parallel-resonant circuit tuned to the frequency of the applied voltage, as shown in Fig. 109*b*. If the plate load impedance to the frequency of the applied voltage is not sufficiently high the plate current will become zero for a portion of the cycle, and during this interval the harmonic voltage produced across the series grid impedance will not be amplified, with the result that the harmonic power output is reduced.

Comparison of Grid- and Plate-distortion Types of Harmonic Generator.—In comparing the plate- and grid-distortion methods of producing harmonics, experiments show that the former method is usually to be preferred at low-order harmonics, such as the second, third and possibly the fourth, while the latter method is much the superior when a large amount of power is desired on a single high-order harmonic, or where appreciable power is to be obtained simultaneously on a large number of high-order harmonics. While the two methods of generation will give about the same power output on the lower order harmonics, the plate-distortion method is to be preferred for such cases because no grid current is drawn and hence the input power is negligible. As the order of the

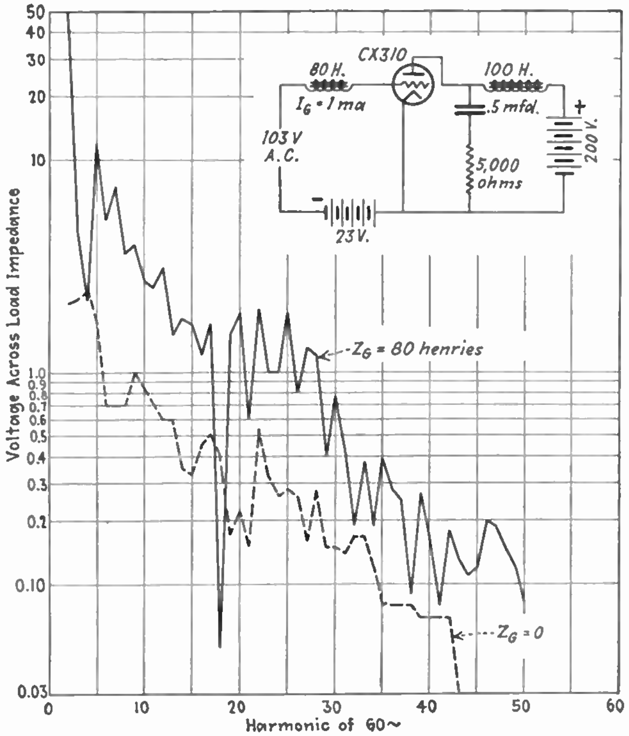


FIG. 111.—Harmonic spectrum produced by grid-distortion type of harmonic generator with an inductance for the series grid impedance. The dotted curve shows the performance of a plate-distortion generator replotted from Fig. 108 for purposes of comparison.

harmonic is increased the power developed by the plate-distortion generator drops off very rapidly, while the grid-distortion harmonic generator produces an output that decreases relatively slowly as the order of the harmonic is increased. Thus it is seen from Fig. 111 that the grid-distortion harmonic generator with a series grid impedance consisting of a series grid inductance gives much more harmonic power on the high-order harmonics than does the plate-distortion generator, and is capable of giving appreciable power on harmonics as high as the fiftieth. The superiority of the grid-distortion generator is even more striking when the series grid impedance consists of a parallel-resonant circuit, as is the case in Fig. 110, where the power output on the twenty-second harmonic is many thousands of times that which could be obtained from a generator of the plate-distortion type operating under the optimum conditions for producing this same harmonic. Another important feature of the grid-distortion harmonic generator is that the power output which it produces at the higher order harmonics is very steady, which is not the case with the plate-distortion type of generator at the higher frequencies.

Harmonic generators are of considerable importance in communication technique because they provide a means of frequency multiplication. Thus, in radio transmitters, harmonics are often employed to derive a high frequency for transmission from a lower frequency oscillator having great frequency stability. The operation of broadcasting stations on synchronized carrier frequencies can also be most easily carried out by transmitting an audio frequency such as 5000 cycles over wires to the various stations and deriving the transmitted carrier frequency from this audio frequency by the use of harmonic generators. Harmonic generators are also of importance in the measurement and comparison of frequencies, since harmonics have a fixed relation to the fundamental frequency and enable frequencies of different order of magnitude to be compared with accuracy.

See I.R.E. Dec. 1931

CHAPTER VII

VACUUM-TUBE OSCILLATORS

49. Vacuum-tube Oscillator Circuits.—A vacuum tube is able to act as an oscillator because of its ability to amplify. Since the power required by the input of an amplifier tube is much less than the amplified output it is possible to make the amplifier supply its input. When this is done, oscillations will be generated and the tube acts as a power converter that changes the direct-current power supplied to the plate circuit into alternating-current energy in the amplifier output. The efficiency of this conversion in properly adjusted oscillators is at least 50 per cent and can run as high as 80 to 90 per cent.

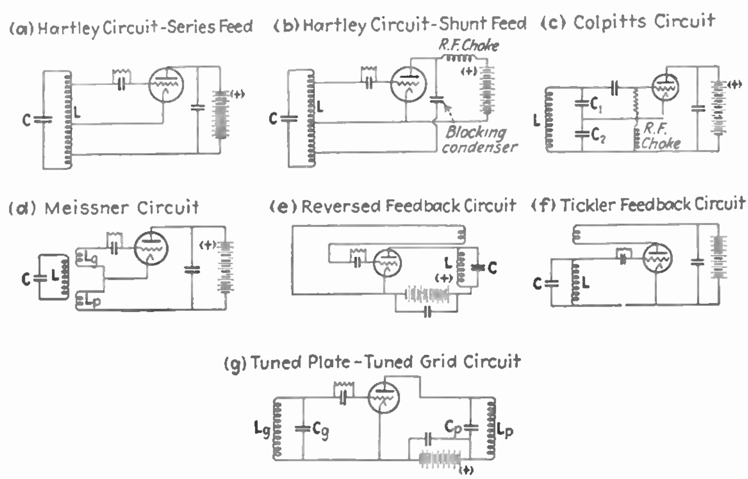


FIG. 112.—Typical oscillating circuits. In each case the frequency is determined by a resonant circuit, and the arrangement is such that the tube acts as an amplifier supplying its own input voltage.

Any amplifier circuit that is arranged to supply its own input voltage in the proper magnitude and phase will generate oscillations. Many circuits can be used for this purpose, of which a number are shown in Fig. 112. In general the voltage fed back from the output and applied to the grid of the tube must be approximately 180° out of phase with the voltage existing across the load impedance in the plate circuit of the amplifier, and must have a magnitude sufficient to produce the output power necessary to develop the input voltage. In the Hartley and

Colpitts circuits this is accomplished by applying to the grid a portion of the voltage developed in the resonant circuit. In the Meissner and the two feed-back circuits, mutual inductances are employed, while the tuned plate-tuned grid circuit transfers energy to the grid tuned circuit through the grid-plate tube capacity. All of the circuits shown in Fig. 112 except that at Fig. 112*b* use series feed in the plate circuit, but it is possible to employ shunt feed, and in fact the shunt feed is usually preferred in practical cases.

The frequency at which the oscillations occur is the frequency at which the voltage fed back from plate circuit to the grid is of exactly the proper phase and magnitude to enable the tube to supply its own input. In oscillators associated in some way with a resonant circuit, as are all those of Fig. 112, the frequency of oscillation approximates the resonant frequency of this circuit very closely.

50. Voltage and Current Relations in Oscillating Circuits.[1]—While the vacuum-tube oscillator is fundamentally an amplifier connected to supply its own input, the oscillator is usually operated under circuit conditions that are seldom found in amplifiers. In the first place, oscillators always operate with a high grid bias, ordinarily so great as to permit the plate current to flow only during the small part of the cycle when the alternating voltage applied to the grid is near its positive crest. By thus causing the plate current to flow in impulses it is possible to make the efficiency with which the direct-current plate power is converted into alternating-current energy reach much higher values than are obtainable with ordinary amplifiers. Secondly, the instantaneous grid potential of an oscillator is allowed to go positive for a small portion of each cycle. This causes the oscillator to draw grid current, the energy for which must be supplied from the oscillating current, but this disadvantage is much more than overcome by the increased efficiency in the conversion of direct-current into alternating-current energy that results. Thirdly, the grid bias is usually obtained by the use of a grid-leak and grid-condenser combination, as illustrated in Fig. 112. The direct-current component of the grid current flows through the grid-leak resistance and so produces a voltage drop that supplies the necessary grid bias, while the grid condenser serves as a by-pass for the alternating-current components of the grid current. The use of a grid leak and grid condenser eliminates the necessity of supplying the grid bias from an external source, and also tends to cause the oscillator automatically to select suitable operating conditions.

Voltage and Current Relations.—The mechanism by which the oscillator converts the direct-current energy supplied to the plate circuit from

[1] An excellent and exhaustive treatment of vacuum-tube oscillators is to be found in the series of articles by D. C. Prince, Vacuum Tubes as Power Oscillators, *Proc. I.R.E.*, vol. 11, pp. 275, 405, 527, June, August, October, 1923.

the plate-supply source into alternating-current energy with a high efficiency can be understood from a study of the voltage and current relations that exist in a properly adjusted oscillator circuit. Thus consider the series-feed Hartley circuit of Fig. 112. The voltage on the plate of the tube is the sum of the direct-current potential of the plate-power source and the alternating voltage developed between the plate and filament. When the oscillator is adjusted to give good efficiency the amplitude of oscillations is such that the crest value of this alternating plate voltage developed across the part of the resonant circuit between plate and cathode is almost, but not quite, equal to the direct-current plate-supply voltage, so that the voltage across the plate of the tube varies periodically as shown at Fig. 113*A*. Likewise the instantaneous potential on the grid is the sum of the grid-bias voltage and the voltage developed between the cathode and grid by the current in the tuned circuit. Since these grid and plate voltages are 180° out of phase, the variation of grid voltage over a cycle is as shown at Fig. 113*B*. The plate and grid currents that flow at any instant are a result of the combined action of the grid and plate voltages at that instant and can be determined from the static curves of the tube. The plate current that this action produces is illustrated at Fig. 113*C* and flows in the form of impulses that last less than a half cycle and come when the instantaneous voltage between plate and cathode is at a minimum, which is also the time when the instantaneous voltage on the grid is at a maximum. These impulses can be considered as being composed of a steady direct-current component upon which is superimposed an alternating current (usually of non-sinusoidal wave form), which when added to the direct-current component gives the actual current. The average value of this plate-current impulse when taken over a full cycle represents the direct-current component of the plate current that is indicated by a direct-current meter in series with the plate-power supply, while the variations of the actual plate current about this average value represent an alternating current that delivers alternating energy to the resonant circuit. The grid current also flows in impulses that last only a very small part of a cycle because the grid current flows only when the grid is positive. The value of this grid current when averaged over an entire cycle is the current that is read by a direct-current galvanometer in series with the grid, and which in flowing through the grid-leak resistance produces a voltage drop that supplies the necessary negative bias.

Power Relations.—The power supplied by the source of plate voltage is equal to the product of this voltage and the plate current and so varies as shown at Fig. 113*D*. Part of this power is delivered to the resonant circuit in the form of alternating-current energy, while the remainder is dissipated at the plate of the tube by the impact of electrons and appears in the form of heat that must be radiated. The instantaneous power supplied to the oscillator divides between the resonant circuit and the plate of the tube in the ratio of the instantaneous voltage drop across the part of the resonant circuit that is connected between plate and cathode to the instantaneous voltage drop between the plate and cathode elec-

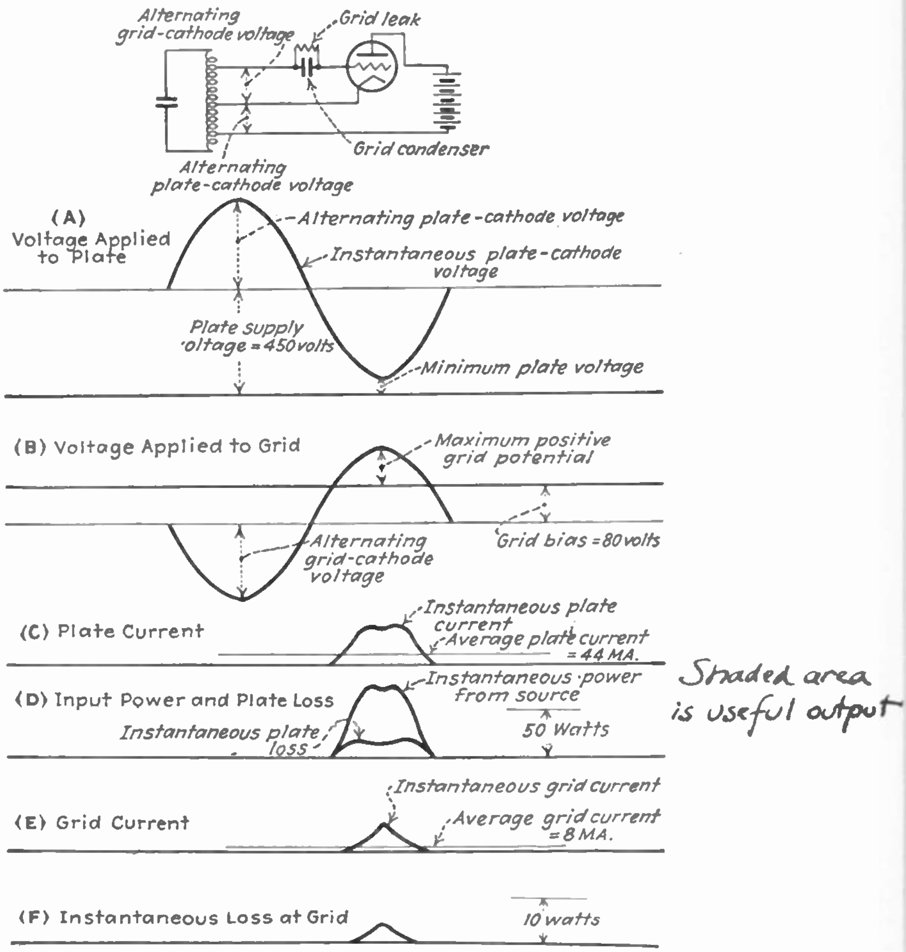


FIG. 113.—Series of oscillograms showing the voltage and current relations existing in a properly adjusted vacuum-tube oscillator.

trodes of the tube. The way in which the power supplied to the oscillator varies during the cycle, and the way in which this power divides between the resonant circuit and the plate of the tube, is shown at Fig. 113*D*. In order that the plate efficiency (*i.e.*, the ratio of power delivered to the oscillating circuit to total power supplied from the source of plate voltage) of the oscillator tube may be high, it is necessary that the plate current be allowed to flow only during that small portion of the cycle when the instantaneous plate-cathode voltage is at or very near its minimum value. Since this condition exists for only a very small portion of the cycle it is seen that a high efficiency requires that the plate current flow in the form of impulses of very short duration, but this gives a low power output since the average plate current, and hence the average power supplied to the oscillator, will be small if the current flows only during a very short interval. The result is that in every oscillator circuit it is necessary to compromise between amount of power output and efficiency with which the output is obtained. It is possible to develop a large output at a low or moderate efficiency, or a small output at a very high plate efficiency, but unless special circuits are provided it is impossible to obtain a large power output at a high efficiency.[1] Under ordinary conditions the plate current is permitted to flow during about one-third to one-sixth of the cycle. In this way an efficiency ranging between 50 and 80 per cent is obtained together with a reasonably large power output.

The plate circuit of a vacuum tube that is generating oscillations acts very much as though it constituted a synchronous switch that was closed for a short interval during each cycle in exactly the right phase to sustain the oscillations in the resonant circuit. The opening and closing of this switch in the actual oscillator are controlled by the alternating voltage applied to the grid, and the portion of the cycle during which the current flows is determined primarily by the grid bias. The effect of varying the grid bias while maintaining the alternating plate voltage substantially constant is shown in Fig. 114, where it is seen that decreasing the grid bias increases the fraction of the cycle during which the plate current flows and so lowers the plate efficiency. The length of the current impulses can be estimated from the fact that a bias corresponding to cut-off for the plate-supply voltage (*i.e.*, $E_c = E_B/\mu$, where E_B is the plate-supply voltage) allows the plate current to flow for exactly one-half of each cycle.

The plate efficiency and power obtained with a given plate current increase as the plate-supply voltage is increased. This is because the minimum plate voltage is approximately the same for a given plate current irrespective of the plate-supply voltage. Increasing the plate-supply voltage while keeping the plate current constant increases the power supplied to the tube in direct proportion to the voltage without changing the plate loss appreciably, with the result that a greater proportion of the power supplied to the plate is delivered to the resonant circuit. The effect of doubling the plate voltage is shown in Fig. 115.

[1] For an example of such a special circuit, see D. C. Prince and F. B. Vodges, A High Efficiency Vacuum-tube Oscillating Circuit, *Proc. I.R.E.*, vol. 12, p. 623, October, 1924.

Conditions Existing at Grid.—The grid of an oscillator is always allowed to become somewhat positive during a portion of each cycle, thus producing impulses of grid current that cause a power loss at the grid

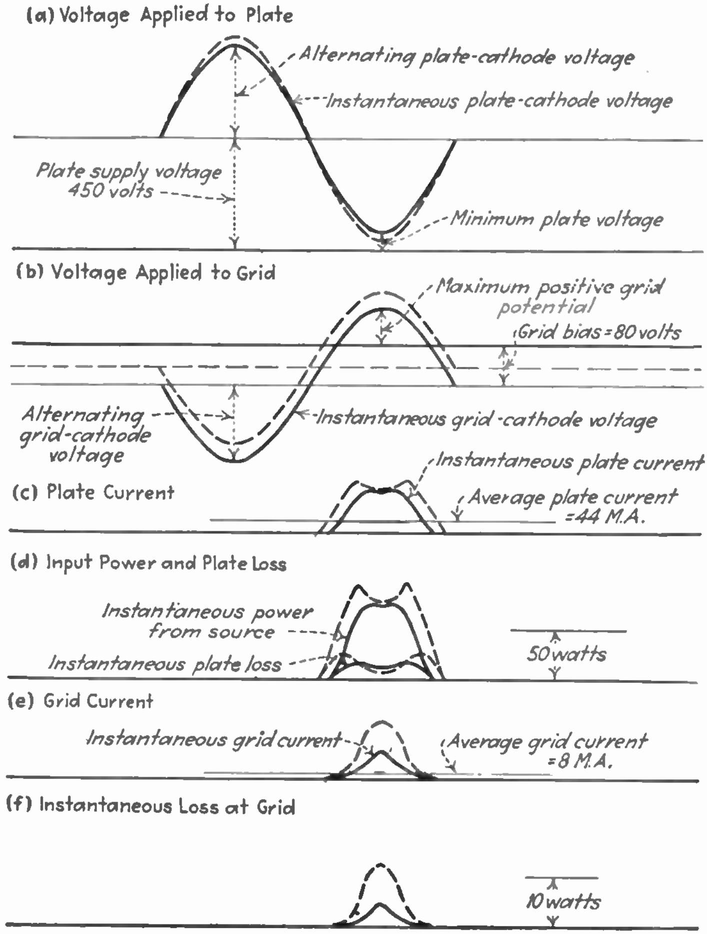


FIG. 114.—Series of oscillograms showing the effect which changing the grid-leak resistance has on the voltage and current relations of the vacuum-tube oscillator. The dotted curves are for the lower value of grid-leak resistance, which is seen to reduce the grid bias and to increase the amount the grid goes positive. The former action increases the fraction of the cycle during which plate current flows and reduces the efficiency, while the latter makes the grid current greater.

that at any instant is equal to the product of the instantaneous grid-cathode voltage and grid current. The power that is thus dissipated at the grid appears in the form of heat, generated by the impact of electrons against the grid, and varies during the operating cycle as shown at Fig. 113*F*. Permitting grid current to flow also results in an additional loss because the grid current heats the grid-leak resistance. These two power losses resulting from the flow of grid current are both supplied by energy obtained from the oscillations in the resonant circuit and

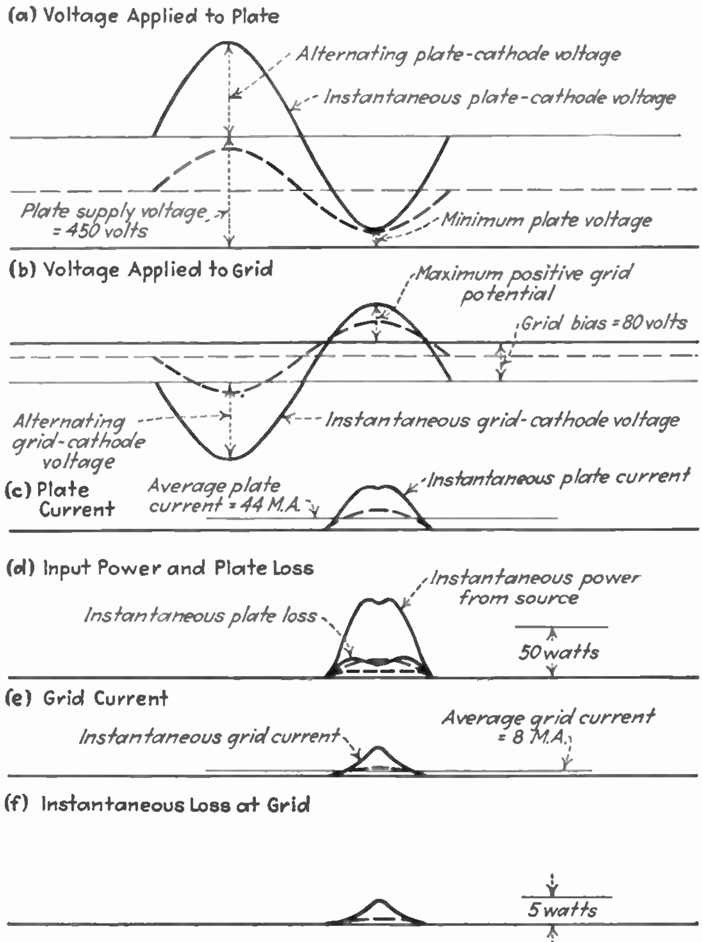


FIG. 115.—Oscillator oscillograms showing the effect of altering the plate-supply voltage. The dotted curves are for a plate-supply voltage one-half that represented by the solid lines. The low plate voltage is seen to reduce the plate current and the power output and also to give a lower efficiency.

represent a total loss that is very nearly equal to the product of the average grid current and the crest value of the alternating grid-cathode voltage.[1]

[1] This is because the grid current flows only when the alternating grid voltage is at or near its crest value, which makes the average energy taken from the oscillations approximate the product of the average grid current and the crest value of the alternating grid-cathode voltage.

While it might be thought desirable to avoid this loss by not allowing the grid to become positive, this is not the case because the grid goes positive at the part of the cycle when the plate current is flowing and thus increases the plate current that will flow with a given plate-cathode voltage. In this way the loss of energy at the plate of the tube can be made lower in proportion to the power supplied from the plate-voltage source than if the grid were not allowed to become positive, and this more than makes up for the grid losses. The extent to which the instantaneous grid potential can be permitted to go positive is limited by the fact that if the grid is more positive than the plate it will rob the plate of space current as shown in Fig. 116, thereby reducing the power output. For this reason it has been found best to limit the maximum positive grid-cathode voltage to not more than approximately 80 to 100 per cent of the minimum instantaneous plate-cathode voltage in all except low-power oscillator.

The grid current can be used to supply the necessary grid bias by requiring that it flow through a grid leak and grid condenser. The condenser acts as a by-pass for the alternating components of the grid-current impulses, while the direct-current component (*i.e.*, the average) is forced to flow through the grid-leak resistance and so puts the grid at a negative bias equal to the product of leak resistance and average grid current. The bias thus produced depends primarily upon the extent to which the grid goes positive (which is the principal factor determining the average grid current) and upon the resistance of the leak.

Factors Controlling the Amplitude of Oscillations.—The oscillations produced in the resonant circuit have an amplitude such that the energy consumed in the resonant circuit is just equal to the energy supplied to it from the source of plate power. The magnitude of the plate-current impulses, and hence the power delivered to the resonant circuit, depends upon the minimum plate voltage that is reached during the cycle and increases rapidly as the amplitude of oscillations is reduced because lowering the amplitude of oscillations increases the minimum plate voltage reached during the cycle, and this larger voltage draws more plate current. The relations are usually such that the full rated output is normally obtained with a minimum plate voltage that is relatively small compared with the plate-supply voltage. The fact that the amplitude of oscillations is such as to make the crest of the alternating plate voltage only slightly less than the plate-supply potential explains many of the characteristics of vacuum-tube oscillators. Thus the current in the resonant circuit is almost exactly proportional to the plate-supply voltage, and will vary with frequency in such a way as to maintain a nearly constant alternating plate voltage across the resonant circuit. With constant frequency the current in the resonant circuit will vary inversely with the effective impedance of the part of the resonant circuit that is connected between plate and cathode.

Varying the resistance of the tuned circuit has relatively little effect on the amplitude of the oscillations but does change the average plate current. When the resistance of the resonant circuit is increased it is impossible for the oscillations to maintain their amplitude, inasmuch

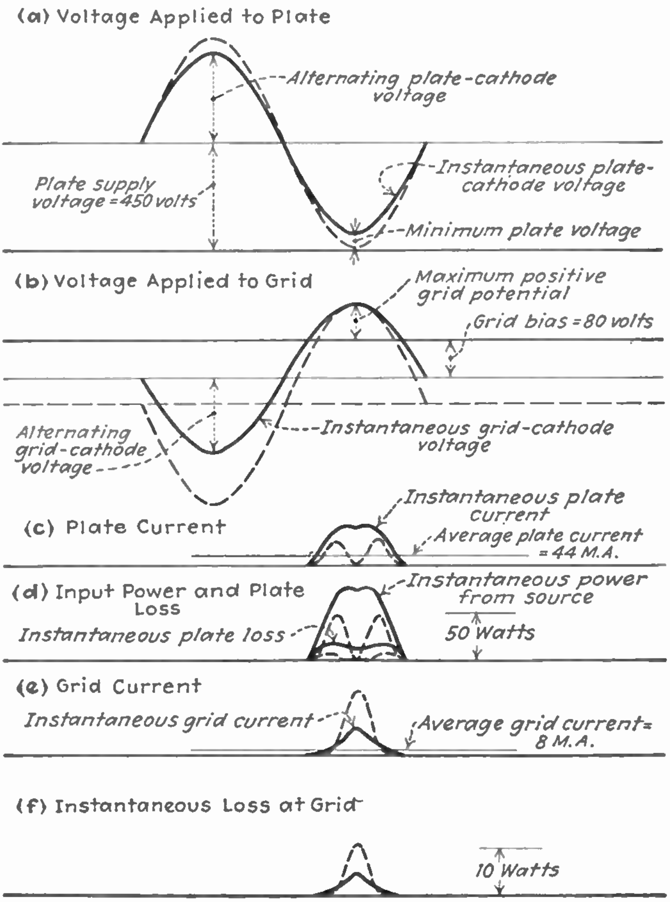


FIG. 116.—Oscillator oscillograms showing the effect of resistance in the resonant circuit. The dotted curves are for an extremely low resistance in the resonant circuit. Reducing the load resistance is seen to increase the amplitude of oscillations slightly, which reduces the minimum plate-cathode voltage to a very small quantity and thus lowers the plate current. An incidental effect is to make the minimum plate voltage much less than the maximum positive grid voltage, with the result that the grid robs the plate of the space current at the instant of minimum plate voltage.

as the original oscillating current in flowing through the added resistance causes more energy to be consumed in the resonant circuit than is supplied from the plate-voltage source. The result is that immediately upon the insertion of additional resistance the amplitude of the oscillations becomes less. This makes the minimum plate voltage larger, increasing the amplitude of the plate-current impulses, and resulting in the resonant circuit receiving additional energy. The amplitude of oscillations then assumes a new equilibrium point at which the enlarged plate-current impulses supply sufficient energy to the resonant circuit to maintain the oscillations with the higher resistance circuit. Inasmuch as the amplitude of the alternating plate-cathode voltage need decrease only a small percentage in order to increase greatly the amplitude of the plate-current impulses, the main result of adding resistance to the oscillating circuit is to increase the average plate current drawn from the plate-supply source, while the amplitude of the oscillations, the grid exciting voltage, and the grid bias are affected relatively much less, as is illustrated by the oscillograms shown in Fig. 116.

Series and Shunt (Parallel) Feed Circuits.—The discussion that has been given of the action taking place in the vacuum-tube oscillator has been based upon a series-feed arrangement. Identical performance is obtained if a shunt feed, such as shown in Fig. 112*b*, is used instead of the series feed. The only difference is that the shunt-feed connection separates the direct-current and alternating-current components of the plate-current impulses, permitting only the former to flow through the plate-supply source and only the latter through the resonant circuit. There is no difference in the power losses at the plate of the tube, or in the energy supplied to the resonant circuit, or in the way in which the instantaneous plate potential varies. A series of oscillograms similar to those of Fig. 113, but for the case of a shunt feed, are given in Fig. 117, and a comparison of these two figures brings out the similarities and differences in the action taking place with the two types of connection.

The Load Impedance.—The alternating-current energy developed by the oscillator is delivered to the resonant circuit and can be usefully employed by associating with the resonant circuit the load impedance that is to consume the power. In some cases this can be accomplished by connecting the load impedance in series with the resonant circuit, but it is usually more convenient to place the load impedance in a separate circuit that is coupled to the resonant circuit of the oscillator. This couples a resistance into the oscillator circuit, and the energy represented by the oscillating current flowing through the coupled resistance is the energy delivered to the load.

51. Adjustment and Design of Oscillator Circuits.—The principles that have been discussed in the preceding section form the basis on which oscillator circuits are designed and adjusted. The first step in adjusting an oscillator to operate at optimum conditions is to tune the resonant circuit to the required frequency. The coupling of this resonant circuit into the plate circuit of the tube is then varied until the desired oscillating current is obtained in the resonant circuit, remembering that increasing the coupling will cause the oscillating current to be lowered.

After this tentative adjustment of the alternating plate-cathode voltage has been made, the alternating grid-cathode voltage (*i.e.*, the grid excitation obtained by coupling to the resonant circuit) is varied until the plate current is a minimum, so that the plate efficiency will be as high as possible. In making this adjustment it is to be noted that when grid

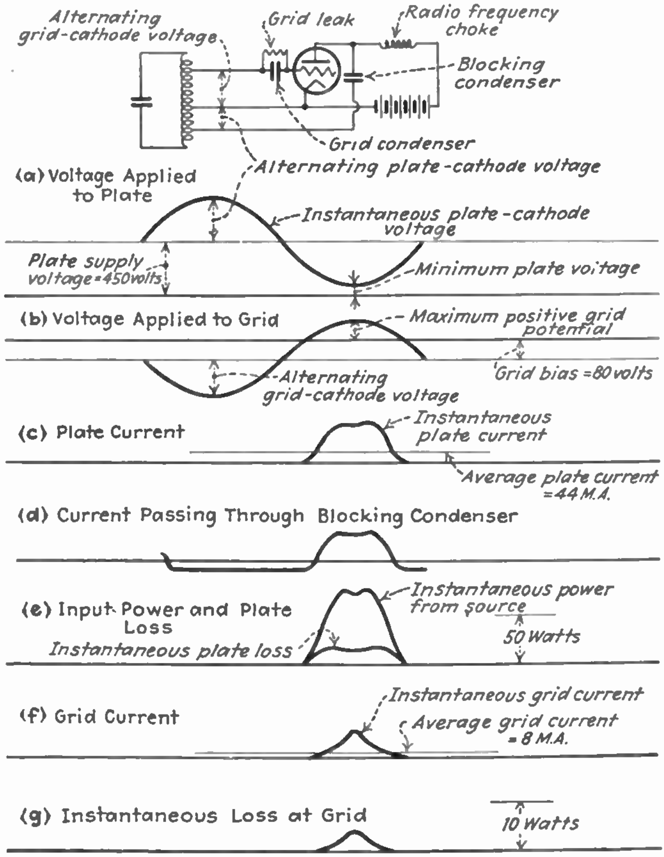


FIG. 117.—Oscillator oscillograms showing voltage and current relations when shunt feed is employed. These curves are identical with those of Fig. 113 except for the fact that separate paths are provided for the direct- and alternating-current components of the plate current.

leak and grid condenser are employed the grid bias is roughly proportional to the amplitude of the grid-exciting voltage, while the extent to which the grid goes positive is affected only very slightly by the grid excitation. As a consequence the principal effect of altering the grid excitation is to vary the fraction of the cycle during which plate current flows, and hence the efficiency, as is brought out by the curves of Fig. 118. The extent to which the grid goes positive is determined primarily by the grid-leak resistance, with a low-resistance leak causing the grid potential to go more positive since a low-resistance leak requires larger impulses of grid current to develop a given bias (*i.e.*, a given voltage drop), and this extra current requires an increased positive grid potential. The effect of altering the leak resistance is shown by the curves of Fig. 114. After

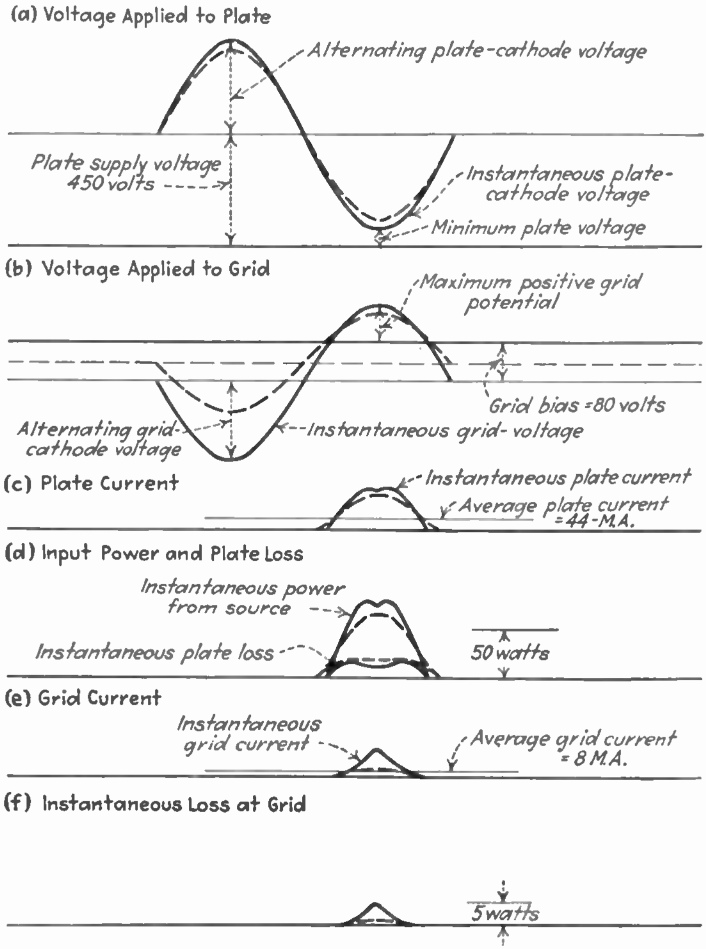


FIG. 118.—Oscillator oscillograms showing effect of varying the grid-exciting voltage. The dotted curves are for a low exciting voltage, which is seen to give a low grid bias without greatly reducing the amount by which the grid goes positive. The lower grid bias increases the fraction of the cycle during which the plate current flows and so lowers the efficiency.

these adjustments have been made the load impedance is then coupled into the resonant circuit by the proper amount to develop the full rated power output from the tube. After this has been done it is usually desirable to repeat this sequence of adjustments at least once in order to make sure that the very best operating conditions have been obtained.

Circuit Design.—In designing the circuits of a vacuum-tube oscillator one ordinarily knows the frequency, the plate-supply voltage, the required power output, and the tube amplification factor, and then proceeds to lay out the circuit to correspond to these factors. The first step is to decide upon the effective value of $\omega L/R$ to be used in the resonant circuit. Experience has shown that the value of $\omega L/R$ should not be less than about 12 if erratic and unstable operation is to be avoided, while values higher than this give excellent operation but unnecessarily increase the current in the resonant circuit and thus waste energy in heating the inductance coil. Since most of the effective resistance of the resonant circuit is the coupled resistance introduced by the load, the value of $\omega L/R$ is determined primarily by the load coupling. The coil resistance itself should be as low as possible, since the ratio of coil resistance to the sum of coil resistance and equivalent resistance contributed by the load represents the fraction of the energy delivered to the resonant circuit that is lost in the circuit itself. When the value of $\omega L/R$ for the oscillating circuit has been decided upon, it is then possible to determine the inductance of the resonant circuit in terms of the voltage desired across the circuit and power consumed in it according to the formula

$$\omega L = \frac{E^2}{P(\omega L/R)} \tag{103}$$

where

ω = 2π times frequency
L = inductance of tuned circuit
E = effective value of alternating voltage across tuned circuit
P = power delivered to tuned circuit
$(\omega L/R)$ = Q of tuned circuit, taking into account the effect of the load resistance on the tuned circuit.

With the tuned circuit determined in this way, the plate and cathode connections are then made so that when the total voltage across the tuned circuit has an effective value E, the crest value of the plate-cathode voltage will nearly, but not quite, equal the plate-supply voltage.

The desired grid bias can next be estimated in accordance with the fraction of the cycle during which it is desired to allow plate current to flow, and a grid-excitation voltage is then provided that has a crest value slightly greater than this grid bias. The grid bias is practically always supplied by the use of a grid leak and grid condenser instead of employing a grid-bias voltage obtained from a separate source. The grid leak and condenser arrangement does away with the necessity of providing an external voltage source and has the very great advantage of tending to make the oscillator self-adjusting. The self-adjusting feature results from the fact that the grid leak always produces a bias that will permit the grid to go positive by a small amount, irrespective of the amplitude of grid excitation, with the result that good operating conditions are realized more or less automatically. The use of a grid leak and condenser also makes the oscillator self-starting, as is discussed in Sec. 52. The grid condenser should have a capacity somewhat larger than the effective grid-cathode capacity of the tube and must also be large enough to have a reactance to the oscillation frequency that is small compared with the grid-leak resistance, but its exact size is unimportant. The exact size of the grid-leak resistance can be determined only by a long step-by-step calculation, but fortunately its value is not critical and can be estimated with satisfactory accuracy on the basis of experience.

In the shunt-feed circuit the only requirement that must be met by the blocking condenser is that it have a low reactance to the generated oscillations, and that it possess sufficient dielectric strength to withstand the direct-current plate-supply voltage. The choke required in the shunt-feed circuit in series with the plate-supply voltage serves to prevent the oscillations from flowing through the plate-supply source and so must have a high impedance to the generated oscillations. A radio-frequency choke of the type described in Sec. 10 is used for this purpose. In the series-feed arrangement it is customary to by-pass the plate-voltage source with a condenser having a low reactance to the oscillations being generated in order to short-circuit what internal impedance the power supply may possess.

The exact behavior of an oscillating circuit in regard to such quantities as plate efficiency, grid and plate currents, power output, etc., can be determined from the characteristic curves of the tube by a series of point-by-point calculations. These are seldom employed, however, because it is less work to set up the circuit according to the approximate methods that have been described and then to make minor readjustments as found necessary. When this has been done the desired quantities can be obtained experimentally from the oscillator performance.[1]

52. Miscellaneous Characteristics of Oscillating Circuits. *Starting of Oscillations.*—When a grid leak is used to produce the grid-bias voltage, oscillations start as soon as the plate voltage is applied because there will always be some variation or irregularity present to act as the initial impulse that starts the oscillations. When the grid is biased with a fixed negative voltage that exceeds the cut-off value for the plate-supply voltage, the oscillator is not self-starting because small irregularities cannot be amplified when no plate current is flowing, and the oscillator is hence unable to supply its own input, but oscillations once established will continue because enough energy is stored up in the resonant circuit to carry over from one plate-current impulse to the next.

[1] A systematic procedure for making these exact calculations is given by D. C. Prince, *loc. cit.*

Harmonics.—The plate current of an oscillator adjusted for high efficiency flows in impulses of short duration that depart from a sine wave shape and hence contain harmonics of the oscillator frequency. These harmonics are discriminated against by the resonant circuit, however, so that the alternating plate and grid voltages produced across portions of the tuned circuit are practically sinusoidal. Harmonics are always present in the output of an oscillator, particularly when adjusted to high efficiency; and where an especially pure wave is desired filtering by the use of selective circuits must be employed.

Intermittent Operation.—When a grid leak is used to produce the grid bias it is sometimes found that the oscillations are periodically interrupted. These interruptions may be at an audible rate, or may be so rapid as to represent a radio frequency, and can be prevented either by reducing the grid-leak resistance, the grid-condenser capacity, or both, or by readjusting the circuit proportions. Intermittent operation in oscillators biased with grid leak and grid condenser is the result of operating under conditions such that oscillations will die out if the grid-bias voltage is maintained at a constant value, and then using such a large time constant RC in the grid leak-condenser combination that the bias developed is not able to change as fast as the oscillations can die out. Under these conditions any infinitesimal reduction in the amplitude of oscillations causes a still further decrease in amplitude because the grid condenser momentarily maintains the bias constant, and this in turn produces a still greater reduction in the output because the grid leak cannot discharge the grid condenser as fast as the oscillations decay, and so on, with the result that the oscillations die out. The grid condenser then discharges until the grid bias is reduced to a point where oscillations can start, at which time the amplitude is built up to normal value and the cycle repeated. Conditions favorable for the production of interrupted oscillations are frequently encountered in power oscillators adjusted for high efficiency.

Oscillating Circuits with More than One Resonant Frequency.—When the circuits associated with an oscillator have more than one resonant frequency it is possible for the oscillator to operate at several frequencies. Where this possibility exists one of the frequencies ordinarily gets started first and suppresses the others unless the frequencies involved are of entirely different orders of magnitude. Two frequencies of oscillating current exist simultaneously in an oscillator adjusted to high efficiency only when one of the frequencies is many times greater than the other, in which case the higher frequency oscillations will exist only when the low-frequency oscillation is allowing plate current to pass; and will die out during the parts of the low-frequency cycle when the plate current is zero. The result is that if a high-frequency oscillation is produced it is modulated at the lower frequency of oscillation.

In radio-frequency oscillators trouble is often experienced from very high frequency oscillations occurring simultaneously with the desired oscillation. Such high-frequency currents are spoken of as *parasitic oscillations* and are the result of subsidiary resonant circuits that involve stray and tube capacities and lead inductances. They absorb energy that would otherwise go into producing currents of the desired frequency and so are to be avoided. Parasitic oscillations can be suppressed by placing resistances in locations that will not affect the desired oscillation and yet which will so increase the resistance of the parasitic resonant circuits as to prevent the parasitic oscillations from being produced. The location of such stabilizing resistances is usually rather hard to predict and must be determined by trial and error, although a resistance in series with the grid of the tube is usually helpful.

When a circuit tuned to the oscillator frequency is coupled to the resonant circuit of the oscillator there will be two resonant frequencies (see Sec. 17 and Fig. 38) and hence two possible oscillating frequencies unless the volt-amperes circulating in the secondary are less than the volt-amperes circulating in the primary (or oscillator) resonant circuit. If no means are employed to control the frequency, the oscillations will pick out the frequency having the lowest resistance and hence the one representing the lowest power output. If the effective resistance to the two possible frequencies is substantially the same, the frequency of the oscillations is a matter of chance, and any interruption or disturbance may cause the frequency to jump suddenly to the other possible value. In order that the volt-amperes in the secondary be less than in the primary the coefficient of coupling, k, between the primary and secondary tuned circuits must not exceed $1/Q$, where Q is the $\omega L/R$ of the secondary taking into account any coupled load that may be present. Stable operation of such an oscillator with a coupling greater than this can be obtained either by detuning one circuit slightly in order to favor one of the resonant frequencies or by obtaining the grid excitation from the secondary circuit, in which case the phase relations are such as to eliminate one of the otherwise possible frequencies.[1]

Blocking.—The phenomenon of blocking appears as a sudden stoppage of oscillations, accompanied by a reversal of grid current, and an increase of plate current to a value much higher than can be obtained with the full direct-current supply voltage and zero grid potential. A high power tube is usually destroyed by blocking, since the energy dissipated at the plate is enormous. Blocking is caused by operating conditions that permit secondary electron emission to take place at the grid to such an extent that the grid loses more electrons by secondary emission than it gains from the cathode by direct flow. This causes a reversal of the grid current, resulting in a positive grid-bias voltage being developed

[1] See D. C. Prince, *loc. cit.*, for a discussion of coupled oscillating circuits.

by the grid leak, and consequently an excessive plate current flows. In order that blocking may exist it is necessary that the minimum instantaneous plate voltage and the maximum instantaneous grid potential obtained during the cycle both be high, and that the grid leak have a high resistance. Blocking, when it occurs, is the result of attempting to force the output of the oscillator by increasing the load resistance coupled into the plate circuit. This results in a reduction in the amplitude of oscillations, which increases the minimum plate voltage. If the grid excitation is then increased, the secondary electron emission at the grid will be increased because of the increased positive potential reached by the grid. Under unfavorable conditions this will cause the net grid current to become less, which reduces the grid bias and makes the maximum positive grid potential still greater, causing a further reduction in the grid current, and so on. Blocking can be entirely avoided by taking reasonable care in the design of the circuits and tube.

Separately Excited Oscillators (Class C Amplifiers).—When the grid-exciting voltage is obtained from an external source instead of from the output of the oscillator, the tube is acting under circuit conditions identical with those existing in an oscillator but is actually functioning as an amplifier. Such an arrangement is called a separately excited oscillator, or a Class C amplifier. The distinguishing characteristic of a Class C amplifier is that the grid bias is greater than the value required to cut off the plate current. This gives a high efficiency in the conversion of direct-current plate-supply power to alternating-current energy but results in a non-linear relation between applied voltage and power output. Class C amplifiers always obtain the negative grid bias from an external potential source rather than by using grid leak and grid condenser. This is because the bias produced by leak and condenser depends upon the alternating-current voltage applied to the grid, and if this excitation is lost the grid assumes zero potential and an excessive plate current flows. The instantaneous grid potential in the Class C amplifier is ordinarily allowed to go positive during a small portion of each cycle just as is the case in oscillators, and for the same reason.

Synchronization of Vacuum-tube Oscillators.—Vacuum-tube oscillators have an inherent tendency to synchronize with any other oscillation of approximately the same frequency that may be present. The behavior of two oscillators loosely coupled together and generating frequencies that are not widely different illustrates what can be expected. If the two frequencies differ by only a small percentage they are both shifted from their normal values in such a way as to reduce the difference. This attraction of the two frequencies becomes more pronounced as the difference between the normal oscillating frequencies is reduced and finally becomes so great that the oscillators pull into synchronism. The extent to which the frequency of an oscillator can be shifted from its normal value by the presence of currents of a slightly different frequency will be greater as the strength of the injected currents is increased, and as the frequency stability of the oscillator is lowered.[1]

It has also been found that when currents having a frequency approximating a harmonic of the generated frequency are injected into the oscillator circuits there is a tendency for the generated oscillation to synchronize its frequency with the injected currents in such a way that a harmonic of the generated frequency is in exact synchronism with the injected frequency. The tendency toward synchronization with a

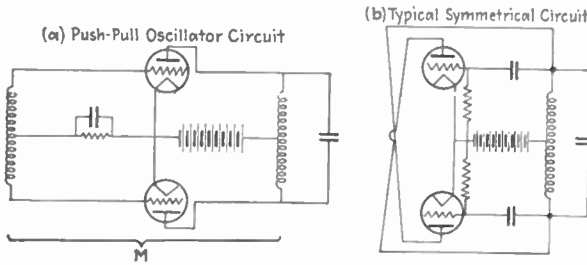


FIG. 119.—Typical symmetrical two-tube oscillator circuits. At *a* is a push-pull circuit of the reversed feed-back type while *b* is a symmetrical oscillator derived from the Hartley circuit.

harmonic of the generated oscillation is not so great as when both frequencies are approximately the same, but when the frequency stability of the oscillator whose frequency is being controlled is relatively low, and the injected energy is relatively strong, it is not difficult to maintain the oscillations at a frequency that is exactly one-half or one-third of the injected frequency, and synchronization has been obtained when the ratio of frequencies is as low as one-sixth.[2]

Oscillators with More than One Tube.—Since vacuum-tube oscillators tend to synchronize automatically there is no difficulty in operating two or more tubes in parallel, and such arrangements can be used where more power is desired than can be obtained from a single tube. It is however usually preferable to use a single tube of larger capacity where this is possible, because parasitic oscillations are very difficult to avoid when tubes are operated in parallel.

The most successful multitube arrangements are those employing two tubes in either a push-pull or some other symmetrical connection, such as shown in Fig. 119, since with such arrangements there is less

[1] A mathematical analysis of this phenomenon of synchronization for one particular type of oscillating circuit is given in the article by E. V. Appleton, Automatic Synchronization of Triode Oscillators, *Proc. Cambridge Phil. Soc.*, vol. 21, p. 231, 1922.

[2] Synchronization between oscillations related by harmonics is discussed in the paper by Isaac Koga, A New Frequency Transformer or Frequency Changer, *Proc. I.R.E.*, vol. 15, p. 669, August, 1927. Also see U. S. Patent 1,527,228 issued to Schelleng.

tendency for parasitic oscillations to be produced. Symmetrically arranged oscillators using two tubes are often employed where the frequency to be obtained is extremely high because they permit the use of very short connecting wires and in effect connect the electrode capacities of the two tubes in series, which increases the highest frequency to which the resonant circuit can be tuned.

53. Frequency of Generated Oscillations.—The alternating current generated by the vacuum-tube oscillator has a frequency such that the voltage which the generated oscillations apply to the grid of the tube is

See Kusanose - IRE Feb 1932
Less deviation for:
high C tank
high Rg
high Rp

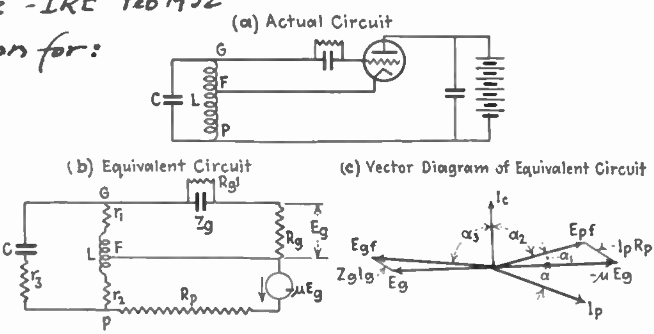


FIG. 120.—Equivalent circuit and vector diagram for determining the frequency of oscillation of the Hartley oscillating circuit. The circuit resistances require that the actual frequency be slightly less than the resonant frequency of the tuned circuit in order to make E_g be exactly 180° out of phase with $-\mu E_g$. In this diagram the tube capacities have been neglected for the sake of simplicity.

of exactly the proper magnitude and phase to produce the oscillations that supply this grid-exciting voltage. This approximates the resonant frequency of the tuned circuit, but the exact value depends upon plate-supply voltage, grid bias, grid-exciting voltage, resistance of tuned circuit, etc., and can be determined with the aid of an equivalent oscillator circuit such as that shown at Fig. 120. In reference to this equivalent circuit, the frequency of the generated oscillations must be such that the equivalent voltage $-\mu E_g$ acting in series with R_p will produce a grid-exciting voltage E_g that is exactly 180° out of phase with $-\mu E_g$. In setting up the equivalent circuit the equivalent plate resistance R_p is taken as the value required for the voltage $-\mu E_g$ to produce the alternating plate voltage E_{pf} actually existing, and will depend upon grid excitation, tuned-circuit resistance, plate-supply voltage, grid bias, etc. Similar considerations lead to a definition of R_g as that resistance which represents the ratio of the exciting voltage applied to the grid of the tube divided by the component of the grid current that is of the frequency of the generated oscillations. The resistance R_g is determined primarily by grid bias, grid excitation, and the minimum plate voltage. When a grid leak is used to obtain the bias, R_g is approximately one-half the grid-leak resistance and tends to be constant with changes in plate voltage, amplitude of oscillations, etc. (Sec. 61, page 288, footnote 1).

Analysis of Equivalent Oscillator Circuit.—The way in which the various circuit elements affect the frequency of the oscillations can be determined by a vector diagram of the equivalent circuit, such as the vector diagram of Fig. 120. This diagram must be drawn so that the grid-exciting voltage E_g produced in the equivalent circuit by the equivalent plate voltage $-\mu E_g$ be exactly 180° out of phase with the latter voltage. In the case of the Hartley circuit this requires that the tuned circuit offer an inductive reactance to the generated frequency, while with the Colpitts circuit the tuned circuit must have a capacitive reactance. In the vector diagram of the Hartley circuit, which is shown in Fig. 120, the plate current I_p therefore lags behind the equivalent plate voltage $-\mu E_g$ by a small angle α, while the alternating plate voltage E_{pf} leads $-\mu E_g$ slightly as a result of the voltage drop of I_p in the effective plate resistance R_p. The alternating plate voltage E_{pf} is applied to the resonant circuit and produces a current I_c through the condenser that leads E_{pf} by an angle α_2 that is a little less than 90° because of the resistances r_1, r_3, R_{gl}, and R_g. The voltage E_{gf} produced by this current leads the current by an angle α_3 which is slightly less than 90° because of the resistances r_1, R_{gl}, and R_g. The grid-exciting voltage E_g is then E_{gf} less the voltage drop I_gZ_{gl} of the grid current I_g in the impedance Z_{gl} of the grid leak-condenser combination. In most cases the difference between E_{gf} and E_g is very small.

In studying the vector diagram of Fig. 120 it will be seen that in this case the frequency of the oscillations must be less than the resonant frequency of the tuned circuit in order to produce the phase shift α_1 between $-\mu E_g$ and the alternating plate voltage E_{pf} that compensates for the fact that the circuit elements r_1, r_2, r_3, R_{gl} and Z_g make the vectors E_{pf} and E_g slightly less than 180° out of phase. The amount by which the generated frequency is less than the resonant frequency of the tuned circuit depends upon the tuned-circuit resistance $r_1 + r_2 + r_3$, impedance Z_g of grid leak and condenser, grid resistance R_g, and plate resistance R_p. The result is that the generated frequency will vary with grid excitation, load resistance coupled into resonant circuit, plate-supply voltage, grid-leak resistance, etc., since all of these things have an influence on at least one factor entering into the vector diagram of the oscillator.

In determining the resonant frequency of the oscillator-tuned circuit it is necessary to include capacities between tube electrodes and other capacities that may be present and that are in effect associated with the resonant circuit. These various capacities while usually small in magnitude will have an appreciable effect on the generated oscillations unless the tuning capacity is extremely large. It is also necessary to take into account any reactance which the load impedance may couple into the resonant circuit, inasmuch as such reactance is effectively a part of the resonant circuit and will help determine its resonant frequency. In the case of shunt-feed circuits the shunt-feed choke and the blocking condenser are also associated with the tuned circuit and affect the resonant frequency.

Frequency Stability of Oscillators.—When an oscillator is turned on and allowed to operate over a period of time it is ordinarily found that the frequency does not stay exactly constant. The factors contributing to this instability are variations in the circuit constants, changes in the impedance which the load couples into the oscillator circuit, and variations in the tube characteristics.

Nearly all variations in constants of the circuits external to the tube are caused by temperature effects and can be minimized by proper design, and in extreme cases by temperature control. Frequency variations resulting from load changes can be made as small as desired by coupling the load extremely loosely to the oscillating circuit, or by using a buffer tube, preferably a screen-grid tube, between oscillator and load, so that the load cannot react upon the oscillator circuits.

The most important and also the most troublesome causes of frequency instability are those which arise within the tube in the form of changes in the effective grid resistance R_g and effective plate resistance R_p caused by the grid, plate, and filament voltages of the tube, amplitude of oscillations, condition of tube, etc. The values of grid resistance R_g and plate resistance R_p that exist in an oscillator are determined by the fact that the amplitude of oscillations increases or decreases until these resistances assume values such that the alternating-current energy generated by the oscillator is exactly equal to the energy which is consumed in the equivalent circuit of the oscillator. This equilibrium point is reached when the ratio R_g/R_p has some particular value determined only by the tuned circuit and the amplification factor of the tube, and the oscillator searches out the particular amplitude that will give this equilibrium condition. When the grid bias has a fixed value, as when a battery is employed, equilibrium results from the fact that as the amplitude of oscillations increases the grid becomes more and more positive, causing the grid resistance R_g to decrease rapidly without producing corresponding variations in the plate resistance R_p. On the other hand, when the grid bias is obtained from a grid condenser and leak, the grid resistance R_g tends to be constant irrespective of the amplitude of oscillations, while the effective plate resistance R_p increases with increased amplitude since the stronger the oscillations the more negative will be the bias produced by the grid leak and the higher will be the plate resistance.

Methods of Stabilizing the Frequency.—Efforts to make the frequency less dependent upon variations in the tube characteristics can be divided into two main classes: First those in which the emphasis is placed on minimizing the changes produced in R_g and R_p by variations in electrode potentials, and second those in which the circuit is arranged in such a way as to make the frequency of oscillation independent of R_g and R_p.

Consider now means of frequency stabilization in which the changes in tube parameters R_g and R_p are kept as small as possible. It will be observed that since with a given tube and tuned circuit the ratio R_g/R_p is a constant at all equilibrium points, it is merely necessary to insure that one of these quantities remain constant, as the other will then take care of itself. The effective grid resistance R_g can be kept substantially constant by obtaining the bias from a grid leak and grid condenser in which the grid-leak resistance is the greatest that can be employed without causing intermittent oscillations. The reason for the constancy of the grid resistance when a grid leak and condenser is used results from the fact that under these conditions the effective grid resistance R_g is approximately half the grid-leak resistance, as demonstrated in Sec. 61, page 288, footnote 1.

The effective plate resistance R_p can be maintained substantially constant by placing a high resistance between the plate and the tuned circuit. The effective plate resistance is then the sum of the tube resistance and the added resistance, so that if the latter, which is independent of electrode voltages, is made much larger than the former, the sum of the two will be substantially constant and will lead to conditions favorable for the production of a constant frequency.

One of the most important factors increasing the stability of the generated frequency is the use of a tuned circuit with a high Q. This requires that the tuned circuit itself have low losses and that the amount of energy transferred to the load be negligibly small. A resonant circuit having a high Q gives a large phase shift in the impedance across its terminals with small changes in frequency and so requires only a small variation in frequency to produce phase shifts sufficiently large to compensate for the effects resulting from changes in the tube characteristics. There is a limit however to the improvement and stability that can be gained from a high Q circuit, since even an infinite Q does not necessarily insure an absolutely constant frequency. This is because the grid resistance R_g and plate resistance R_p are associated with the tuned circuit and cause it to have an effective Q that can never be infinite. The best that can be done is to use the highest value of Q that is possible and then to operate the tube in such a way that the effective plate resistance R_p is high. This will make R_g likewise large and will reduce the effective Q by the smallest possible amount.

Circuits in Which the Generated Frequency Is Independent of Tube Characteristics.—Investigation has shown that if the actual Q of the tuned circuit is infinite, *i.e.*, if the tuned circuit can be assumed to have zero losses, it is possible to make the generated frequency absolutely independent of the tube resistances R_g and R_p provided suitable phase-shifting networks are placed between the tube and the tuned circuit.[1] Thus in the Hartley oscillating circuit shown at *a* in Fig. 121 the insertion in the plate circuit of the condenser C_5 having the size indicated in the figure will make the frequency theoretically independent of the conditions existing within the tube. This method of stabilization is based on the fact that the ratio R_g/R_p has a fixed value at the equilibrium condition, and that the alternating currents flowing into the grid and out of the plate have a definite ratio that is independent of tube conditions. It is then possible by the use of a suitable phase-shifting impedance to produce just the necessary phase shift to make the generated frequency independent of the tube constants.

Figure 121 shows a number of stabilized oscillator circuits of this type, together with the relations that are required to give stability. When the circuit proportions are as indicated, the frequency is absolutely independent of the tube voltages provided the actual oscillator circuit is identical with that shown in the figure and that the actual Q of the tuned circuit is extremely high. In applying the circuits of Fig. 121 it must be remembered that at all except the lower radio frequencies the electrode capacities of the tube will modify the results and prevent complete stabilization unless the circuit is of such a character as to make it possible to include these electrode capacities as part of the oscillator network.

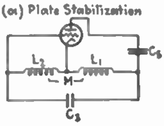
While the above discussion applies strictly to the ideal case where the Q of the oscillator circuit is infinite the results are qualitatively the same when the oscillator tuned circuit has a reasonably high but entirely practical value of Q. Under such conditions the effect of the finite Q is to alter slightly the value of stabilizing impedance required and to prevent the frequency from being entirely independent of tube characteristics. Practically however the analysis derived on the basis of ideal conditions will still give qualitatively correct results.

Practical Circuits Having High Frequency Stability.—The best method of applying the principles that have been discussed to practical oscillators depends upon the frequency being generated. At low frequencies, where coils with magnetic cores can be employed, the preferred method of obtaining high frequency stability is by the use of the first two circuits shown in Fig. 121*a*. Here the effective plate resistance R_p of the tube

[1] For an excellent discussion of oscillators in which the frequency is independent of tube characteristics, see F. B. Llewellyn, Constant-frequency Oscillators, *Proc. I.R.E.*, vol. 19, p. 2063, December, 1931. A discussion of oscillators arranged to minimize changes in tube characteristics is given by J. W. Horton, Vacuum-tube Oscillators—A Graphical Method of Analysis, *Bell System Tech. Jour.*, vol. 3, p. 508, July, 1924.

HARTLEY OSCILLATOR

(a) Plate Stabilization

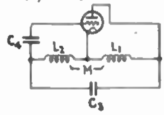


$$C_5 = C_3\left[\frac{L_0}{L_1 + L_2A^2 - 2MA}\right]$$

$$L_0 = L_1 + L_2 + 2M$$

$$A = \frac{L_1 + M}{L_2 + M}$$

(b) Grid Stabilization

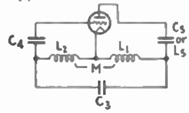


$$C_4 = C_3A^2\left[\frac{L_0}{L_1 + L_2A^2 - 2MA}\right]$$

$$L_0 = L_1 + L_2 + 2M$$

$$A = \frac{L_1 + M}{L_2 + M}$$

(c) Plate and Grid Stabilization



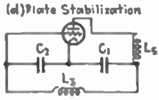
$$\frac{1}{C_5} + \frac{A^2}{C_4} = \frac{1}{C_3}\left[\frac{L_1 + L_2A^2 - 2MA}{L_0}\right]$$

$$L_5 = L_0\frac{C_3}{C_4}A^2 - L_1 - L_2A^2 + 2MA$$

$$L_0 = L_1 + L_2 + 2M;\ A = \frac{L_1 + M}{L_2 + M}$$

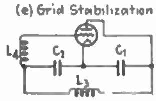
COLPITTS OSCILLATOR

(d) Plate Stabilization



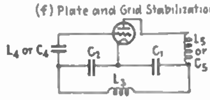
$$L_5 = L_3\frac{C_2}{C_1}$$

(e) Grid Stabilization



$$L_4 = L_3\frac{C_1}{C_2}$$

(f) Plate and Grid Stabilization

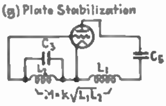


$$L_4\left(\frac{C_2}{C_1}\right) + L_5\left(\frac{C_1}{C_2}\right) = L_3$$

$$L_5 = L_3\frac{C_2}{C_1}\left[1 + \frac{C_2}{C_4}\left(\frac{C_2}{C_1 + C_2}\right)\right]$$

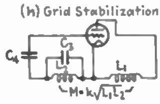
FEED-BACK OSCILLATOR

(g) Plate Stabilization



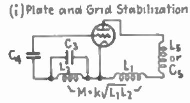
$$C_5 = C_3\frac{L_2}{L_1}\left(\frac{1}{1 - k^2}\right)$$

(h) Grid Stabilization



$$C_4 = C_3\left(\frac{k^2}{1 - k^2}\right)$$

(i) Plate and Grid Stabilization

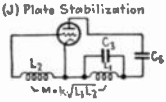


$$L_5 = L_1\left[k^2\left(1 + \frac{C_3}{C_4}\right) - 1\right]$$

$$C_5 = C_3\frac{L_2}{L_1}\left[\frac{1}{1 - k^2\left(1 + \frac{C_3}{C_4}\right)}\right]$$

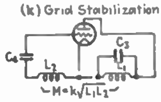
REVERSED FEED-BACK OSCILLATOR

(J) Plate Stabilization



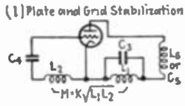
$$C_5 = C_3\left(\frac{k^2}{1 - k^2}\right)$$

(k) Grid Stabilization



$$C_4 = C_3\frac{L_1}{L_2}\,\frac{1}{1 - k^2}$$

(l) Plate and Grid Stabilization

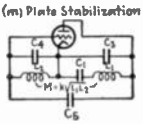


$$L_5 = L_1\left[1 + \frac{1}{k^2}\left(\frac{L_1C_3}{L_2C_4} - 1\right)\right]$$

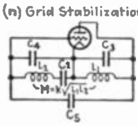
$$C_5 = \frac{C_3}{\frac{1}{k^2}\left(1 - \frac{L_1C_3}{L_2C_4}\right) - 1}$$

UNITY COUPLED OSCILLATOR

(m) Plate Stabilization

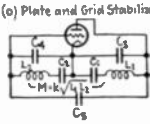


(n) Grid Stabilization

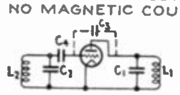


$$C_1 = \left(\frac{L_2}{L_1}\right)\left[\frac{C_3k^2\frac{L_1}{L_2} + C_4 + C_5\left(1 + k\sqrt{\frac{L_1}{L_2}}\right)^2}{1 - k^2}\right] \qquad C_2 = \left(\frac{L_1}{L_2}\right)\left[\frac{C_4k^2\left(\frac{L_2}{L_1}\right) + C_3 + C_5\left(1 + k\sqrt{\frac{L_2}{L_1}}\right)^2}{1 - k^2}\right]$$

(o) Plate and Grid Stabilization



(p) TUNED PLATE-GRID OSCILLATOR WITH NO MAGNETIC COUPLING



$$\omega^2k^2L_1L_2 = \left(\omega L_1 - \frac{1}{\omega C_1}\right)\left(\omega L_2 - \frac{1}{\omega C_2}\right) \qquad C_1 = \frac{L_2}{L_1}\left[C_2 + \frac{(1 + \mu)C_3C_4}{C_4 + (1 + \mu)C_3}\right] - C_3$$

FIG. 121.—Circuits and circuit proportions that will make the generated frequency independent of the tube constants provided the Q of the resonant circuit is very high (theoretically infinite).

is kept constant by means of a high series resistance, while the generated frequency is made substantially independent of tube conditions by employing a high Q tuned circuit with a unity coefficient of coupling between the tuned circuit and the grid-exciting coil. The series resistance in the plate circuit (often called the feed-back resistance) insures the minimum possible variation in tube constants, while the unity coupling between grid and tuned circuit satisfies the theoretical condition required for making the oscillator frequency independent of what variations

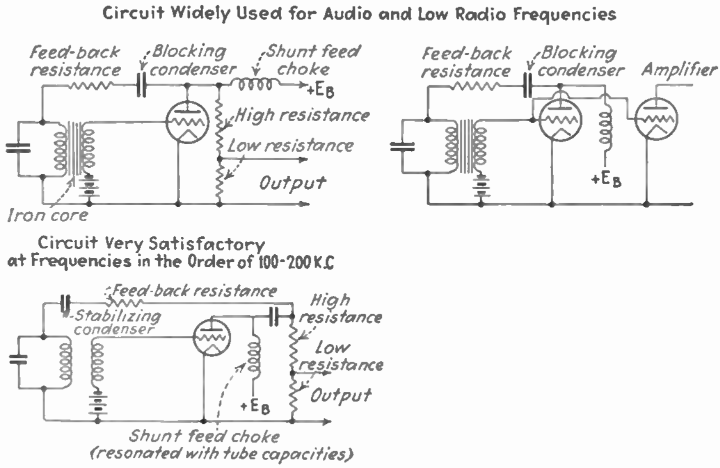


FIG. 121a.—Practical circuits used at low and moderately low frequencies when high frequency stability is desired.

there are in the tube constants R_g and R_p. A unity-transformation ratio is usually, but not always, employed between the two coils. In order to obtain the best results the ratio of feed-back resistance to parallel impedance which the tuned circuit develops between the plate and filament connections should be only slightly less than the amplification factor of the tube, while the feed-back resistance itself should at the same time be at least twice, and preferably ten to one hundred times, the plate resistance of the tube. The usual method of adjusting the resistance is to use the highest possible value that will sustain the oscillations over the range of plate voltages to be encountered. This not only gives high frequency stability, but also insures good wave form.

At higher frequencies, where coils with magnetic cores cannot be employed, the first two circuits of Fig. 121a are not satisfactory because it is impossible to obtain unity coupling between the primary and secondary of an air-cored transformer. At moderate radio frequencies, such as 100 to 200 kc, it is hence better to employ one of the stabilized circuits shown in Fig. 121 in combination with a moderate feed-back resistance in series with the plate of the tube. There is ordinarily little trouble from tube capacities at these frequencies, especially if a shunt feed choke is resonated with the plate-cathode capacity of the tube in the middle of the frequency range to be generated. A practical circuit of this type intended for use at 100 kc is shown at Fig. 121*a*. The series plate resistance keeps the effective R_p virtually independent of electrode voltages, while the stabilizing condenser in the plate circuit makes the frequency largely independent of what variations do take place in the tube constants.

At higher radio frequencies, such as the broadcast and higher frequencies, it is desirable to use the stabilized tuned plate-tuned grid circuit or a stabilized unity-coupling circuit in order that the electrode capacities may be made an integral part of the oscillator network. At these higher frequencies it is necessary to dispense with the stabilizing plate resistance because of the tube capacities. Instead, the grid bias should be obtained by means of a grid-leak and grid-condenser combination having a reasonably high grid-leak resistance and the largest time constant RC that will not cause intermittent oscillations.

Oscillators adjusted to deliver large power outputs at a good efficiency cannot be depended upon to maintain their frequency constant to closer than 1 per cent over long periods of time under the ordinary changes in operating conditions that can be expected. In contrast with this the stable frequency oscillator circuits of Fig. 121*a* are capable of generating oscillations that are but little affected by changes external to the resonant circuit. Thus variations in the plate-supply voltage of 10 per cent will not produce a frequency change greater than one part in several thousand, while variations in other conditions of operation have corresponding effects. The result is that such an oscillator will maintain its frequency constant to better than 0.1 per cent over long periods of time if the ambient temperature is constant.

Frequency Range.—The range of frequencies obtainable from tube oscillators is limited only by the frequency range that can be covered by resonant circuits having a reasonably high Q. The lowest frequency is limited by the difficulty of obtaining low-frequency resonant circuits with low resistance, while the highest frequency is determined by the electrode capacities of the tube, which are part of the resonant circuit and fix the minimum tuning capacity that is present. With small commercial tubes the limit is about 100,000 kc and is reached when the resonant circuit consists of the inductance of the shortest possible leads tuned by the tube capacities. It is possible to obtain slightly higher frequencies with the push-pull connections of Fig. 119*a* and *b* because these connections permit the use of shorter connecting wires and at the same time place the electrode capacities of the two tubes in series. Even with tubes of special construction it is very difficult to generate frequencies higher than 300,000 kc (1 meter wave length).[1]

[1] For further information see C. R. Englund, The Short Wave Limit of Vacuum-tube Oscillators, *Proc. I.R.E.*, vol. 15, p. 914, November, 1927.

54. Characteristics of Tubes Suitable for Use in Oscillators.—The tubes used in oscillators are essentially amplifier tubes that have size as their distinguishing characteristic. When the amount of power to be generated does not exceed a few watts it is possible to employ the ordinary small vacuum tubes commonly used in radio receivers; but larger tubes are required for greater powers. The amount of power that a tube can handle is determined by the plate voltage that may be applied to the tube with safety, by the electron emission of the cathode, and by the amount of power that can be dissipated within the tube without overheating.

Filament Size.—The plate current that an oscillator tube can draw is determined by the electron emission of the cathode and by the fraction of the cycle during which the plate current is allowed to flow. The maximum instantaneous current that can flow cannot exceed the emission of the cathode, but since this maximum current flows for only a fraction of the cycle the average value of the plate current is much less than the maximum. It is therefore apparent that the smaller the fraction of the cycle during which the plate current flows the greater must be the electron emission from the cathode if a given average plate current is to be obtained. In order to allow ample factor of safety the electron emission of the cathode when the tube is new is commonly from fifty to one hundred times the rated value of average, *i.e.*, direct-current, plate current.

Voltage Requirements.—The voltage that is applied to the plate of the tube consists of the direct-current plate-supply potential plus an alternating component that has a crest value approximately equal to the direct-current voltage, so that during operation the instantaneous voltage between plate and cathode varies from nearly zero to nearly twice the plate-supply potential. The voltage that can be applied safely to the plate is limited by the plate-cathode insulation and the degree of vacuum existing within the tube, and in order to withstand potentials ranging from 1000 to 20,000 volts it is necessary that extreme care be taken in the design and construction of the tube.

In tubes generating alternating current of very high frequencies, *i.e.*, above 3000 kc, the alternating component of the plate potential produces large dielectric losses in the glass walls because of the very high frequency and the relatively high temperature of the glass during operation of the tube. The result is that when a tube is operated at a high plate voltage and is generating very high frequencies there is a tendency for the dielectric losses in the glass to produce local overheating that may soften the glass and destroy the vacuum.[1] Vacuum tubes which are to operate with very high plate voltage must therefore provide

[1] A discussion of punctures resulting from dielectric losses in the glass walls of the tube is to be found in the article by Yujiro Kusunose, Puncture Damage through the Glass Wall of a Transmitting Vacuum Tube, *Proc. I.R.E.*, vol. 15, p. 431, May, 1927.

ample plate insulation and be arranged to minimize the dielectric stress in the glass walls. Even with the best designs it is usually necessary to use a lower plate voltage when operating at extremely high frequencies than when operating at low or moderate frequencies.

Heat Energy to Be Dissipated.—The plate power supplied to an oscillating tube is equal to the product of plate-supply voltage and average plate current and so is limited by the direct-current plate voltage that may be used without danger of a breakdown, and by the allowable direct-current plate current. A certain fraction of this power supplied to the plate, usually more than half, is delivered to the resonant circuit in the form of alternating-current energy, while the remainder appears at the plate in the form of heat which the plate must be capable of radiating to the walls of the tube without becoming excessively hot. There is also a somewhat smaller power loss at the grid which the grid must be capable of radiating without reaching an excessive temperature. The total power dissipated inside the tube consists of these grid and plate losses plus the power used in heating the cathode and must be carried away through the outside wall of the tube. Tubes generating 1 kw. or less of alternating-current power can transfer the energy loss in the tube to the surrounding air with glass walls of reasonable size, but since glass softens at relatively low temperatures the amount of energy that can be radiated per unit area of glass surface is low, and larger tubes would have to be enclosed in glass bulbs of prohibitive size. The cooling of tubes with glass walls is ordinarily obtained by allowing free circulation of air although in the largest sizes a forced draft supplied by a fan is sometimes employed, and occasionally the tubes are immersed in oil.

A large proportion of the energy dissipated in tubes having glass bulbs is produced at the plate, which must therefore be capable of radiating without damage all the heat generated at its surface plus that fraction of the filament-heating power which the filament radiates to the plate. In order to facilitate this radiation of energy from the plate to the glass walls of the tube it is desirable to blacken the plate to increase the rate of heat radiation, and to use a material that will stand relatively high temperatures. In some types of tubes the plates are actually at a dull red heat when operating under normal conditions.

Water-cooled Tubes.—In tubes having power ratings in excess of 1 kw. the problem of carrying away the energy dissipated in the tube has been solved by using water-cooled plates. Such tubes can be made in several ways. One type of construction is shown in Fig. 122 and employs a cylindrical copper plate which is dropped into a water jacket through which cooling water is circulated. The plate serves as part of the wall of the tube as well as acting as an anode, and since it is in direct contact with the cooling water many kilowatts can be dissipated at the plate without an appreciable rise in temperature. Another type of construction that has been commercially successful encloses the tube with a glass bulb but employs a hollow plate through which water is circulated by means of pipes sealed through the glass. The most important feature of these water-cooled tubes is the metal-to-glass seal, which can be made in a number of ways.[1]

While water-cooled plates effectively eliminate the problem of dissipating the energy developed at the plate there is still the energy dis-

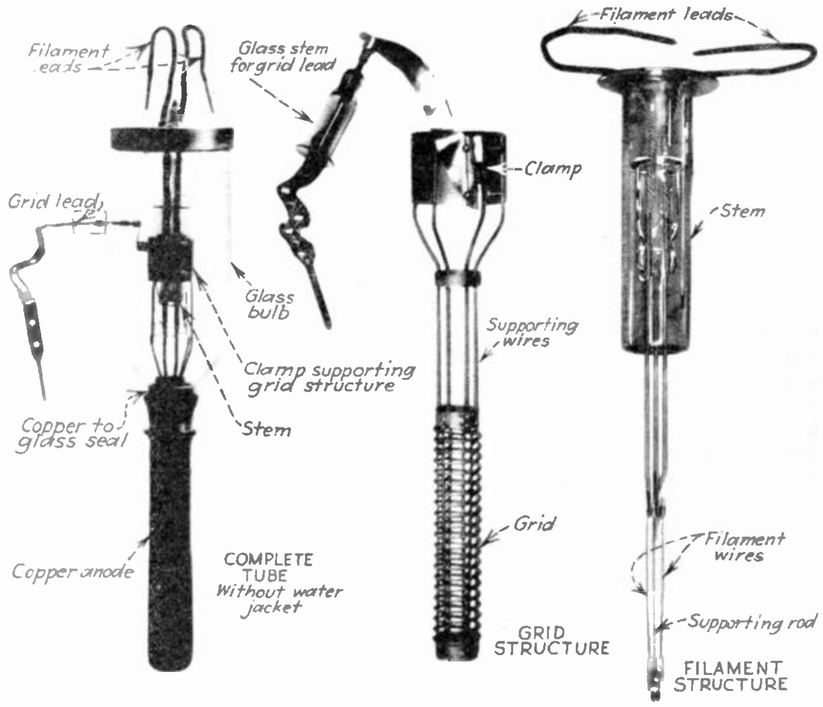


FIG. 122.—A water-cooled type of vacuum tube having a power rating of 20,000 watts output when acting as an oscillator. The anode is made in the form of a copper cylinder which is cooled by immersion in a water-jacket.

sipated at the grid of the tube, which while much less than that developed at the plate, is so great in large tubes as to cause the grid to operate at a red heat under normal conditions and may even heat the grid to a temperature at which thermionic emission of electrons takes place. Tubes with water-cooled grids have been devised and while not now in commercial use will probably be employed when ratings in excess of several hundred kilowatts are reached.

[1] There are a number of ways of sealing metal to glass, some of which have been known for many years. For a description of a type of seal that has been extensively used in water-cooled tubes see: William G. Housekeeper, The Art of Sealing Base Metals through Glass, *Trans. A.I.E.E.*, vol. 42, p. 870, 1923.

Emitters.—The electron emitter used in high-power tubes operating at high plate voltages is practically always a tungsten filament, while oxide-coated and thoriated-tungsten filaments, though having a high thermionic efficiency, are employed only in the smaller tubes. This is because of the effects produced by the gas molecules left in the tube after evacuation. The electrons in traveling from cathode to plate ionize these gas molecules by collision, producing positive ions that bombard the cathode with a velocity that is very great when the plate potential is high. While the number of gas ions produced is small because of the high vacuum, their velocity is so great in tubes operating at a high plate voltage as to strip the thorium layer from the surface of thoriated-tungsten emitters faster than the layer can be replaced by diffusion, and to cause rapid disintegration of oxide-coated cathodes. With the oxide-coated emitters troubles from gas are aggravated by the fact that the oxide coating is damaged if the tube electrodes are heated to temperatures sufficiently high to give up all occluded gas. For these reasons tungsten filaments are always used in the largest tubes, while thoriated-tungsten and oxide-coated filaments are employed in moderate-size tubes designed to operate at moderate plate voltages.

The amount of power consumed in heating the cathode of a high-power vacuum tube is very large because the required electron emission is great, and because the emitter is usually tungsten. Data on the cathode heating power of a number of typical tubes designed for oscillator purposes are to be found in Table VII, which shows that the cathode power is in excess of 1 kw. in the largest tubes.

In large tubes it is customary to place a resistance in series with the filament when the filament circuit is first closed in order to limit the rush of current that would otherwise flow because of the low resistance of the cold filament. Without this starting resistance the initial current in large tubes would burn out fuses and might even damage the lead wires passing through the glass seal.

Construction and Rating of Tubes.—One of the most difficult problems encountered in the manufacture of high-voltage, high-power tubes is the production of the vacuum. The air originally in the tube can be readily pumped out, but the removal of the gas that is occluded in the metal and glass parts of the tube that are in contact with the vacuum is not a simple matter. Such solid materials will absorb large quantities of gas which clings very tenaciously and can be removed only by the application of heat. All metal parts entering into the tube construction are always degassed by being heated to a high temperature before being placed in the tube, and the pumping procedure is carried out with the entire tube in an oven that is heated to a temperature just below the softening point of the glass in order to remove as much as possible of the gas occluded in the glass parts. Finally, while the tube is still on the pump, the metal parts are brought to temperatures above those that would be produced during normal operation. By taking these precautions it is possible to produce a suitable vacuum with a reasonable certainty that it will be maintained throughout the operating life of the tube. The difficulty of obtaining such a vacuum can be understood when it is realized that the

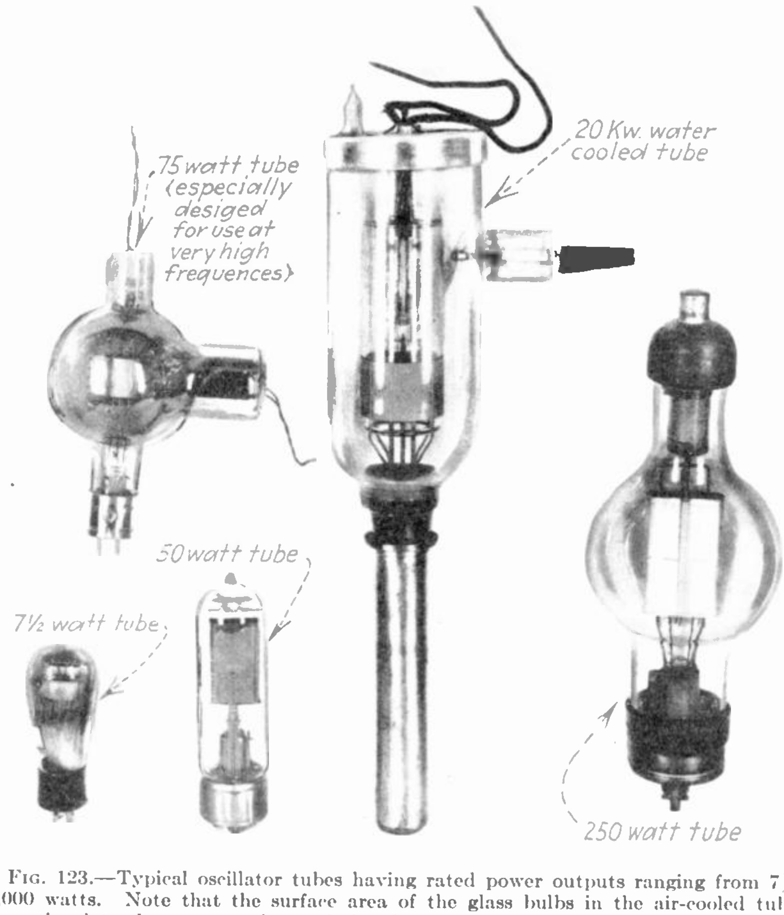


FIG. 123.—Typical oscillator tubes having rated power outputs ranging from $7\frac{1}{2}$ to 20,000 watts. Note that the surface area of the glass bulbs in the air-cooled tubes is proportional to the power rating and that in the larger tubes (*i.e.*, those requiring high plate voltages) the plate leads are brought out through separate seals remote from the grid and filament seals.

large water-cooled tubes require continuous pumping for approximately 24 hr. to remove the occluded gas.

Materials suitable for the plate and grid electrodes of high-power vacuum tubes are few in number because of the high temperatures to which these electrodes may be subjected during operation. The grid is generally of tungsten because of the refractory nature and high mechan-

TABLE VII.—CHARACTERISTICS OF TYPICAL OSCILLATOR TUBES

Type	Nominal rating, watts output	Maximum allowable plate loss, watts	Maximum allowable direct-current plate voltage, volts	Maximum allowable direct-current plate current, milliamperes	Filament data				Amplification factor	Type of cooling
					Type	Volts	Amperes	Watts		
210	7½	15	425	60	Thoriated	7½	1.25	9	8	Air
203-A	50	100	1,250	175	Thoriated	10	3.25	32	25	Air
852	75	100	3,000	100	Thoriated	10	3.25	32	12	Air
204-A	250	250	2,500	275	Thoriated	11	3.85	42	25	Air
206	1,000	350	15,000	100	Tungsten	11	14.75	162	350	Air
207	20,000	10,000	15,000	2,000	Tungsten	22	52	1,144	20	Water
862	100,000	100,000	20,000	10,000	Tungsten	33	207	6,831	48	Water

NOTE: With most tubes the allowable plate dissipation will be exceeded if the maximum allowable plate current is drawn when the maximum allowable plate voltage is applied to the tube.

ical strength of this material, while molybdenum is also sometimes used. The plates of air-cooled tubes are nearly always made of molybdenum for the reason that this material is capable of being operated at a high temperature and can be readily worked. Tantalum is occasionally used for the plates of power tubes and has the desirable property of absorbing gas in a certain range of temperatures. Tantalum plates thus help maintain the vacuum by "cleaning up" any gas that may be given out by the metal and glass parts during operation. The plates of water-cooled tubes need meet no special requirements inasmuch as the operating temperature is low. Copper is generally employed because of its high thermal conductivity, but other materials can be used, and in at least one instance commercial tubes have been built using welded steel plates.

The amplification factors of power tubes do not differ greatly from those of ordinary receiving tubes, although the tendency is to use somewhat higher values because of the much greater plate voltages involved. The plate resistance of large tubes is somewhat lower than for small tubes having similar amplification factors, but the differences are not striking since the increased size of the tube increases the spacing as well as the electrode area. The chief distinguishing features of high-power tubes as compared with small receiving tubes are the very high voltages that the tubes can stand, the large space currents that they are capable of producing, and their ability to dissipate large power losses. The principal characteristics of a representative series of power tubes are given in Table VII, and a number of the same tubes are shown in Fig. 123. It will be noted in the latter figure that the size of glass bulb of the air-cooled tubes is proportional to the power rating, and that while the low-power tubes employ a type of construction similar to that used in receiving tubes, the tubes intended for service at high voltages are arranged so that the plate connection enters the tube through a special seal remote from the grid and cathode connections in order to make the insulation strength as great as possible. In many tubes the grid connection is also brought out through a separate seal, which has the advantage of reducing the electrode capacities and increasing the insulation strength still more.

Oscillator tubes are rated on the basis of power output under conservative operating conditions. Thus a 50-watt oscillator tube will develop 50 watts of radio-frequency output under ordinary conditions. In most cases, particularly with small tubes, ratings are purely conventional, being made on the basis of an anode efficiency of 50 per cent with a conservative figure taken as the allowable plate loss. Hence a 75-watt tube has a nominal allowable plate loss of 75 watts but can be operated safely with a 100-watt loss, and so at a plate efficiency of 67 per cent will develop 200 watts of radio-frequency output.

A number of special problems are involved in oscillator tubes intended to be used at very high frequencies. In addition to the necessity of avoiding excessive dielectric loss in the glass walls of the tube, tubes intended for high-frequency service must have low interelectrode capacities in order to allow for external tuning capacity in the resonant circuits. To obtain this low capacity the electrode areas must be small and the spacing greater than usual, while the different electrodes are preferably mounted from different presses. The 75-watt tube of Fig. 123 is an example of a tube embodying these constructional features. The grid and plate lead-in wires of tubes used at very high frequencies must be capable of carrying heavy currents, because at the frequencies involved the capacity currents flowing to the grid and plate may reach many amperes even though the interelectrode capacities are small.

55. Crystal Oscillator.[1]—The frequency stability of an oscillator can be made very high by replacing the usual resonant circuit with a mechanically vibrating piezo-electric quartz crystal and utilizing the piezo-electric effect to obtain the connection between the electrical circuits and the mechanical vibrations. Quartz crystals from which the piezo-electric material used in crystal oscillators is obtained come in the form of crystals which when perfect have a hexagonal cross section and pointed ends, as illustrated in Fig. 124.[2] The properties of such a crystal can be

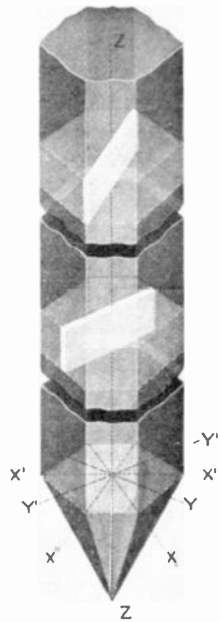


FIG. 124.—Illustrations howing the natural quartz crystal and the relation of the electric or X, the mechanical or Y, and the optical or Z axes to the crystal structure. The upper section shows a Y (or 30°) cut plate while the plate in the center section is X (or Curie) cut. The third Y-axis $Y''Y''$ is not shown because the perspective of the drawing makes it coincide with the ZZ-axis.

[1] The literature on crystal oscillators and related piezo-electric phenomena is too extensive to be covered with any thoroughness in a book of this type. The following selected bibliography is recommended to the reader desiring additional background: W. G. Cady, The Piezo-electric Resonator, *Proc. I.R.E.*, vol. 10, p. 83, April, 1922; Bibliography on Piezo-electricity, *Proc. I.R.E.*, vol. 16, p. 521, April, 1928; A. Crossley, Piezo-electric Crystal-controlled Transmitters, *Proc. I.R.E.*, vol. 15, p. 9, January, 1927; August Hund, Uses and Possibilities of Piezo-electric Oscillators, *Proc. I.R.E.*, vol. 14, p. 447, August, 1926; Note on Quartz Plates, Air-gap Effect, and Audio-frequency Generation, *Proc. I.R.E.*, vol. 16, p. 1072, August, 1928; F. R. Lack, Observations on Modes of Vibration and Temperature Coefficients of Quartz-crystal Plates, *Proc. I.R.E.*, vol. 17, p. 1123, July, 1929; K. S. Van Dyke, The Piezo-electric Resonator and Its Equivalent Network, *Proc. I.R.-E.*, vol. 16, p. 742, June, 1928. Crystal oscillators are the invention of Dr. W. G. Cady.

[2] There are many crystalline substances which have piezo-electric properties, such as Rochelle salts, tourmaline, quartz, etc., but of these quartz is used exclusively in crystal oscillators because of its cheapness, mechanical ruggedness, and low temperature coefficient.

expressed in terms of three sets of axes. The axis joining the points at the ends of the crystal is known as the optical axis, and electrical stresses applied in this direction produce no piezo-electric effect. The three axes X', X'', and X''' passing through the corners of the hexagon that forms the section perpendicular to the optical axis are known as the electrical axes, while the three axes Y', Y'', and Y''', which are perpendicular to the faces of the crystal, are the mechanical axes.

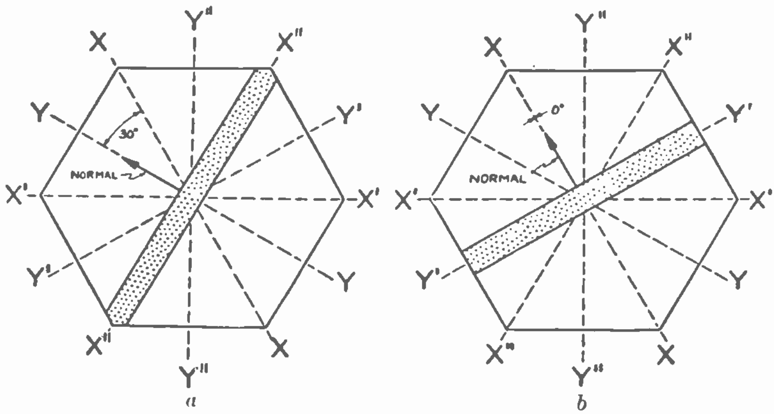


FIG. 125.—Cross sections of the quartz crystal shown in Fig. 124 taken in planes taken perpendicular to the optical axis ZZ. The plate at a has the Y cut (also called 30° cut) because its face is perpendicular to a Y (*i.e.*, mechanical) axis, while the plate at b is an X cut plate (Curie cut) because its face is perpendicular to an X (*i.e.*, electric) axis.

If a flat section is cut from a quartz crystal in such a way that the flat sides are perpendicular to an electrical axis as indicated in Fig. 125*b* (X or Curie cut) it is found that mechanical stresses along the Y-axis of such a section produce electrical charges on the flat sides of the crystal section. If the direction of these stresses is changed from tension to compression or *vice versa* the polarity of the charges on the crystal surfaces is reversed. Conversely, if electrical charges are placed on the flat sides of the crystal by applying a voltage across these faces, a mechanical stress is produced in the direction of the Y-axis. This property by which mechanical and electrical properties are interconnected in a crystal is known as the piezo-electric effect and is exhibited by all sections cut from a piezo-electric crystal. Thus, if mechanical forces are applied across the faces of a crystal section having its flat sides perpendicular to a Y-axis, as in Fig. 125*a* (which is known as the Y or 30° cut), piezo-electric charges will be developed because forces and potentials developed in such a crystal have components across the Y- and X-axes, respectively.

When an alternating voltage is applied across a quartz crystal in such a direction that there is a component of electric stress in the direction of an electric axis, alternating mechanical stresses will be produced in the direction of the Y- (or mechanical) axis which is perpendicular to the X-axis involved. These stresses will cause the crystal to vibrate, and if the frequency of the applied alternating voltage approximates a frequency at which mechanical resonance can exist in the crystal the amplitude of the vibrations will be very large. In the vicinity of such a resonant frequency the current that is drawn by the crystal as a result of the vibrations is exactly the same current that would be drawn by a series circuit composed of resistance, inductance, and capacity. In addition to this current representing the vibrational characteristics of the piezo-electric crystal there is also a component of leading current resulting from the electrostatic capacity between the points of application of the exciting voltage.

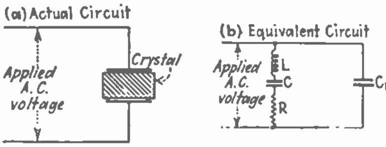


FIG. 126.—Equivalent electrical network that represents the effect which a vibrating quartz crystal has on the electrical circuits associated with it.

Equivalent Electrical Circuit of Quartz Crystal.—As far as the electrical circuits associated with the vibrating crystal are concerned, the crystal can be replaced by the electrical network of Fig. 126, in which C_1 represents the electrostatic capacity between the crystal electrodes when the crystal is not vibrating, and the series combination L, C, and R represents the electrical equivalent of the vibrational characteristics of the material.[1] The inductance L is the electrical equivalent of the crystal mass that is effective in the vibration, C is the electrical equivalent of the effective mechanical resilience, while R represents the electrical equivalent of the coefficient of friction. The frequency at which L and C are in series resonance is also the frequency of mechanical resonance. The electrical energy drawn by the equivalent L-C-R series circuit represents energy which the electrical circuit supplies to maintain the crystal vibrations. Below resonance this energy contains a leading reactive component because the elastic forces control the crystal vibration, while above resonance the inertia is the dominating factor and lagging reactive energy is required to sustain vibrations. At resonance the vibrations consume no reactive energy, and the crystal consumes power at unity power factor. Electrical circuits involving piezo-electric crystals can therefore be analyzed by replacing the crystal with its equivalent electrical network and then determining the behavior of the resulting circuit.

[1] The fact that a vibrating quartz crystal can be replaced by an equivalent electrical network was first discovered by Dr. H. J. Ryan and his co-workers during war researches on submarine detection carried on in 1918, but the report representing this work has not yet been released by the military authorities. The equivalent electrical circuit of the vibrating piezo-electric crystal has since been discovered by several other investigators. For further information see K. S. Van Dyke, The Piezo-electric Resonator and Its Equivalent Network, *Proc. I.R.E.*, vol. 16, p. 742, June, 1928.

The magnitudes of L, C, R and C_1 that enter into the equivalent electrical network of the vibrating quartz crystal depend upon the way in which the crystal is cut, the size of the crystal, and the type of vibration involved. The numerical values can be calculated from the crystal dimensions for the simpler modes of vibrations, and typical values for several crystals are given in Table VIII.[1]

TABLE VIII.—EQUIVALENT ELECTRICAL CHARACTERISTICS OF TYPICAL QUARTZ CRYSTALS

Crystal No.	Dimensions, cm.			Type of cut	Type of vibration	Resonant frequency, kc	Equivalent electrical quantities				
	t	w	l				L, henrys	C, $\mu\mu f$	R, ohms	C_1, $\mu\mu f$	Q, approximately
I	0.15	3.0	0.40	X	Width	About 90	137	0.0235	About 7,500	3.54	10,300
II	0.25	2.5	2.5	X	Thickness	About 1,100	0.33	0.065	About 2,700	1.0	844
III	0.636	3.33	2.75	X	Thickness	451.5	3.656	0.0316	9,036	5.755	1,147

The frequency at which mechanical resonance takes place is the frequency at which L and C are in resonance, and the magnitude of the resonance effect is determined by the ratio $\omega L/R$ of the equivalent electrical network (*i.e.*, by the equivalent Q of the crystal). The outstanding characteristics of the crystal vibrator are that the resonant frequency varies inversely with the dimensions of the crystal in the direction in which the principal vibration is taking place, that the ratio L/C (*i.e.*,

[1] In the case of X- (Curie-) cut crystals in which the principal vibration is in the direction of the X-axis (thickness vibration) the electrical quantities are given to a fair approximation by the following formulas:

$$L_t = 130\frac{t^3}{lw} \text{ henrys} \tag{104a}$$

$$C = 0.0022\frac{lw}{t}\ \mu\mu f \tag{104b}$$

$$C_1 = 0.40\frac{lw}{t}\ \mu\mu f \tag{104c}$$

$$R = 130{,}000\frac{t}{lw} \tag{104d}$$

When the principal vibration is in the direction of the Y-axis (width vibration) the formulas for C, C_1, and R are the same, while the equivalent inductance L becomes

$$L_\omega = 130\frac{wt}{l} \tag{104e}$$

The dimensions w, l, and t are measured in centimeters in the directions of the Y-axis parallel to the surface of the crystal, of the X-axis perpendicular to this Y-axis, and the Z-axis, respectively.

the stiffness), of the equivalent quartz resonator is enormously higher than could conceivably be obtained with coils and condensers, and finally that the effective Q of the crystal vibrator varies inversely with frequency and is extremely high with a low resonant frequency.

Quartz Oscillators.—Since the vibrating quartz crystal is equivalent to a resonant circuit it can be used as the frequency-controlling element in a vacuum-tube oscillator in place of the usual tuned circuit. There are a number of ways in which this can be done, one of the most common of which is shown in Fig. 127. In this circuit the piezo-electric crystal is connected between the grid and cathode of the tube, while an inductive reactance (ordinarily obtained by tuning a parallel-resonant circuit to a frequency slightly higher than the resonance frequency of the crystal) is provided in the plate circuit. The input impedance of a tube operated in this way with an inductive load impedance consists of a capacity

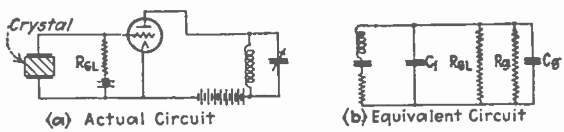


FIG. 127.—Circuit of a crystal oscillator together with equivalent grid circuit of tube. When the tuned circuit in the plate is resonant at a frequency slightly higher than the resonant frequency of the crystal the tube input resistance R_g is negative and neutralizes the resistance of the crystal, and the crystal is set into vibration.

shunted by a negative resistance as explained in Sec. 43. When this negative resistance is sufficiently low it neutralizes the resistance of the electrical network representing the crystal, and oscillations are produced at a frequency of mechanical resonance in the crystal. These oscillations result from the fact that the piezo-electric voltages developed by the crystal vibrations are applied to the grid of the tube and produce amplified power in the plate circuit. A portion of this amplified energy is then fed back through the grid-plate tube capacity and is supplied to the crystal in the proper phase to maintain the vibrations.

Frequency Stability of Quartz Oscillators.—The outstanding property of an oscillator in which the tuned circuit is supplied by a vibrating crystal is that the frequency generated is remarkably constant under varying conditions in the circuits associated with the crystal. One important factor contributing to this high frequency stability is the fact that a quartz vibrator has a relatively high Q, particularly when the resonant frequency of the crystal is a low frequency, in which case the crystal Q is many times that obtainable with ordinary electrical circuits. Another feature contributing to the frequency stability of the quartz oscillator is the high stiffness possessed by the equivalent electrical network of the vibrating crystal. As a result of this high ratio of L/C the coupling between the quartz crystal and the associated electrical circuits is extremely low, and these circuits consequently have relatively little effect on the resonant frequency of the quartz vibrator. This is apparent when it is noted that changes in the capacity C_1, which shunts the crystal vibrator, have negligible effect on the resonant frequency of the equivalent circuit, since even a few micro-microfarads is practically equivalent to a short circuit compared with the capacitive reactance of the equivalent crystal capacity C. Another valuable property of the crystal vibrator is that quartz is an extremely durable material, which is unchanged by ordinary handling and which furthermore has a very low temperature coefficient, so that the properties of a quartz crystal are affected to a much smaller degree by aging, handling, and temperature

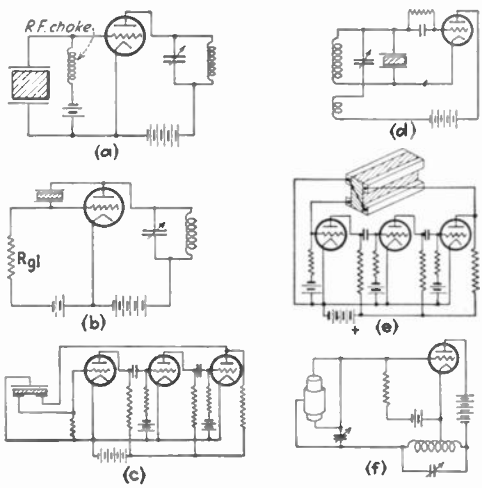


FIG. 128.—Typical crystal oscillator circuits. The first four set up longitudinal vibrations while e and f set up flexural and torsional vibrations, respectively.

changes than are the properties of ordinary coils and condensers. All of these factors combine to cause the crystal oscillator to generate a frequency that depends almost solely on the dimensions of the quartz crystal and which is relatively independent of the circuit constants, tube characteristics, or electrode voltages associated with the crystal oscillator. The exact influence which such factors exert on the frequency of the crystal oscillator can be determined by replacing the crystal by its equivalent electrical network and then analyzing the resulting system of electrical circuits by the methods of Sec. 53.[1] It is possible to make the frequency practically independent of the tube by using one of the stabilized circuits in Fig. 121, the quartz crystal taking the place of one of the tuned circuits.

[1] For an excellent example of this type of analysis see Earle M. Terry, The Dependence of the Frequency of Quartz Piezo-electric Oscillators upon Circuit Constants, *Proc. I.R.E.*, vol. 16, p. 1486, November, 1928.

Circuits for Producing Different Modes of Vibration.—A plate cut from a quartz crystal represents an extremely complex vibrating system having a large number of resonant frequencies. These frequencies depend upon the size of the crystal plate, the orientation of its crystal axes, and the way in which the crystal is excited, and may be classified into three general types, namely longitudinal (or straight tension or compression across the crystal plate), flexural (or bending), and torsional (or twisting) vibrations.

Self-oscillations may be set up in piezo-electric crystals by many vacuum-tube arrangements, a few of which are shown in Fig. 128. The circuit at Fig. 128*a* is a modification of Fig. 127 in which the grid leak has been replaced by a radio-frequency choke, while Fig. 128*b* represents a type of oscillator in which the tuned circuit is connected between grid and plate (sometimes called the ultra-audion circuit). The circuit shown at Fig. 128*c* is unique in that the input and output circuits of the vacuum-tube amplifier associated with the crystal are coupled through the crystal by the use of two pairs of electrodes. The circuit at Fig. 128*d* is an ordinary oscillator circuit in which a piezo-electric crystal has been shunted across the tuning condenser. Under such conditions the oscillator frequency will tend to stabilize at the resonant frequency of the crystal, and over a limited range of adjustments is substantially independent of the resonant frequency of the circuit formed by the coil and condenser. These first four circuits of Fig. 128, together with the arrangement given in Fig. 127, set up longitudinal vibrations in the crystal, while the last two circuits of Fig. 128 show arrangements for producing flexural and torsional vibrations, respectively.[1] With most of these circuits it is possible to excite several resonant frequencies in the crystal, and the particular one which appears depends upon the circuit adjustments.

X- and Y-cut Crystals.—Crystal oscillators used in practice nearly always make use of longitudinal vibrations in crystal plates cut from the natural crystal in such a way as to approximate the X (Curie) cut or the Y (30°) cut, and generally employ the circuit of Fig. 127 or a modification thereof. In the X-cut plate there are two principal response frequencies, one high and one low. The high frequency is a function of the thickness of the plate in the direction of the electric axis and is given to a good approximation by the equation

$$\text{Thickness frequency of } X\text{-cut plate} = \frac{2.860 \times 10^6}{t} \qquad (105a)$$

[1] See J. R. Harrison, Piezo-electric Resonance and Oscillatory Phenomena with Flexural Vibrations in Quartz Plates, *Proc. I.R.E.*, vol. 15, p. 1040, December, 1927; August Hund and R. B. Wright, New Piezo-electric Oscillators with Quartz Cylinders Cut along the Optical Axis, *Proc. I.R.E.*, vol. 18, p. 741, May, 1930.

where t is the thickness in millimeters as measured in the direction of the X-axis perpendicular to the crystal surface.

The low frequency is determined by the width of the crystal in the direction of the Y, or mechanical, axis and is given by Eq. 105*a* by substituting for t the width w of the crystal in millimeters as measured in the direction of the Y-axis. The temperature coefficient of these two resonant frequencies is negative and has a value of approximately twenty parts in a million per degree centigrade. The width and thickness vibrations are established by the use of the same circuits, and the one which is obtained in any particular case depends upon the adjustments.

In the Y (or 30°) cut crystal there are also two principal resonant frequencies corresponding to those obtained with the X-cut, but while the width frequency is given by Eq. (105*a*), the thickness frequency depends to a certain extent upon the ratio of thickness to width and also possesses a number of other peculiar properties. When this ratio of thickness to width is very small the high frequency is a function of the thickness t (*i.e.*, thickness as measured in the direction of the Y-axis) and is given approximately by the equation

$$\text{Thickness frequency in } Y\text{-cut crystal} = \frac{1.96 \times 10^6}{t} \qquad (105b)$$

The thickness frequency of the Y-cut crystal often appears as a doublet consisting of two frequencies differing by a relatively small percentage. There are also certain thicknesses at which no thickness resonant frequency is found, and varying the thickness of the crystal while maintaining the other dimensions constant tends to cause the resonant frequency to vary in discontinuous steps. Furthermore the thickness resonant frequency is affected by the width of the crystal, and for certain widths the thickness frequency disappears. These effects are particularly pronounced when the thickness and width are of the same order of magnitude, and the explanation for the peculiar behavior is that the width and thickness vibrations in the Y-cut crystal are coupled to each other so that when the two dimensions are of comparable magnitude there are certain ratios of thickness to width at which the width vibration, or one of its harmonics, is the same as the thickness resonant frequency. Thus the crystal can be considered as involving two or more coupled circuits, and when the crystal size is such that one of the secondary circuits has the same resonant frequency as does the thickness vibration, the secondary (or width vibration) couples a high resistance into the primary (or thickness) vibration and prevents the latter from existing. The doublets that sometimes appear at the thickness resonant frequency represent the two resonant frequencies that exist in coupled circuits tuned to approximately the same frequency, and can usually be avoided by grinding the edge of the crystal to give a different width, or by changing the temperature of operation.

The temperature coefficient of the width vibration of the Y-cut crystal is negative and is of the same order of magnitude as the temperature coefficient of the X-cut. The temperature coefficient of the thickness vibration for Y-cut plates having a very small ratio of thickness to width is positive and normally ranges from 35 to 40 parts in a million per degree centigrade. When the ratio of thickness to width is not small, however, the temperature coefficient can lie almost anywhere between +100 parts in a million per degree centigrade, and −20 parts in a million per degree centigrade, and in special cases can be made zero for a limited range of temperatures. The behavior of the Y-cut crystal at its thickness frequency also often changes radically with slight changes in the crystal temperature. These variations in the temperature coefficient of the thickness frequency of the Y-cut crystal result from the fact that the normal thickness vibration has a positive coefficient, but the width vibration which is coupled to the thickness vibration has a negative coefficient, so that the resultant effect depends to a considerable extent upon the degree of coupling between the two types of vibrations and can vary over wide limits. An important consequence of this rather complicated situation is that, in Y-cut crystals in which the ratio of thickness to width is not too small, the frequency of oscillation will often vary discontinuously with the temperature, there being certain critical temperatures at which the frequency makes a sudden jump.

Crystal Mountings.—The crystal mounting should be so arranged that the vibrations are restrained as little as possible. It is customary to place the crystal between two flat plates, with the upper plate either resting lightly on the crystal or with a slight clearance between upper plate and crystal. The latter arrangement is preferred where the oscillations must have a predetermined frequency, since changing the spacing between the upper plate and the crystal permits a slight adjustment of the frequency. When there is an air gap between the crystal and the upper plate, supersonic air waves will exist in this space as a result of the crystal vibrations, and, if the space is a multiple of a half wave length of the air wave, resonances that absorb large quantities of energy from the crystal will be set up in the air, and they may even damp out the crystal vibrations. It is therefore necessary to use care in choosing the length of air gap. Where extremely high frequency stability is desired the crystal holder is enclosed in a container that is maintained at constant temperature by a thermostat in order to prevent the crystal frequency from changing with the weather.[1]

[1] For details of units for maintaining constant crystal temperature see J. K. Clapp, Temperature Control for Frequency Standards, *Proc. I.R.E.*, vol. 18, p. 2003, December, 1930. See Marrison - IRE July 1928

Frequency Range and Power of Crystal Oscillators.—The frequency range of crystal oscillators runs from about 25 kc as the lower limit to about 4000 kc as the practical upper limit. The width vibration is used for frequencies below about 500 kc, and the lowest frequency that is possible is limited by the size of crystal that can be obtained. The highest frequency that can be generated by crystal oscillators is determined by the minimum thickness to which crystals can be ground. Crystals generating frequencies greater than about 4000 kc are so thin as to be impracticable, and it has been found best to obtain these higher frequencies by generating harmonics of oscillations obtained from lower frequency crystals.

The power that can be obtained from a crystal oscillator is limited at high frequencies by the heating of the crystal, and at low frequencies by the strains which the vibrations set up in the crystal structure and which will crack the crystal if the vibrations are too intense. A crystal having a large area will develop more power than a small crystal with the same resonant frequency, but the extent to which one can go in this direction is limited by the difficulty of obtaining large crystals and grinding their surfaces absolutely parallel. An ordinary crystal will easily drive a tube delivering 5 to 10 watts output, and at the higher frequencies can be used to drive a 50-watt oscillator tube.

Frequency Stability of Crystal Oscillators.—A particular crystal will generate a frequency that does not vary more than several parts in a thousand irrespective of the crystal temperature and the circuit conditions associated with the crystal. When carefully operated under commercial conditions and when used with the same circuit constants, but without temperature control, a given crystal will develop a frequency that is constant to within several parts in 10,000 over long periods of time, while the addition of equipment to maintain the crystal temperature constant will increase the frequency stability that can be expected to several parts in 100,000. Finally, when every possible effort is made to eliminate all causes of frequency variation, it has been found possible to maintain the frequency of an oscillator constant to within a few parts in ten million over periods of days.[1]

56. Magnetostriction Oscillators.—The frequency stability of an ordinary oscillator can be made high by associating a vibrating magnetostriction rod with the electrical circuit and utilizing the magnetostriction effect to obtain the connection between the electrical circuit and the mechanical vibration. Magnetostriction is the distortion of a body

[1] For examples of crystal oscillators with unusually high frequency stability see L. M. Hull and J. K. Clapp, A Convenient Method for Referring Secondary Frequency Standards to a Standard Time Interval, *Proc. I.R.E.*, vol. 17, p. 252, February, 1929; W. A. Marrison, A High Precision Standard of Frequency, *Proc. I.R.E.*, vol. 17, p. 110, July, 1929.

produced by a magnetic field, and is the result of stresses produced within the material by the magnetic flux. There is also an inverse effect, as a result of which a mechanical stress in the magnetostriction material changes the magnetic permeability. Materials showing large magnetostriction effects include monel metal, nickel, stoic metal, invar, stainless steel, and other alloys.

When a rod of magnetostriction material is placed inside a coil through which alternating current is flowing, as illustrated in Fig. 129, the rod will vibrate longitudinally, *i.e.*, will lengthen and shorten, at the frequency of the alternating current provided there is a steady polarizing flux supplied by either residual magnetism, a permanent magnet, or a direct current in the coil.[1] When the frequency of the supply current approxi-

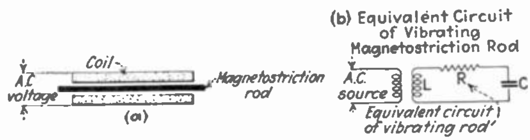


FIG. 129.—Magnetostriction rod arranged to be set into vibration by alternating flux set up by a coil, and equivalent electrical circuit of the combination.

mates the mechanical resonant frequency of the rod to longitudinal vibrations, the intensity of the vibrations becomes very great. As far as the electrical circuits are concerned, such a vibrating rod acts as though it were an electrical circuit tuned to the resonant frequency of the rod and coupled to the coil through which the alternating current passes. Thus the coil with the vibrating magnetostriction rod shown in Fig. 129*a* is electrically equivalent to the circuit of Fig. 129*b* in which the rod has been replaced by the secondary circuit L-C-R.

The most satisfactory type of magnetostriction vibrators are those in which longitudinal vibrations are set up in rods by means of magnetic flux parallel to the axis. The rod is usually pivoted or clamped at its center, which is a nodal point for the vibrations, while the ends are left free. Under these conditions the rod is resonant at a frequency which makes it exactly one-half of a wave length long, so that:

$$\text{Resonant frequency of magnetostriction rod} = \frac{v}{l} \qquad (106)$$

where

v = velocity of sound in the rod
l = length of rod.

The circuit commonly used for magnetostriction oscillators is shown in Fig. 130. While this arrangement resembles a Hartley circuit with

[1] If this polarizing flux is omitted, the rod will vibrate at twice the applied frequency, since positive and negative flux have identical magnetostrictive effects.

a magnetostriction rod coupled to the plate and grid inductances it differs in that the plate and grid coils are preferably connected so that the mutual inductance between them opposes the production of oscillations, with the result that oscillations cease when the magnetostriction rod is removed. The behavior of the circuit of Fig. 130*a* can be analyzed by the equivalent electrical circuit of Fig. 130*b* by considering the vibrating rod to be a resonant circuit inductively coupled to both the plate and grid inductances. The direction of couplings between the rod and the two inductances is such that the indirect coupling between L_g and L_p by way of M_1 and M_2 is of the proper polarity to sustain the self-oscillations which will be generated in the rod when the capacity C_1 tunes the series circuit C_1-L_g-L_p to a frequency that approximates the resonant frequency of the rod. Over a considerable range of adjust-

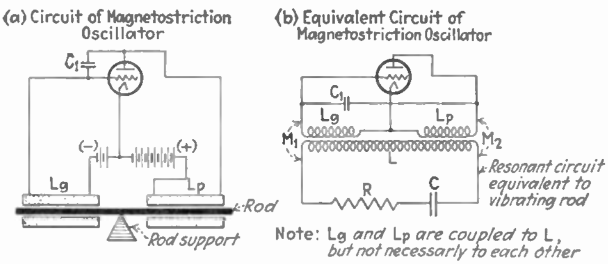


FIG. 130.—Circuit of magnetostriction oscillator, together with equivalent electrical circuit that can be used in analyzing the action taking place.

ments centering about the resonant frequency of the rod the frequency of the oscillations thus generated is determined almost entirely by the vibrating rod and only to a very slight extent by circuit adjustments, even including variations in the tuning condenser C_1.

Oscillator Frequency.—The exact frequency generated by a magnetostriction oscillator can be determined by an analysis based on the equivalent circuit of Fig. 130*b*, and will be found to be very close to the resonant frequency of the rod irrespective of the constants of the associated electrical circuits because the magnetostriction vibrator has an extremely high effective Q, in the neighborhood of 5000 to 10,000. Any slight deviation of the frequency from the resonant frequency of the rod therefore causes large changes in the reactance which the vibrating rod couples into the circuit L_g-L_p-C_1, and this coupled reactance is always of such a sign as to bring the resonant frequency of the circuit L_g-L_p-C_1 nearer to that of the rod. Changes in the electrical circuits, such as variation in the condenser C_1 or in the plate resistance of the tube, etc., which ordinarily tend to alter the frequency of the oscillations, therefore have relatively little effect on the frequency of the magnetostriction oscillator. The extent to which the frequency of a magnetostriction oscillator does vary with changes in circuit conditions can be calculated with the aid of the equivalent electrical circuit of the oscillator shown at Fig. 130b. The frequency stability of the magnetostriction oscillator is much greater than the frequency stability obtained when coils and condensers are used to control the frequency, but is somewhat less than that of the best piezo-electric oscillators.

The frequency range of magnetostriction oscillators is limited at low frequencies by practical considerations relative to the length of the vibrating rod that it is practicable to use, and at high frequencies by the fact that at very high frequencies the magnetic skin effect in the magnetostriction material prevents penetration of the magnetic flux into the material and so reduces the resulting stresses to such an extent that the vibrations are too feeble to control the frequency. The useful range of the magnetostriction oscillator is from about one thousand cycles as a lower practical limit, to several hundred thousand cycles as an upper limit. The magnetostriction oscillator therefore fills in the gap between the highest practical frequency obtainable from tuning forks, and the lowest frequencies that can be obtained by using piezo-electric oscillators.[1]

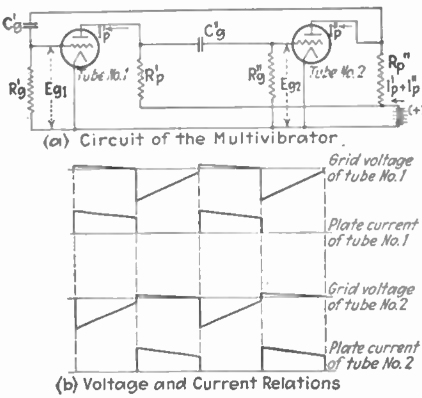


FIG. 131.—Circuit of the multivibrator, together with oscillograms showing the way in which the instantaneous grid potentials and plate currents vary during the cycle of operation.

57. The Multivibrator.—The multivibrator is a two-stage resistance-coupled amplifier in which the voltage developed in the output of the second tube is applied to the input of the first tube, as shown in Fig. 131. Such an arrangement will oscillate because each tube produces a phase shift of 180°, causing the output of the second tube to supply an input voltage to the first tube that is in exactly the right phase to sustain oscillations. Since there is no resonant circuit, the generated frequency is unstable, and the wave form of the oscillating currents is very distorted. It is because of these properties that the multivibrator finds its usefulness

The operation of the multivibrator can be understood by reference to the series of oscillograms shown in Fig. 131. The operation is started

[1] This discussion of magnetostriction oscillators is merely an introduction to the subject. For further information see: George W. Pierce, Magnetostriction Oscillators, *Proc. I.R.E.*, vol. 17, p. 42, January, 1929; and E. H. Lange and J. A. Myers, Static and Motional Impedance of a Magnetostriction Resonator, *Proc. I.R.E.*, vol. 17, p. 1687, October, 1929. Dr. Pierce is the inventor of magnetostriction oscillators.

by an irregularity at the grid of one of the tubes, say tube I. This voltage is amplified by the two tubes and then reapplied to the grid of tube I in a very much enlarged form, to be again amplified and so on. This action takes place almost instantly, and if the initial irregularity places a positive voltage on the grid of tube I the result of the repeated amplification is to cause the grid of tube I to increase its potential suddenly and to make the grid potential of tube II decrease just as suddenly (*i.e.*, become more negative), as is seen in Fig. 131*b*. The rapid change ceases when the plate current of one of the tubes, in this case tube II, is cut off; for then all amplification ceases, and for the moment one tube is drawing a heavy plate current while the other tube takes little or no plate current. This situation cannot endure permanently, however, because the leakage through the grid-leak resistances gradually brings the grid potentials back to normal. As the grid voltage of tube II becomes less negative a point is finally reached where any minute irregularity that increases the plate or grid voltage slightly will make amplification again possible, in which case this irregularity will be amplified and will build up a very sudden positive voltage on the grid of the second tube while causing the grid of the first tube to go so negative that amplification is rendered impossible. This action is clearly evident in the oscillograms and is exactly the same as the initial action except that the relative functions of the two tubes have been interchanged. The potentials that are thus suddenly built up on the two grids gradually die away as a result of the action of the grid leaks, just as before, and finally reach a point at which the cycle repeats itself.

The frequency of the multivibrator oscillation is determined primarily by the grid-leak resistance and grid-condenser capacity, but is influenced somewhat by the remaining circuit elements. The time required to complete one cycle is determined by the rate at which the grid voltages decay as a result of the grid-leak resistances discharging the grid condensers (*i.e.*, by the time constants $R_g'C_g'$ and $R_g''C_g''$) and can be estimated from the formula

Frequency proportional to

$$\text{Approximate frequency of oscillation} = \frac{1}{R_g'C_g' + R_g''C_g''} \qquad (107)$$

The notation follows Fig. 131. The multivibrator can be adjusted to generate frequencies ranging anywhere from 1 cycle every few seconds to about 100,000 cycles per second. The upper limit is the highest frequency at which resistance-coupled amplification is possible, while the lower limit is fixed by the leakage resistance of the grid condenser in relation to the condenser capacity. When very low leakage condensers are available it is possible to employ a leak-condenser combination of such a large time constant as to make the frequency as low as 1 cycle per minute. The exact value of the frequency also varies a certain amount from cycle to cycle since the precise instant at which the reversal of amplification takes place depends upon chance factors which are not necessarily the same from cycle to cycle.

Synchronization with Injected Voltage.—The frequency of a multivibrator oscillator can be controlled by injecting in series with the plate-supply voltage, or in series with the grid-leak resistance, a potential of the frequency to be generated by the multivibrator or of a frequency that is a harmonic of the frequency to be generated by the multivibrator. The mechanism by which this injected potential controls the multi-

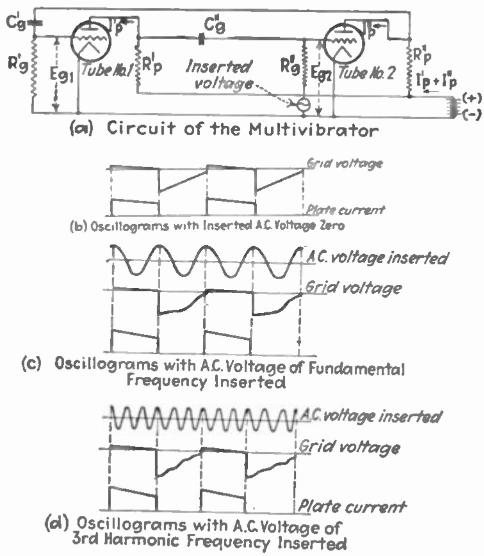


FIG. 132.—Oscillograms of grid voltage similar to those of Fig. 131b, but showing how the multivibrator frequency can be controlled by injecting a voltage in series with the grid leak.

vibrator frequency is apparent from an examination of Fig. 132b, where it is seen that the injected voltage determines the instant at which the grid voltage suddenly changes, with the result that the multivibrator synchronizes with the injected frequency. The frequency of the multivibrator can also be controlled in such a way as to be a subharmonic of the controlling frequency. Thus if the frequency of the uncontrolled oscillation is approximately one-third of that of the injected oscillation the multivibrator frequency will naturally synchronize with the injected voltage in such a way as to have exactly one-third as high a frequency. The mechanism of this control by a frequency that is a harmonic of the stabilized multivibrator oscillation is shown in Fig. 132c, in which it is seen that the instant of reversal of multivibrator currents is determined by the injected oscillation, exactly as in the case of Fig. 132b.

It has been found possible to stabilize a multivibrator oscillation at a frequency as low as one-fiftieth of the injected frequency, but the control is more stable when the frequency difference is less. Ratios as great as 1:10 are readily obtainable, and are extensively used in standard frequency equipment.

It is not necessary that the circuit constants of the two tubes constituting the multivibrator be identical, and unsymmetrical combinations in which the time constants of the two grid circuits differ, or in which the coupling resistances are not the same, can be employed. These unbalanced forms of the multivibrator have characteristics that are in general similar to those of the balanced symmetrical arrangement, but which at the same time differ in numerous respects.[1]

The principal use of the multivibrator is in the measurement of frequency. The multivibrator oscillations have a wave that is rich in harmonics, so by using a standard frequency of known value to control the frequency of the multivibrator it is possible to obtain many frequencies related to the standard. If the fundamental frequency of the multivibrator is controlled by the fundamental frequency or a harmonic of the standard, the multivibrator will then produce higher harmonics of the standard frequency. On the other hand when the frequency of the multivibrator oscillations is controlled in such a way as to be an exact subharmonic of the standard frequency, the multivibrator produces frequencies that are less than the standard. Thus when the multivibrator oscillation has a fundamental frequency that is exactly one-tenth of the standard controlling frequency, the harmonics of the multivibrator are exactly 1/10, 2/10, 3/10, etc., of the standard frequency, and the arrangement is in effect a frequency-reducing system.

"*Motor-boating.*"—Oscillations of the multivibrator type can occur in any audio-frequency amplifier employing a grid leak and grid condenser, provided only that there is some connection between the amplifier output and input by which the amplifier can supply its own input in the proper phase and in sufficient magnitude. Thus multivibrator type of oscillations can be generated in resistance- and impedance-coupled amplifiers as a result of energy fed back through a common plate impedance, when the magnitude and phase of the feed back are such as to make the amplification infinite according to Eq. (82). Such oscillations are frequently called "motor-boating" because their frequency is very low, usually in the order of 1 cycle per second, with the result that the plate current changes give a "put-put" sound when flowing through a telephone receiver or loud-speaker. The changes in plate current are

[1] See L. M. Hull and J. K. Clapp, A Convenient Method for Referring Secondary Frequency Standards to a Standard Time Interval, *Proc. I.R.E.*, vol. 17, p. 252, February, 1929. This paper discusses the controlled multivibrator with both balanced and unbalanced circuit arrangements.

so slow that the presence of motor-boating in ordinary amplifiers can be detected by placing a milliammeter in the plate circuit of one of the tubes and noting whether or not the needle vibrates. Motor-boating in audio-frequency amplifiers is to be avoided under all circumstances and when present can be eliminated either by the use of a source of plate voltage having low internal impedance (as fresh batteries), or by the use of filters as discussed in Sec. 40. Amplifiers receiving their plate voltage from a common dry battery have a tendency to motor-boat when the batteries are near the end of their useful life, particularly when the amplification is high.

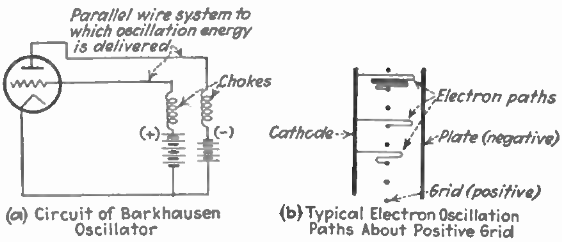


FIG. 133.—Circuit for generating Barkhausen oscillations, together with sketch showing how electron oscillations take place about a positive grid.

58. Barkhausen Oscillations.[1]—The high-frequency limitations of the ordinary type of oscillator are overcome in the method of generating high-frequency oscillations discovered by Barkhausen. In the Barkhausen oscillator the grid is maintained at a positive potential commonly in the neighborhood of 100 to 200 volts, while the plate is kept at a small negative potential, as shown in Fig. 133. With this arrangement there is relatively little space charge in the vicinity of the cathode because of the high positive voltage on the grid, and the electrons are therefore pulled away from the cathode as fast as they are emitted and ultimately reach the grid. In doing so, however, most of the electrons make one or more oscillations about the grid before they finally are drawn to it, as illustrated in Fig. 133. Each individual electron oscillating about the grid generates oscillations that have a frequency determined by the rate at which the electron passes through the grid, and which is an extremely high frequency determined primarily by the grid potential. It might be thought that the phase differences between the oscillations produced by the different electrons would be so completely at random as to cause the average effect of all of the electrons to be practically nil through concellation. This is not the case, however, because the condition of random distribution of motion of the electrons is unstable, and

[1] An excellent treatment of electron oscillations of the Barkhausen type is given by H. E. Hollman, On the Mechanism of Electron Oscillation in a Triode, *Proc. I.R.E.*, vol. 17, p. 229, February, 1929.

any slight deviation from this condition sets into operation forces that tend to accentuate the deviation and thus in the end destroy the random condition.

Barkhausen oscillations can be readily produced in wave lengths ranging from about 30 cm to a few meters, and with special tubes operated under the proper conditions the range can be extended to wave lengths as short as 10 cm, which corresponds to frequencies in the vicinity of 3,000,000,000 cycles per second. The power of the oscillations is quite small, however, the efficiency of such oscillators being enormously less than that of more conventional types operating at lower frequencies.

Gill-Morrell Oscillations.—If the circuits associated with a Barkhausen type of oscillator are approximately in tune with the frequency of the oscillations it is found that the currents in the resonant circuit produce a variation in the potential of the grid which is able to exert a modifying influence on the electron oscillations. The result is that under these conditions the frequency that is generated depends upon the tuned-circuit adjustment as well as upon the grid potential and tube geometry. Oscillations of this type are sometimes called Gill-Morrell oscillations, after their discoverers, but are essentially Barkhausen oscillations generated under special circumstances.

CHAPTER VIII

VACUUM-TUBE DETECTORS

59. Detection of Radio Signals.—Detection is the process of reproducing the transmitted signal from the modulated radio wave. Since all systems of radio communication in practical use transmit intelligence by varying the amplitude of the radiated wave, the detection process, which is also sometimes spoken of as demodulation, must produce currents that vary in accordance with the amplitude of the modulated radio signals. This is done by rectifying the modulated signal to obtain a pulsating direct current varying in magnitude in accordance with the original signal.

Rectification is obtained by applying the modulated radio wave to a circuit in which the current that flows is not proportional to the impressed voltage. During the history of radio communication many types of detectors have been employed, including mechanical, electrolytic, crystal, magnetic (utilizing hysteresis), and others, but all of these have been displaced by the modern vacuum tube, which is vastly superior in both sensitivity and stability.

Rectification can be obtained with vacuum tubes by making use of the non-linear relation that exists between grid voltage and plate current when the plate current is small, or by utilizing the non-linear relation between grid current and grid voltage that exists at grid potentials approximating zero. Detectors of the former sort are variously called anode, plate, or C-bias rectifiers, while those of the latter type are called grid rectifiers. Detectors are also classified into power and weak-signal rectifiers according to whether they are intended to rectify large or small radio-frequency voltages, with amplitudes of about 1 volt being considered as the dividing point between strong and weak signals. Weak-signal rectifiers always develop a rectified current that is proportional to the square of the voltage being rectified and are frequently spoken of as square-law detectors. Most power detectors on the other hand produce a rectified current that is directly proportional to the signal being rectified, and when this condition is approximated the detector is said to be linear.

The detector should reproduce the original signal from the modulated radio wave, and when this is done the detection is said to be distortionless. All practical detectors are more or less imperfect, however, and as a result deliver an output that is not exactly the same as the intelligence that is contained in the modulated wave. This distortion can take several forms. The detector output may include frequencies that were not contained in the original modulation, thus giving rise to amplitude distortion. The detector may also discriminate between modulation frequencies, giving an output that depends upon the modulation frequency and thus introduce frequency distortion. Finally a detector may reproduce the different components in the original modulation in altered phase relations, resulting in phase distortion.

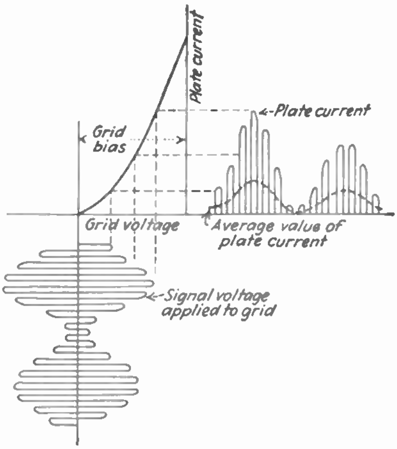


FIG. 134.—Details of action taking place in plate rectifier, showing how a modulated wave applied to the grid adjusted to cut-off will cause plate-current impulses having an average value that varies in accordance with the modulation envelope.

60. Power Detector Using Anode Rectification.[1]—Anode rectification makes use of the non-linear relation that exists between grid voltage and plate current, or what is the same thing, the non-linear relation that exists between plate voltage and plate current. An examination of the characteristic curves of a vacuum tube, such as those of Fig. 134, shows that if the tube is operated with a grid-bias potential approximating the cut-off value, and an alternating voltage is then applied to the grid superimposed upon this grid bias, there will be an impulse of plate current resembling a half sine wave for each positive half-cycle of alternating voltage applied to the grid, while during the negative half-cycles no plate current will flow. This represents a rectification in the plate circuit, as a result of which there is produced a rectified current that varies in accordance with the amplitude of the alternating voltage applied to the grid. The details of this action are shown in Fig. 134 for the case of a wave with sinusoidal modulation, and it is apparent that the average value of the plate current varies in accordance with the envelope of the modulated wave.

Equivalent Circuit of Power Anode Detector.—In order to utilize the rectified current in the plate circuit of the anode detector it is necessary to place a load impedance, such as a telephone receiver, an audio-frequency transformer, or some similar device, in the plate circuit of the detector. When such a load impedance is present the conditions under which the detector operates are different from those existing with no

[1] This treatment of power anode detection is a modification of the method of analysis originated by Stuart Ballantine, Detection at High Signal Voltages—Plate Rectification with the High Vacuum Triode, *Proc. I.R.E.*, vol. 17, p. 1153, July, 1929.

load in the plate circuit because the rectified current produces a voltage drop across the load, and this alters the voltage actually applied to the plate of the detector tube. *When the degree of modulation of the wave*

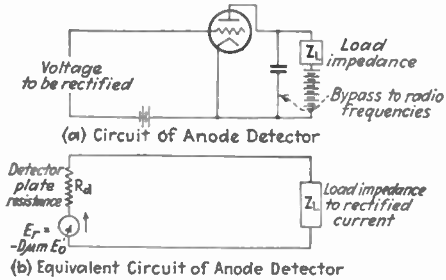


FIG. 135.—Equivalent circuit of plate rectifier, in which the modulation-frequency current produced by rectification is considered as resulting from the action of a generator developing a voltage E_r and having an internal plate resistance R_d.

being rectified by the plate detector is small, i.e., such as 20 per cent or less, the performance of the anode power detector can be calculated by assuming that the rectification process produces a hypothetical modulation-frequency voltage E_r that acts in a circuit consisting of an equivalent detector plate resistance R_d in series with the plate load impedance.[1] This leads to the equivalent circuit of the plate power rectifier shown in Fig. 135, which is analogous in form to the equivalent circuit of the amplifier.

[1] This proposition can be derived as follows: When a carrier voltage E_o modulated to a degree m is applied to the grid of the tube, the amplitude of the alternating input voltage varies sinusoidally about E_o by an amount mE_o. This variation in the signal voltage produces a variation in the rectified plate current of ΔI_p.

If the degree of modulation is not too great, then to a first approximation one can write

$$\Delta I_p = \frac{\delta I_p}{\delta E_s} mE_o + \frac{\delta I_p}{\delta E_p} \Delta E_p$$

where $\delta I_p/\delta E_p$ represents the rate of change of rectified plate current with plate voltage as evaluated in the presence of an unmodulated carrier voltage E_o, ΔE_p represents the change of plate voltage produced by the rectified current ΔI_p flowing through the load impedance Z_L in the plate circuit, and $\delta I_p/\delta E_s$ is the rate of change of rectified plate current with carrier voltage as taken at constant plate voltage and evaluated about a carrier voltage E_o. Substituting $\Delta E_p = -Z_L \Delta I_p$ in the equation gives

$$\Delta I_p = \frac{\delta I_p}{\delta E_s} mE_o - Z_L \Delta I_p \frac{\delta I_p}{\delta E_p}$$

Solving this for ΔI_p results in

$$\Delta I_p = \frac{mE_o\left(\dfrac{\delta I_p/\delta E_s}{\delta I_p/\delta E_p}\right)}{\left(Z_L + \dfrac{1}{\delta I_p/\delta E_p}\right)}$$

This equation then represents a voltage

$$mE_o \frac{\delta I_p/\delta E_s}{\delta I_p/\delta E_p}$$

acting in a circuit having an impedance comprising a load impedance Z_L in series with a resistance $\delta E_p/\delta I_p$.

Equivalent Rectified Voltage.—The equivalent rectified voltage E_r that can be considered as producing the modulation-frequency component of the plate current of the detector when the applied voltage is a carrier wave of crest amplitude E_o modulated to a degree m is given by the expression

$$E_r = D\mu m E_o \tag{108}$$

where

μ = amplification factor of tube

$$D = \frac{[\partial I_p/\partial E_s]_{E=E_o}}{\mu[\partial I_p/\partial E_p]_{E=E_o}} = \text{efficiency of rectification.}$$

The significance of the detection coefficient D that appears in Eq. (108) can be deduced from the fact that in a radio wave having a carrier amplitude E_o and a degree of modulation m, the component of the envelope of the modulated wave that varies at the modulation frequency is mE_o, so that $\mu m E_o$ represents the equivalent modulation-frequency component of the signal envelope that is acting in the plate circuit. *The ratio which the equivalent rectified voltage E_r bears to the modulation-frequency component in the envelope of the wave being rectified, is then D, which is therefore the efficiency of rectification.* The exact value of D depends upon the characteristics of the detector tube, upon the grid and plate voltages, and upon the amplitude of the carrier voltage that is applied to the detector, and will always be less than the value of 1.00 which corresponds to a perfect rectifier.

The efficiency of detection D can be measured by applying to the grid of the detector an alternating voltage (of any convenient frequency) equal to the carrier voltage E_o at which D is desired, and noting the direct-current component of the rectified plate current. This alternating grid voltage is then increased by a small increment, after which the direct-current plate voltage is altered until the direct-current plate current is the same as before the addition of the grid-voltage increment. The ratio of the plate-voltage increment to the alternating (crest value) grid-voltage increment is then μD. The way in which the efficiency of detection D varies with signal voltage in a typical tube is shown in Fig. 136, where it is seen that the efficiency of rectification is low with small carrier amplitudes but increases rapidly as the carrier amplitude is increased up to moderate values, after which it is not particularly critical with respect to the carrier voltage. Increasing the grid bias is also seen to raise the efficiency of detection, but bias values greater than cut-off are to be avoided.

Equation (108) gives the equivalent rectified voltage that can be considered as acting in the plate circuit when a sinusoidally modulated signal is being rectified. If the modulation is complex, that is, if it contains a number of alternating components of different frequencies, the equivalent rectified voltage will have a component for each frequency contained in the modulation envelope, and for each of these components the equivalent rectified voltage will be $D\mu m'E_o$, where m' represents the amplitude of the particular frequency component in the modulation envelope.

Detector Plate Resistance.—The detector plate resistance R_d represents the ratio of direct-current plate-voltage increment to direct current that this plate-voltage increment produces when the unmodulated carrier

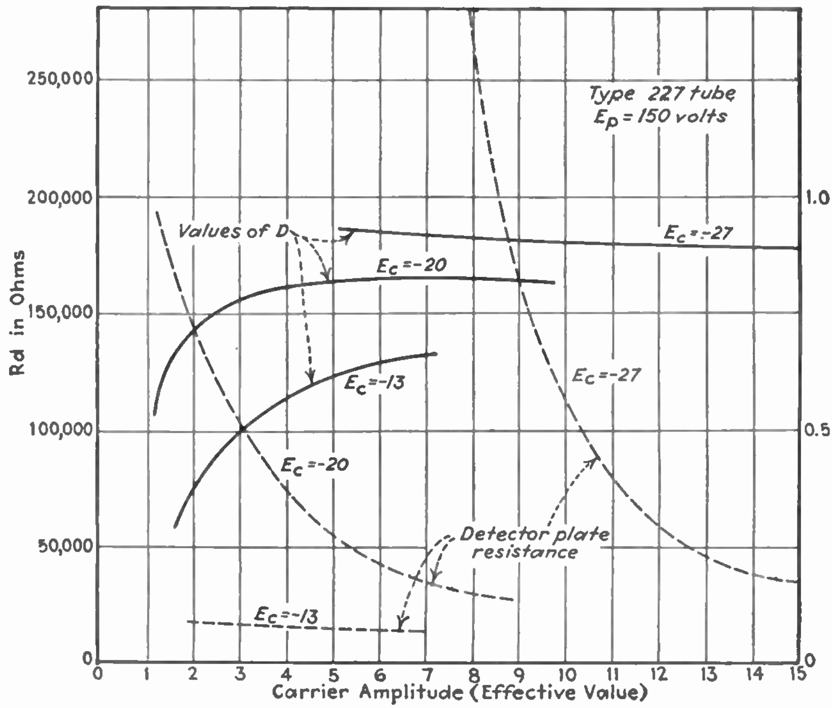


FIG. 136.—Variation of efficiency of detection D and detector plate resistance R_d as a function of carrier amplitude for a typical anode power detector with different grid-bias potentials.

wave E_o is applied to the grid, *i.e.*, $R_d = (\partial E_p/\partial I_p)_{E=E_0}$. The detector plate resistance is therefore identical with the dynamic plate resistance of the tube except for being measured in the presence of the carrier voltage. The value of R_d is determined by the usual technique for evaluating the plate resistance, with the addition of an inaudible voltage of proper amplitude applied to the grid. The detector plate resistance depends very greatly upon the amplitude of the carrier voltage being rectified, having a value that when the signal is small approximates the dynamic plate resistance for the grid bias and plate potentials present, and decreases with increased signal amplitude. The detector plate resistance for a representative tube is shown as a function of carrier voltage in Fig. 136 and illustrates what can be expected in actual cases.

Circuit Design and Calculation of Performance.—The design of the circuits utilizing the rectified output of the detector is based upon the equivalent plate circuit of the anode power detector which is given in Fig. 135, and since this is equivalent in form to the equivalent plate circuit of the amplifier the design problems are exactly the same in both cases. The only feature that is peculiar to the design of the detector circuit is that in detection the load impedance must be shunted with a small by-pass condenser that has a high reactance to the modulation frequency but a low reactance to the high-frequency carrier. Failure to make the load impedance to the carrier frequency low will cause a portion of the equivalent radio-frequency voltage acting in the plate circuit to be consumed in forcing radio-frequency current through the load, and this reduces the signal voltage acting on the non-linear part of the plate characteristic.

The modulation-frequency voltage and power that the rectifier develops as a result of the application of the modulated signal can be calculated from the equivalent plate circuit of the detector exactly as one would in the case of an amplifier, and the frequency distortion, *i.e.*, the variation in output with modulation frequency, can be determined in the same way. The results will depend upon the amplitude of the carrier wave, and in particular it will be found that the frequency distortion will tend to be greater as the amplitude of the carrier voltage is reduced because the detector plate resistance becomes larger as the amplitude decreases. The maximum carrier voltage that can be applied to a plate rectifier is limited by the fact that the grid must never be allowed to become positive, and since 100 per cent modulation must be allowed for, the crest value of the carrier input voltage cannot exceed one-half the grid-bias potential that is being used.

The analysis that has been given of plate detection assumed that the degree of modulation was small, and gives quantitatively correct results only when the degree of modulation is not so great as to carry the amplitude of the signal wave to operating regions where the detector plate resistance R_d and the efficiency of detection D vary during the modulation cycle. This is equivalent to saying that the small-modulation theory holds only so long as the detector is strictly linear, and that the departure from linearity is a measure of the extent to which the small-modulation analysis fails to be correct. With a given degree of modulation the detector will be more nearly linear the larger the carrier voltage, while with a given carrier voltage the departure from linearity becomes greater as the degree of modulation is increased, particularly when the degree of modulation approaches 100 per cent.

Tubes for Anode Power Detection.—The characteristics of tubes suitable for use as anode power detectors are the same as those desired for amplifier tubes. Thus when resistance-coupled amplification is to be used in the plate circuit of the detector the tube should have a high amplification factor compared with the value employed with transformer coupling, while when power output is the chief consideration a power tube should be employed. The grid bias that should be used with anode power rectification is approximately the cut-off value, or slightly less, but the exact value is not highly critical provided it does not exceed cut-off. Since the maximum carrier voltage that can be handled with 100 per cent modulation is one-half the grid bias E_c, the amplitude of signal that can be applied to the grid of an anode power rectifier, and hence the modulation-frequency voltage that can be developed by the output, is proportional to the plate-supply voltage.

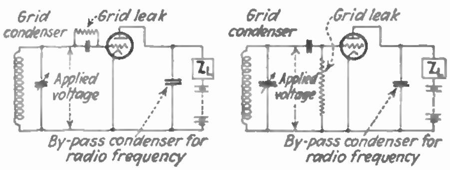


FIG. 137.—Circuits of grid-leak power rectifiers.

61. Power Detection Using Grid Rectification.[1]—In this type of detector the radio-frequency signal voltage is rectified in the grid circuit by taking advantage of the fact that the grid voltage-current characteristic of a tube is non-linear in the region of zero grid potential. Circuits for grid rectification are shown in Fig. 137 and comprise a grid-leak and grid-condenser combination *R-C* in the grid circuit, with the grid return lead brought directly to the cathode without the use of a grid-bias potential. When a modulated wave is applied to the grid of the tube a rectified grid current which varies in amplitude in accordance with the modulation flows in the grid circuit and develops a modulation-frequency voltage across the grid leak and condenser. This voltage is in effect applied to the grid of the detector and so is amplified in the plate circuit by ordinary amplifier action to produce the detector output.

The way the details actually work out can be seen from Fig. 138. At each positive crest of the signal the instantaneous grid potential goes slightly positive, resulting in an impulse of grid current that charges the grid condenser negatively, thereby putting a negative bias on the grid. During the interval between the grid-current impulses some of the accumulated charge on the condenser leaks off through the grid leak, to be

[1] A somewhat more extended discussion of grid power detection is to be found in, F. E. Terman and N. R. Morgan, Some Principles of Grid-leak Power Detection, *Proc. I.R.E.*, vol. 18, p. 2160, December, 1930.

replenished by the next spurt of grid current. The total charge on the grid condenser adjusts itself in relation to the signal amplitude in such a way that the amount by which the instantaneous grid potential goes positive is just sufficient to enable the grid-condenser charge to maintain the proper negative grid potential. As the modulation envelope of the signal changes in amplitude, the amount by which the grid goes positive during each cycle, the size of the resulting grid-current impulse, and the accompanying bias put on the grid by the grid-condenser charge, all vary together, as shown in Fig. 138. The voltage which is applied to the detector grid as a result of the charge on the grid condenser follows the path of the heavy line in Fig. 138*C*. Neglecting the radio-frequency components introduced by the leaking off of charge between cycles, this voltage varies in almost exactly the same way as does the modulation

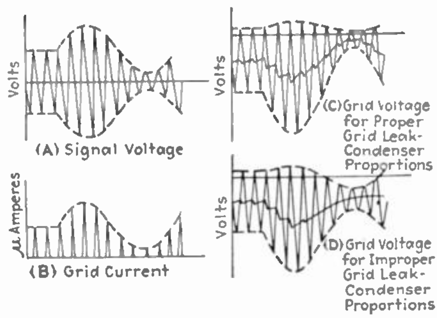


FIG. 138.—Details of action taking place in grid circuit of grid-leak power detector of Fig. 137 when a modulated radio-frequency voltage is applied to the detector.

envelope of the signal, provided the proper proportions of grid-leak resistance and grid condenser are used. The net result of the rectifying action that takes place in the grid-leak power detector is therefore to apply to the grid of the tube a negative voltage that varies in amplitude with the modulation envelope of the signal, and that affects the plate current by amplifier action.

Proportions of Grid Leak and Condenser.—The voltage across the grid condenser of the power detector can decrease only as fast as the condenser charge leaks off through the grid leak. If this leakage is slower than the rate at which the modulation envelope decreases, then the condenser voltage cannot follow the modulation envelope, as is the case in Fig. 138*D*, and the result is frequency and amplitude distortion; but as long as the grid-condenser charge does decrease as fast as the modulation envelope, experiments show that the distortion is small. The rate of decay of the grid-condenser charge is determined by the grid leak-condenser combination, while the rate of decrease of the modulation envelope depends upon the modulation frequency f and the degree of modulation m of the signal. Taking these factors into account, it can be shown that the distortion will be small provided the grid leak-condenser proportions satisfy the relation[1]

$$\frac{X}{R} \geqslant \frac{m}{\sqrt{1 - m^2}} \tag{109}$$

where R is the grid-leak resistance, m the degree of modulation of the signal, and X is the reactance of the *effective* grid-condenser capacity at the modulation frequency involved. The effective grid-condenser capacity is the actual capacity plus the input capacity of the tube, taking into account the plate-load reaction. In order to permit ready visualization of the conditions at which distortion begins to increase rapidly Eq. (109) has been plotted in Fig. 139, where it is apparent that for a given grid leak and condenser there is for every modulation frequency a maximum degree of modulation beyond which distortion begins to be large, but that the permissible value of m increases as the modulation frequency is reduced.

Completeness of Detection.—The ratio of modulation-frequency voltage applied to the grid of the detector by the grid-condenser charge, to the modulation-frequency voltage contained in the signal envelope, is a measure of the completeness of rectification being obtained from the grid-leak detector. This ratio is not greatly different for different types of tubes and is approximately constant for any particular tube type over a wide range of signal amplitudes, degrees of modulation, plate voltages, and grid leak and condenser proportions. When a carrier voltage of E_o modulated to a degree m is applied to a grid-leak power detector

[1] This equation can be derived as follows: The rate of change of voltage across a condenser C charged to a voltage E', caused by leakage through a resistance R, is $-E'/RC$. If the equation of the envelope of the modulated signal wave is

$$e = E(1 + m \cos 2\pi ft)$$

then the rate of change of envelope is

$$\frac{de}{dt} = -mE2\pi f \sin 2\pi ft$$

and at the time $t = t_o$ the envelope magnitude e_o is

$$e_o = E(1 + m \cos 2\pi ft_o)$$

The grid condenser rate of discharge can follow the rate of change of modulation envelope at any time t_o of the modulation cycle provided $-E'/RC \geqslant de/dt$, where de/dt is evaluated at $t = t_o$ and $E' = e_o$, or when $-mE2\pi f \sin 2\pi ft_o \leqslant -E(1 + m \cos 2\pi ft_o)/RC$. This can be reduced to

$$\frac{X}{R} = \frac{1}{2\pi fCR} \geqslant \frac{m \sin 2\pi ft_o}{(1 + m \cos 2\pi ft_o)}$$

The point on the modulation cycle where it is most difficult for the grid-condenser charge to keep up with the envelope change is at a time t_o that makes the right-hand member of this last equation maximum, which is when $\cos 2\pi ft_o = -m$. The maximum value of the right-hand member of the equation is therefore $m/\sqrt{1 - m^2}$, and the charge on the grid condenser can decrease at least as fast as the modulation envelope changes when $X/R \geqslant m/\sqrt{1 - m^2}$.

having a completeness of rectification β, the modulation-frequency voltage that the detection process causes to be applied to the detector grid is $\beta m E_o$, which is then amplified in the usual manner. Values of β as a function of signal amplitude are given in Fig. 140 for a number of representative tubes. The completeness of rectification depends upon the sharpness with which the grid current-voltage characteristic rises. This depends primarily upon the voltage drop in the cathode of the tube, since with an equipotential cathode the grid goes positive with respect to all parts of the grid at the same time, which is not the case if there is a considerable voltage drop in the filament.

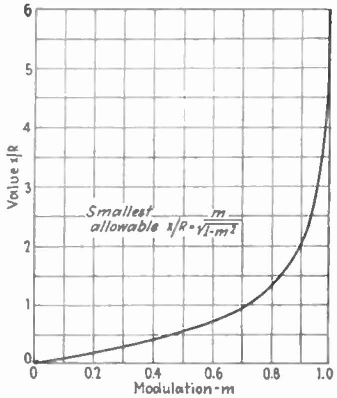


FIG. 139.—Values of the smallest allowable ratio X/R that will prevent frequency distortion, plotted as a function of degree of modulation.

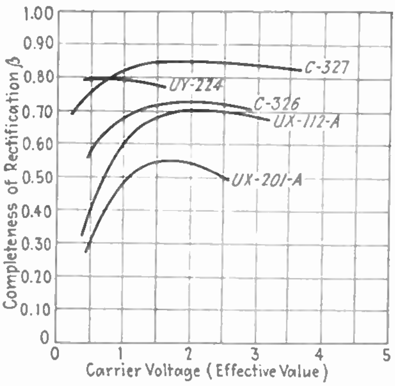


FIG. 140.—Variation of completeness of grid rectification as a function of carrier voltage for a number of representative tubes.

Design of Grid-leak Power Detectors.—It is apparent from Eq. (109) that the frequency distortion in a grid-leak power detector will be least when the grid-leak resistance and the grid-condenser capacity are both small, but there is a limit to the extent to which one may go in this direction. Since the modulated wave is applied to the grid of the tube through the grid-condenser capacity it is necessary that this capacity be at least several times the effective grid-cathode capacity of the tube to the high-frequency signal if loss of signal by voltage drop in the grid condenser is to be avoided. With ordinary tubes the effective grid-cathode capacity is between 10 and 20 $\mu\mu$f, which makes the minimum allowable grid-condenser capacity something between 50 and 100 $\mu\mu$f. The minimum value of grid-leak resistance that can be used is limited by the fact that the effective input resistance of the detector grid circuit to the modulated wave is proportional to the grid-leak resistance and approximates $R_g/2\beta$.[1] It is therefore necessary to compromise between

[1] The grid current in the grid-leak detector flows only when the signal voltage is at or near its crest value, as is clearly shown at Fig. 138*B*. The power absorbed by

low energy consumed by the detector from the signal, *i.e.*, a high input resistance, and a small amount of frequency distortion. Under practical circumstances this ordinarily calls for a grid leak approximating ¼ to ½ megohm, in which case the equivalent input resistance to the modulated wave is between 150,000 and 300,000 ohms. With such a grid-leak resistance, and a grid-condenser capacity of 100 $\mu\mu$f or less, it is possible to rectify high-quality radio-telephone signals without introducing appreciable distortion and at the same time keep the losses in the detector grid circuit reasonably low.

The load impedance that is placed in the plate circuit to make use of the rectified output of the detector is designed exactly as in the case of an amplifier, since the detector action is to produce a modulation-frequency voltage that is applied to the detector grid and amplified in the ordinary manner. The only special features that must be considered in the case of detection are first that the effective plate resistance depends upon the amplitude of the carrier voltage and is higher the larger the carrier voltage because of the greater average negative voltage that is applied to the grid when a large carrier is present, and second that the impedance offered by the plate circuit to the carrier frequency should be low in order that the input impedance of the tube will be as high as possible (see Sec. 43).

Power Capacity of Grid Power Rectifier.—The maximum signal voltage that may be applied to a grid-leak power detector adjusted to satisfy Eq. (109) is the input at which plate rectification begins to be appreciable. Such plate detection produces distortion frequencies and also prevents the undistorted portion of the output from being proportional to the signal amplitude and the degree of modulation. Plate rectification takes place only at such high signal amplitudes that the grid bias produced by the grid-condenser charge is sufficiently great to cause the operating region to extend into the curved part of the grid voltage plate current characteristic. Stated in another way, this means that if amplitude distortion is to be avoided in the rectified output, the modulated wave must be amplified in the plate circuit without distortion. The power output available from a grid-leak power detector therefore varies with the plate-supply voltage in almost the same way as does the power rating of an amplifier, and it is just as impossible to handle large inputs in a

---

the detector input is accordingly slightly less than the product of the crest signal voltage and the average grid current. Since the average grid current is equal to $\beta E/R$, where β is the completeness of rectification, E the crest value of signal voltage, and R the grid-leak resistance, one can write

$$\text{Grid power loss} = \frac{\beta E^2}{R} = \frac{(\text{effective signal})^2}{(R/2\beta)}$$

The denominator of this last term represents the equivalent input resistance to the signal, which is accordingly $R/2\beta$.

grid-leak power detector without high plate voltages as it is in an amplifier.

It is convenient to express the maximum carrier voltage that may be applied to a grid-leak power detector in terms of the voltage that may be applied to the same tube when acting as a power amplifier having the same plate voltage. If the plate load impedance to the radio-frequency carrier is low, and 100 per cent modulation is allowed for then[1]

$$\frac{\text{Allowable carrier voltage}}{\text{Allowable amplifier input}} = \frac{k}{1 + \beta} \tag{110}$$

where β is the completeness of rectification and usually lies between 0.7 and 0.9, while k represents the ratio of proper grid bias for amplifier operation with no load impedance, to proper grid bias for amplifier operation with normal load impedance, as calculated from Eq. (74). If the allowed amplifier input is taken as the value recommended by the tube manufacturers for amplifier operation, k may be expected to be very close to 0.70.

With ordinary tubes the carrier voltage that can be handled by the grid-leak power detector without introducing amplitude distortion is at

[1] When a signal of carrier amplitude E_o modulated to a degree m is rectified with a completeness β, the maximum potential across the grid condenser is $\beta(1 + m)E_o$ and takes place when the signal is at the crest of the modulation cycle, when the amplitude reaches $(1 + m)E_o$. The most negative instantaneous potential on the grid is the maximum voltage across the grid condenser plus the signal voltage, which sum, if distortion is to be avoided, must not exceed twice the proper grid bias for amplifier operation with zero plate load. Since the proper grid bias depends upon plate load impedance, being less for lower impedances, the bias E_c recommended for amplifier operation at the plate potential employed in the detector must be multiplied by a factor k, having a value approximately 0.70, to give the bias for amplifier operation with no plate load. The maximum value of E_o that will not produce distortion must then satisfy the relation

$$2kE_c = (1 + \beta)(1 + m)E_o$$

Remembering that E_c is the allowable input for normal amplifier operation, this equation can be rewritten as

$$\frac{\text{Allowable carrier } E_o}{\text{Allowable amplifier input}} = \frac{2k}{(1 + \beta)(1 + m)} \tag{111}$$

This reduces to Eq. (110) when $m = 1.0$.

When the plate circuit of the grid-leak detector contains an impedance to the modulation frequency, the carrier voltage that may be applied to the detector is somewhat greater than the value given by Eq. (110). This is because the effective plate voltage, so far as the radio-frequency signal currents are concerned, is higher during the crest of the modulation cycle than at other times as a result of the voltage developed across the load impedance by the modulation-frequency currents in the plate circuit. Such a load impedance may increase the allowable detector input by as much as one-third if the completeness of rectification is great and the load has a very high impedance, but will ordinarily have somewhat less effect, so that Eq. (110) is satisfactory for ordinary calculations, particularly as it errs on the conservative side.

least one-third and not more than one-half of the maximum input that the same tube will handle as an amplifier at the same plate-supply voltage. This is evident when reasonable values of β and k are substituted in Eq. (110), and estimates are made to correct for the effect of plate load impedance. The ratio of maximum undistorted modulation-frequency voltage that can be delivered by the output of a grid-leak power detector, to the output that can be obtained from the same tube operated as an amplifier, is β times the ratio of inputs obtained from Eq. (110) after this equation has been corrected for the plate load impedance. With ordinary tubes this ratio of voltage outputs ranges from 0.30 to 0.40 with sufficient certainty to enable these figures to be used for tentative design purposes.

Miscellaneous Aspects of Grid Power Rectifiers.—Since the completeness of rectification β does not differ greatly with different types of tubes, the choice of a tube for grid-leak power detection is determined primarily by the amplification characteristic, so that the tubes best suited for grid-leak power detection are also those preferred for amplification purposes. In selecting the operating conditions for the tube it is to be kept in mind that the maximum signal voltage which the tube can handle is proportional to the plate-supply voltage, but this plate voltage cannot be increased without limit because when no carrier voltage is present the tube operates at substantially zero grid bias, and if the plate voltage is very high the plate current will be excessive. As a result the plate voltage of a grid-leak power detector is limited to values in the order of one-half to two-thirds of the maximum plate voltage that can be used when the tube is operating as an amplifier.

In the case of filament-type cathodes the grid return lead can be brought to either the positive or negative side of the filament, as with large signals the performance is substantially the same in either case.

The frequency distortion of a grid-leak power detector will be small provided Eq. (109) is satisfied, while the amplitude distortion will be low if the input voltage does not exceed that given by Eq. (110). The grid-leak power detector is usually somewhat more linear (*i.e.*, its rectified output is more nearly proportional to the amplitude of the signal envelope) than is the plate power detector, and like the power anode rectifier, with constant carrier voltage the linearity becomes less as the degree of modulation is increased, particularly as 100 per cent modulation is approached, while with constant degree of modulation the linearity becomes greater as the carrier voltage is increased.

It would be possible to analyze the action that takes place in the power grid rectifier by means of an equivalent grid circuit analogous to the equivalent circuit used in connection with the power anode rectifier. This method of analysis is, however, not very satisfactory when applied to grid-leak power detection because the high resistance which the grid leak offers to the direct-current component of the rectified grid current places a negative potential on the grid of the tube that very nearly cuts off the grid current, with the result that the equivalent circuit of the grid power rectifier possesses an internal resistance that is very sensitive to the degree of modulation. The equivalent-circuit method of analyzing

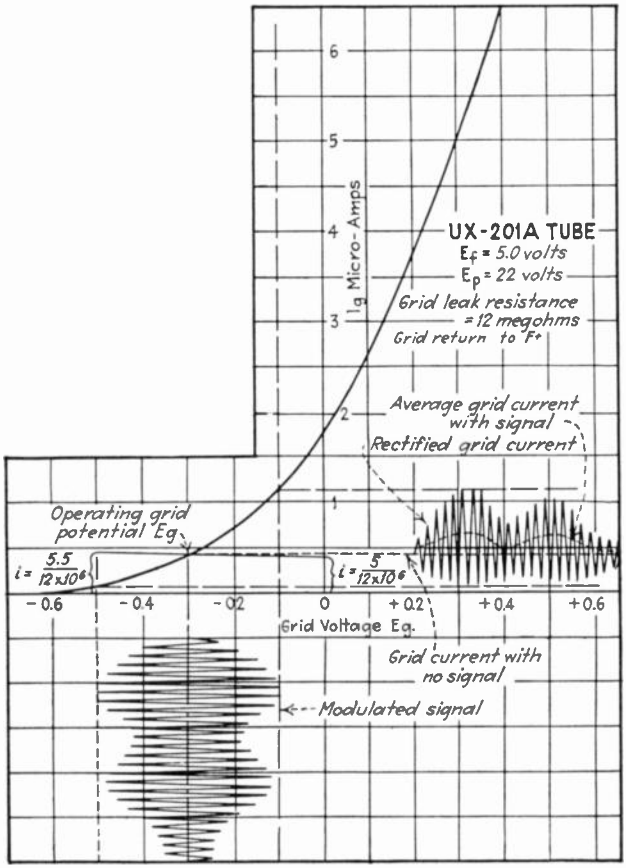


FIG. 141.—Typical grid current-voltage characteristic, together with waves showing how such a characteristic is capable of rectifying a small alternating voltage. The flow of grid current at a slightly negative grid voltage is made possible by the emission velocity of the electrons.

power rectification in the grid circuit has its usefulness limited to very small degrees of modulation and is not nearly so satisfactory with large signals as is the method of analysis that has been given. These objections do not apply to the anode power rectifier, where the direct-current resistance of the plate load impedance is relatively small.

62. Grid Rectification of Small Signals.[1]—Grid rectification of small signals makes use of the same circuit as grid rectification of large signals,

[1] The following is a list of references selected from the extensive literature on the detection of weak signals: Stuart Ballantine, Detection by Grid Rectification with

but the action taking place differs in many respects because when the signal being rectified is small, *i.e.*, 0.25 volt or less, the curvature of the grid voltage-current curve is such that the grid current is only partially suppressed during the negative half-cycles of the signal, with the result that the grid rectification is only partially complete. The mechanism of grid rectification under these conditions of small signals can be understood by reference to Fig. 141, where the adjustments are such that with no signal present the grid is at an operating potential marked as E_g. An alternating voltage applied to the grid of the detector is superimposed upon the operating grid potential E_g and causes the instantaneous grid potential to swing alternately more and less positive than E_g, as shown in Fig. 141. This alternating grid voltage produces corresponding variations in the grid current, but as a result of the curvature of the grid voltage-current characteristic the grid current increases more during the positive half-cycles than the current decreases during the negative half-cycles. The net result is consequently a rectified grid current produced by the alternating signal voltage and varying in magnitude with the amplitude of the signal, as shown in Fig. 141.

When a grid leak and grid condenser are inserted in the grid circuit the rectified grid current produced by the radio-frequency signal voltage must flow through the leak-condenser combination, and in doing so there will be developed across this impedance a voltage drop which is in effect applied to the grid of the detector tube and so is amplified in the plate circuit. Since the magnitude of the rectified current depends upon the amplitude of the voltage being rectified, the voltage drop which this rectified current produces across the grid leak-condenser combination varies in accordance with the envelope of the signal voltage and therefore reproduces the original signal which was modulated upon the radio wave. The details of the action taking place in grid-leak detection of weak signals can be understood by an examination of Fig. 142, which shows the way in which the instantaneous grid current, grid voltage, and voltage drop across the grid leak-condenser combination vary when a modulated wave is applied to the detector. The rectified grid current and the voltage drop across the leak vary with the modulation envelope

---

the High-vacuum Triode, *Proc. I.R.E.*, vol. 16, p. 593, May, 1928; John R. Carson, The Equivalent Circuit of the Vacuum-tube Modulator, *Proc. I.R.E.*, vol. 9, p. 243, June, 1921; E. L. Chaffee and G. H. Browning, A Theoretical and Experimental Investigation of Detection for Small Signals, *Proc. I.R.E.*, vol. 15, p. 113, February, 1927; F. M. Colebrook, The Rectification of Small Radio-frequency Potential Differences by Means of Triode Valves, *Exp. Wireless and Wireless Eng.*, vol. 2, p. 946, Dec., 1925; F. E. Terman, Some Principles of Grid Leak-Grid Condenser Detection, *Proc. I.R.E.*, vol. 16, p. 1384, October, 1928; F. E. Terman and T. M. Googin, Detection Characteristics of Three-element Vacuum Tubes, *Proc. I.R.E.*, vol. 17, p. 149, January, 1929.

and so contain alternating components that represent the intelligence carried by the modulated wave.

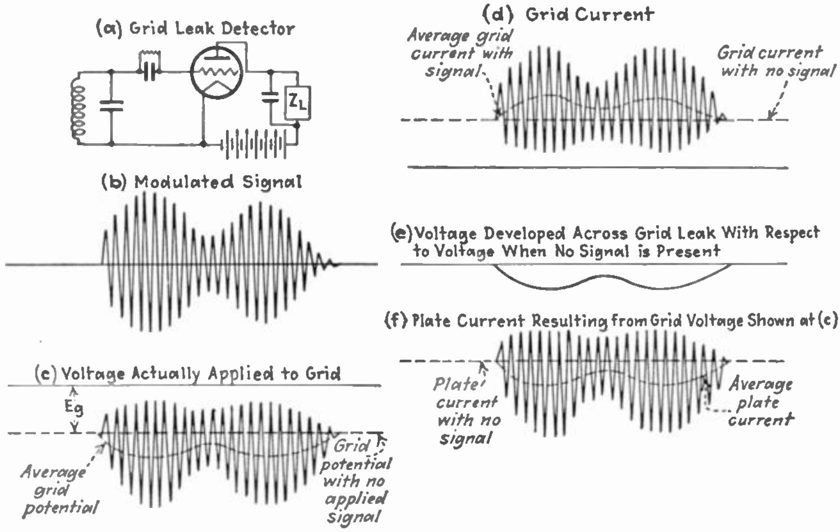


FIG. 142.—Details of voltage and current relations existing in a weak-signal grid-leak detector, showing how the rectified grid current develops a voltage drop across the grid leak that causes the average value of plate current to vary in accordance with the modulation envelope of the signal voltage.

Equivalent Circuit of Weak-signal Grid Rectifier.—The voltage drop across the grid leak-condenser combination that is produced by a given radio-frequency signal can be determined with the aid of the equivalent circuit of the grid leak-condenser detector shown in Fig. 143. This equivalent circuit is based on the following proposition: *The rectified grid current produced by application of a small signal voltage, such as 0.05 volt effective or less, to the grid of a detector tube is exactly the same current that would be produced by a series of suitable generators acting between the grid and cathode in series with the grid-cathode resistance of the tube. There is one such generator for each frequency component of the rectified current. The proper amplitude and phase of each generator depend only upon the operating grid potential, and the nature of the signal, and are given in Table IX. The current that would be produced by these fictitious generators if they were actually present in the grid circuit is the rectified current that the applied radio-frequency signal voltage produces.*[1]

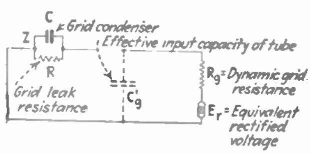


FIG. 143.—Equivalent grid circuit of weak-signal grid rectifier, in which the rectified grid current is considered as being produced by an equivalent generator having a voltage E_r and an internal resistance of R_g.

[1] This proposition is obtained by considering that a small section of the grid voltage-current curve can be represented by an equation of the form $I_g = a + be_g +$

The equivalent rectified voltage E_r that can be considered as producing the rectified grid current ordinarily contains a number of components of different frequencies. In addition to the frequency components contained in the modulation envelope, there are also second harmonic and sum and difference distortion frequencies present. Thus in the case of a wave with simple sinusoidal modulation there is a direct-current and a modulation-frequency component representing the modulation envelope, and a distortion component having twice the modulation frequency. The components present in any particular case can be determined with the aid of Table IX.

TABLE IX.—FORMULAS FOR RECTIFIED GRID VOLTAGE E_r

1. Unmodulated radio-signal voltage of crest value E_o, having equation $e_s = E_o \sin wt$.
   Components of rectified grid voltage:

   $$\text{Direct-current component} = +E_o^2/2v$$

2. Two superimposed alternating-current voltages of amplitudes E_1 and E_2, and frequencies f_1 and f_2, respectively, having equation $e_s = E_1 \sin 2\pi f_1 t + E_2 \sin (2\pi f_2 t + \phi)$.
   Components of rectified grid voltage:

   $$\text{Direct-current component} = +(E_1^2 + E_2^2)/2v$$
   $$\text{Component of frequency } f_1 - f_2 = +[(E_1E_2)/v] \cos [2\pi(f_1 - f_2)t - \phi]$$

3. Modulated wave with carrier crest amplitude E_o and degree of modulation m, having equation $e_s = E_o(1 + m \sin qt) \sin wt$.
   Components of rectified grid voltage:

   $$\text{Direct-current component} = +(E_o^2 + m^2E_o^2/2)/2v$$
   $$\text{Modulation-frequency component} = +(E_o^2m/v) \sin qt$$
   $$\text{Double modulation frequency component} = -(m^2E_o^2/4v) \cos 2qt$$

4. General formula: Write down e_s as a function of time, square the function, and divide by v. The low-frequency components of the result are the components of E_r effective in the equivalent detector grid circuit.

Notation and formulas:

$$R_g = dE_g/dI_g = \text{dynamic grid resistance.}$$
$$v = 2R_g/(dR_g/dE_g) = \text{detection constant.}$$

Note: All amplitudes given in this table are crest values.

The magnitude and character of the rectified grid voltage E_r are determined only by the signal being rectified, and a tube constant v which takes into account the characteristics of the tube. The size of grid condenser and grid leak has no effect upon the rectified voltage E_r except in so far as the grid-leak resistance affects the operating grid

---

ce_g^2. This is equivalent to assuming that the $e_g - i_g$ curve over a limited range of values is a section of a parabola. A proof of the proposition is given by John R. Carson, The Equivalent Circuit of the Vacuum-tube Modulator, *Proc. I.R.E.*, vol. 9, p. 243, June, 1921.

potential. The equivalent grid circuit of the weak-signal grid rectifier can therefore be treated as an ordinary electrical circuit in which a voltage E_r is acting. This rectified voltage E_r usually contains a number of components of different frequencies, and each one of these components acts independently to produce its own component of current according to the impedance which the equivalent grid circuit offers to it. If E_r'

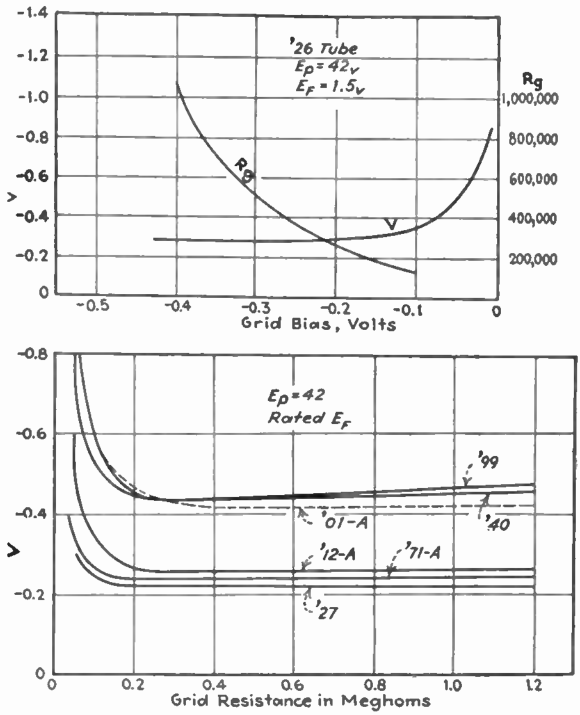


FIG. 144.—Relationship between grid bias, detection constant v, and grid resistance R_g in a typical vacuum tube, and also v as a function of R_g in a number of typical tubes. Note that all tubes with the same type of cathode have similar characteristics, and that the oxide-coated cathode gives a distinctly lower v than does the thoriated-tungsten emitter.

is a particular component of the rectified grid voltage, and Z' is the impedance offered to this frequency by the grid leak-condenser combination, reference to Fig. 143 shows that

$$\text{Voltage produced across grid leak by } E_r' = E_r' \frac{Z'}{Z' + R_g} \qquad (112)$$

This voltage, which is applied to the grid of the tube and then amplified in the plate circuit, will be greatest when the grid leak-condenser impedance Z' is large compared with R_g, but in no case can it exceed E_r'.

Detection Constants R_g and v.—The grid-cathode resistance R_g that appears in the equivalent detector circuit is the resistance offered by the grid circuit of the tube to a small increment in grid voltage and is analogous to the dynamic plate resistance, so that

$$R_g = \frac{dE_g}{dI_g} \tag{113}$$

The value of R_g depends upon the grid and plate voltages, the cathode temperature, and tube construction, and, with a particular tube operated at a given plate voltage and cathode temperature, will increase as the grid potential becomes more negative, as shown in Fig. 144. The grid resistance R_g can be evaluated either by taking the slope of the $E_g - I_g$ curve, or by direct measurement of the dynamic resistance by an alternating-current bridge, as explained in detail in Sec. 141.

The grid-detection constant v is the factor that takes into account the effect which the tube characteristics have upon the magnitude of the rectified voltage E_r. Its numerical value depends upon the grid resistance R_g and upon the rate of change of grid resistance with grid voltage, according to the formula

$$\text{Detection constant} = v = \frac{2R_g}{dR_g/dE_g} \tag{114}$$

The detection constant can be most easily evaluated by determining the grid resistance at the operating point, and then at grid potentials greater and less than that at the operating point by a small increment ΔE_g. If the grid resistances under these conditions are R_g, R_g' and R_g'', respectively, then $dR_g/dE_g = (R_g' - R_g'')/2\Delta E_g$, and $v = 4\,R_g\Delta E_g/(R_g' - R_g'')$.

The detection constant v depends primarily upon the cathode temperature, and the dynamic grid resistance R_g at the operating point, and only to a secondary extent upon the tube construction and plate and grid voltages which happen to give this grid resistance. The way in which the size of v varies with grid resistance in typical tubes is shown in Fig. 144, where it is seen that as the grid resistance is increased from low to high values, v is initially large but decreases rapidly to a low and constant value that is approximately proportional to the absolute cathode temperature and so is roughly the same for different tubes with the same type of cathode.[1] The values over this flat minimum region of v

[1] When the grid current is very small, as it is on the low flat part of the $v - R_g$ characteristic, the grid current is determined almost solely by the distribution of the emission velocities of the electrons, which depends upon the cathode temperature rather than the tube construction, plate voltage, or voltage drop in the cathode. This is a Maxwellian distribution and causes the grid current to vary with grid voltage according to an equation of the type $i_g = I_o\epsilon^{k(E_g - E_o)}$, where k is a constant that is proportional to absolute cathode temperature and is related to v according to the equation $v = -2/k$, and I_o and E_o are constants determined by tube construction, plate voltage, etc. Where this exponential relation between i_g and E_g holds, the detector constant v is constant irrespective of grid voltage (or grid resistance) at a value that is proportional to the cathode temperature.

Plot I_g curve as easy way to get v

are always very close to -0.25, -0.45 and -0.55 volt, for oxide-coated, thoriated-tungsten, and tungsten cathodes, respectively, operated at their normal temperatures, irrespective of the plate voltage, voltage drop in the cathode, or tube construction.[1] The detection constant v has the dimension of a voltage and is negative in ordinary tubes. Since the magnitude of the rectified grid voltage E_r is inversely proportional to the numerical value of v, the flat minimum part of the $v - R_g$ characteristics represents the desirable operating region, and the efficiency of rectification (*i.e.*, the sensitiveness of the detector) that can be obtained is much greater with oxide-coated cathodes than it is with those of the thoriated-tungsten type.

The Operating Point.—Satisfactory operation of weak-signal grid rectifiers depends primarily upon operating at the proper point on the grid voltage-current characteristic. If the grid resistance R_g at the operating point is too low, the detector constant v is large, and the efficiency of rectification is therefore low. At the same time the energy consumed by the detector input will be large because the dynamic grid resistance at the operating point represents the effective input resistance of the detector. On the other hand if the grid resistance R_g is very high at the operating point the detector constant v will be small, giving a high sensitivity, *i.e.*, a large E_r, and the input resistance will be high, but the higher modulation frequencies in the signal envelope will not be reproduced as well as the lower frequencies, thus resulting in frequency distortion.

In the proper operating region for weak-signal grid rectifiers, the grid resistance R_g depends almost solely upon the grid current at the operating point and upon the type of cathode (*i.e.*, whether tungsten, thoriated tungsten, or oxide coated) and is relatively independent of the tube construction or the plate voltage and grid potential giving rise to this grid current. The amount of grid current flowing at the operating point is determined by the grid-leak resistance and the potential of the grid return lead. This is because the grid potential at the operating point is more negative than the potential of the grid return lead by the voltage drop produced in the grid leak by the grid current. The grid current therefore decreases as the grid-leak resistance is made greater and as the potential of the grid return lead is made more negative. When the $E_g - I_g$ characteristic is known, the operating point obtained with any particular grid-leak resistance R_{gl} can be determined by the following construction: Draw a straight line that intersects the I_g-axis at $I_g = E_c/R_{gl}$, where E_c is the potential of the grid return lead, and which goes through the point corresponding to a grid voltage E_g' and a grid current $(E_c - E_g')/R_{gl}$, where E_g' is any convenient grid potential. The intersection of this line with the $E_g - I_g$ curve gives the grid operating point

[1] See Terman and Googin, *loc. cit.*

for the resistance and grid-return potential used. In Fig. 141 this construction has been carried out for a 12-megohm leak when the grid return lead is brought to the positive side of the filament ($E_c = +5.0$). The usual procedure in weak-signal grid rectifiers is to bring the return grid lead to the positive side of the filament in filament-type tubes, or directly to the cathode in heater-type tubes, and then to employ whatever grid-leak resistance is required to give the desired grid resistance at the operating point.[1]

Selection of Grid-condenser Capacity and Leak Resistance.—The size of the grid condenser is the result of a compromise between two conflicting requirements. The capacity should be large in order that it will consume a negligible part of the radio-frequency signal that is applied through the grid condenser to the grid of the tube, and should be small in order to offer a high impedance to the rectified grid currents. The fraction of the signal voltage applied to the detector that is actually impressed upon the grid of the tube is $C/(C + C_g)$, where C is the capacity of the grid condenser, and C_g is the effective input capacity of the tube to radio frequencies plus the capacity of the lead running from the grid condenser to the grid. It is important that the fraction $C/(C + C_g)$ be large since the sensitivity of a detector, *i.e.*, the rectified voltage E_r, is proportional to the square of the signal impressed upon the grid of the tube. In order to avoid appreciable loss of sensitivity it is necessary that the grid condenser have a capacity that is at least five times the tube capacity C_g, and since the value of C_g will usually be not less than 15 $\mu\mu$f, the grid-condenser capacity should be at least 75 to 100 $\mu\mu$f and preferably twice this value.

With the grid-condenser capacity determined by these considerations, the grid resistance at the operating point is then selected by keeping in mind that: (1) the grid resistance must be high enough to lie on the low flat part of the $v - R_g$ characteristic if the full sensitivity of the detector is to be realized; (2) the grid resistance constitutes the input resistance of

[1] In the region where the $v - R_g$ curve has the flat minimum, and which is always the proper operating region, the dynamic grid resistance can be determined by measuring the grid current with a microammeter. This is because under these conditions the current and voltage are related by an equation of the type

$$i_g = I_0 \epsilon^{k(E_g - E_0)} \tag{115}$$

where I_0 and E_0 are constants depending upon the tube construction. The grid resistance $R_g = dE_g/dI_g$ at a grid current i_g is then given by the equation

$$R_g = \frac{1}{ki_g} = -\frac{v}{2i_g} \tag{116}$$

By means of Eq. (115) the approximate grid current required to give a desired grid resistance can be calculated for any type of cathode, and the grid-leak resistance can then be changed until a microammeter in the grid circuit shows that the desired current is present. In this way any desired operating conditions can be readily approximated using only a microammeter.

the detector to the signal and so the higher the input resistance the lower will be the energy consumption of the detector; and (3) the higher the grid resistance the greater will be the frequency distortion of the high-frequency components of the rectified current. These requirements set the minimum feasible value of R_g at about 150,000 ohms, which is also nearly the highest value that can be tolerated if frequency distortion is to be avoided in the detection of high-quality radio-telephone signals.

Calculation of Frequency Distortion.—Frequency distortion appears in the weak-signal grid rectifier as the result of the fact that the grid condenser offers a lower impedance to rectified currents of high frequency than to rectified currents of low frequency. This causes the rectifier to have its maximum sensitivity at the lower modulation frequencies, and to discriminate against the higher modulation frequencies. At low frequencies the reactance of the grid condenser is so high as to be substantially infinite, and the potential which is developed across the grid-leak resistance by the rectified voltage is

$$\text{Voltage across leak at low frequencies} = E_r \frac{R_{gl}}{R_{gl} + R_g} \tag{117}$$

where R_{gl} is the grid-leak resistance, and R_g is the grid resistance at the operating point. Since the leak resistance is always much higher than R_g, the voltage that is developed across the grid leak, and which is also the voltage applied to the detector grid, is only very little less than the equivalent rectified voltage E_r for low modulation frequencies. As the frequency of the rectified current increases, the fraction of E_r that appears across the leak resistance becomes less because of the increased shunting effect which the grid condenser has as the frequency is raised. The extent to which the response falls off as the frequency of the rectified current increases depends upon the ratio which the resistance formed by the grid leak R_{gl} and dynamic grid resistance R_g in parallel, bears to the reactance of the effective grid-condenser capacity, and becomes greater as this ratio is increased. The exact analysis can be carried out by the use of the equivalent circuit of the detector which is shown in Fig. 143 and leads to the following result:

$$\frac{\text{Actual response}}{\text{Response at low frequencies}} = \frac{1}{\sqrt{1 + (R_{eq}/X)^2}} \tag{118}$$

where R_{eq} is the resistance of the combination formed by the grid-leak and dynamic grid resistance in parallel, and X is the reactance $1/\omega C$ of the effective grid-condenser capacity. Equation (118) is exactly the same as Eq. (64) since the two are based on identical circuits. The extent of the falling off can therefore be determined directly from Fig. 63, in terms of the ratio R_{eq}/X. The grid-condenser capacity C that is effective in determining X is the capacity C of the grid condenser actually present, plus the added capacity C_g' that represents the input capacity of the tube to the frequency of the rectified current. Since the tube

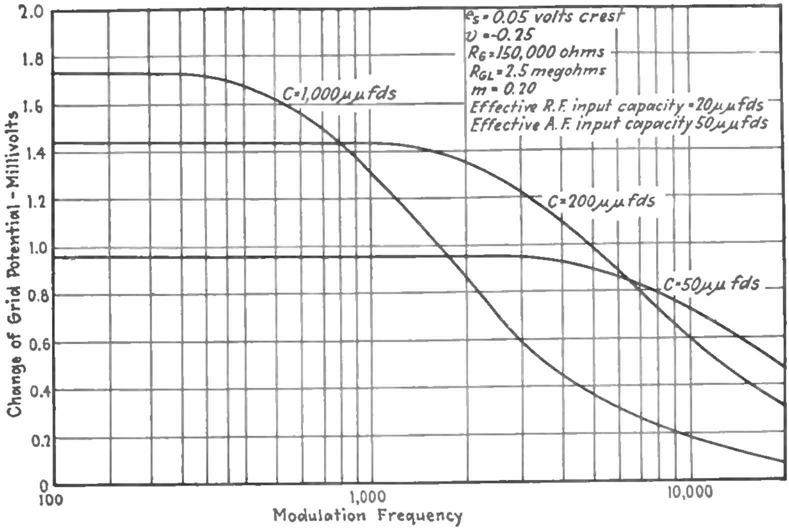


FIG. 145.—Curves showing the effect which varying the grid-condenser capacity has on the response of a weak-signal grid-leak rectifier. A large capacity causes the high notes to be discriminated against, while a very small capacity results in loss of sensitivity.

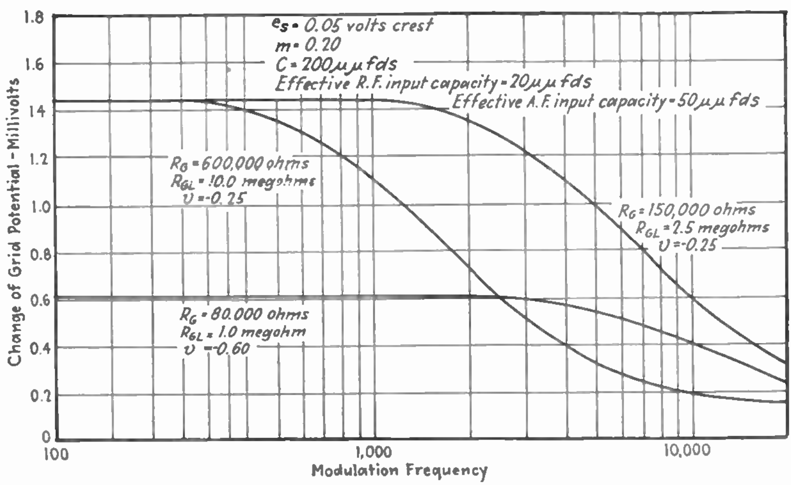


FIG. 146.—Curves showing effect which varying the grid-leak resistance has on the response of a weak-signal grid rectifier. A low-resistance leak gives poor sensitivity but low distortion, while a high-resistance leak causes the high frequencies to be discriminated against.

acts as an amplifier to this frequency the capacity C_g' will be greater than the actual tube capacity by an amount depending upon the load impedance in the plate circuit, as explained in Sec. 43.

Miscellaneous Comments.—The effects produced by varying the grid-leak resistance and grid-condenser capacity of a weak-signal grid-leak detector are shown in Figs. 145 and 146. Figure 145 brings out the effect which the grid-condenser capacity has upon the sensitivity and frequency distortion. If the grid condenser is small the full output is obtained up to high frequencies, but the sensitivity is seriously reduced by the loss of signal voltage in the grid condenser. On the other hand a condenser of very large capacity consumes a negligible fraction of the signal voltage but introduces excessive frequency distortion because of its tendency to by-pass the higher frequency components of the rectified grid current. The effect of varying the grid-leak resistance is shown in Fig. 146. If the leak resistance is low the dynamic grid resistance is also low, and the detection constant v is large, resulting in low sensitivity although the frequency distortion is low. If the leak resistance is great the operating point will be at a high value of grid resistance, which gives full sensitivity but increases the frequency distortion. Because of the shape of the $v - R_g$ characteristic, increasing the grid-leak resistance beyond the value that just places the operating point on the flat part of the $v - R_g$ curve has little effect on the response, while increasing the input resistance at the expense of added frequency distortion.

In selecting a tube to be used for grid rectification of weak signals, it is to be remembered that the rectifying properties of the grid circuit, *i.e.*, the $v - R_g$ characteristic, are determined by the type of cathode and are substantially independent of the tube construction and the plate voltage. Tubes with oxide-coated cathodes are the best grid rectifiers, *i.e.*, have the lowest v, and so should be used whenever possible. The remaining considerations are the same as in audio-frequency amplifiers, since the detector tube amplifies the voltage developed across the grid-leak resistance by the rectified grid current. This amplification may employ either resistance, impedance, or transformer coupling, and the tube and associated circuits should be selected accordingly. The only special features are that the plate load impedance to the radio-frequency signal should be low, so that the input capacity of the tube to the signal will be small; and that the plate-supply voltage must be somewhat less than is usually employed in amplifier operation since the grid-leak detector operates at substantially zero grid bias.

The weak-signal grid-leak detector has a square-law characteristic, that is, the rectified grid current is proportional to the square of the amplitude of the envelope of the wave being rectified. In an ordinary modulated wave this envelope varies in accordance with the signal that is to be reproduced, so that the rectified current is not a distortionless reproduction of the intelligence modulated on the signal. In the case of simple sine-wave modulation the rectified current is seen from Table IX to contain a direct-current component (which is unimportant), a modulation-frequency component (which is the desired output), and an additional component having twice the modulation frequency and representing a distortion frequency introduced by the square-law action of the rectifier. The ratio which the amplitude of this second-harmonic distortion frequency bears to the amplitude of the desired modulation-frequency component is $m/4$ and reaches a maximum of 25 per cent when the degree of modulation is 1.00. Another consequence of the square-law action is to make the rectifier less efficient with small signals than with large signals. The efficiency of rectification in the case of sine-wave modulation can be considered as the ratio of modulation-frequency component of the equivalent rectified voltage to mE_o, and is E_o/v.

63. Weak-signal Anode Rectification.—Weak signals can be rectified in the plate circuit by operating on the curved part of the grid-voltage

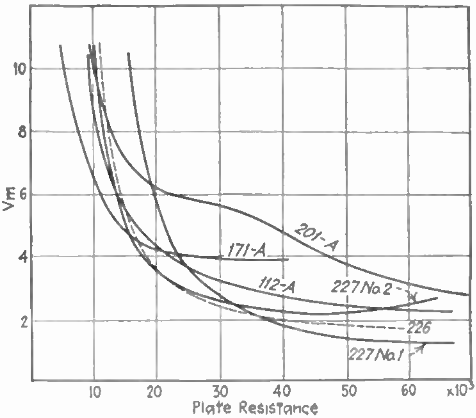


FIG. 147.—Variation of constant v_m for weak-signal anode rectification as a function of plate resistance R_p, for a number of typical tubes. The relative rectification efficiency of grid and anode rectifiers varies as v/v_m, so that a comparison of Fig. 144 with this figure shows the very great superiority of grid rectification for weak signals.

plate-current characteristic, and the theory of such detection is similar to the theory that has been developed for weak-signal grid-leak detection; that is, the rectifying action taking place in the plate circuit is exactly the same as though it were produced by an equivalent rectified voltage acting in a circuit consisting of the dynamic plate resistance in series with the load impedance connected in the plate circuit. The magnitude of the equivalent rectified voltage that can be considered as producing the rectified plate current is determined by the signal being rectified, and by a tube constant that is analogous to the constant v for grid rectification. This constant for plate rectification can be represented by the symbol v_p which is defined by the relation $v_p = 2R_p/(dR_p/dE_p)$, and when divided by the amplification factor μ of the tube gives a constant v_m [*i.e.*, $v_m = v_p/\mu = 2R_p/(\mu dR_p/dE_g)$], which is directly comparable with the constant v for grid rectification. Values of v_m for a number of tubes of different types are shown in Fig. 147. The constant v_m for weak-signal plate detection has a value that ranges from four to ten times v, which indicates that weak-signal anode rectification is correspondingly less efficient than weak-signal grid detection. The detector constant v_p (or v_m) for weak-signal anode rectification varies greatly for different tubes of the same type, and in many cases it is found that individual tubes of otherwise excellent characteristics are extremely poor weak-signal anode rectifiers. This is to be contrasted with the situation for grid rectification, where different tubes of the same type have almost

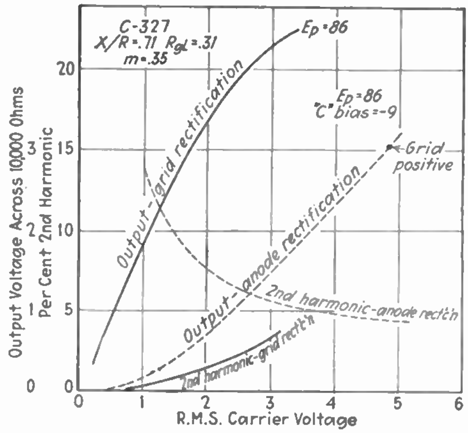


FIG. 148.—Curves showing power output and percentage of second harmonic as a function of carrier amplitude in grid-leak and anode power detectors. The grid rectifier has less distortion and is more efficient, particularly at moderate amplitudes, while the plate rectifier has a greater power capacity because it can be operated with higher plate-supply voltages.

identical characteristics. As a result of its low efficiency the weak-signal anode rectifier is practically never used.

64. Comparison of Detection Methods.—In weak-signal detection the grid-leak rectifier is so superior to the anode rectifier as to be used almost exclusively. The amplitude distortion is exactly the same for both types of weak-signal detectors because both have a square-law characteristic.

Satisfactory power detection may be obtained with either grid or plate rectification. Grid rectification has the advantage of delivering somewhat more output from a given signal voltage, particularly when this signal is only moderately large, but has a lower input resistance than the plate rectifier, and has a power capacity that is limited by the fact that a high plate-supply voltage cannot be used in grid rectifiers. While both plate and grid power rectifiers are approximately linear, the latter type is superior in this respect and so produces less amplitude distortion when the signal is a modulated wave. The type of performance that can be expected in power rectifiers is shown in Fig. 148, which gives the rectified output and the second-harmonic distortion component for typical anode and grid rectifiers. The grid rectifier is seen to have greater sensitivity, particularly when the signal voltage is only moderately large, and also has less amplitude distortion, as indicated by the amount of second harmonic produced in the output. Under conditions where the detector is not seriously overloaded, practically all of the amplitude distortion appears as the second harmonic of the modulation frequency, which therefore can be used as an indication of the degree of linearity present.

The chief differences in the performance of weak-signal and power detectors are in the efficiency and linearity of rectification. Power rectification is much more efficient, giving at least several times as much output voltage in proportion to the signal amplitude as does weak-signal detection, and so makes it possible to obtain a desired output with less amplification. At the same time power detection requires more amplification before detection (*i.e.*, more radio-frequency amplification), and less after (*i.e.*, less audio-frequency amplification), than does weak-signal rectification, which may or may not be an advantage, depending upon the circumstances. When the signal being rectified is a modulated wave, power detection is always preferred because power detectors have a linear characteristic and so give a rectified output that reproduces the modulation envelope with the minimum amount of amplitude distortion.

65. Heterodyne Detection.—When two signals of slightly different frequencies are superimposed, the envelope of the resulting oscillation varies in amplitude at a frequency that is equal to the difference between the frequencies of the two alternating currents, and swings through an amplitude range equal to the amplitude range of the smaller of the two voltages, as is shown in Fig. 149. This result is obtained because at one moment the two waves will be in phase and so will add together, while a short time later the higher frequency wave will be one-half cycle ahead of the other wave and so will combine with it in phase opposition. The rate at which the amplitude of the envelope varies is called the beat frequency (or the difference frequency), and the production of such beats by combining two waves is known as heterodyning. Since rectification of such a heterodyne signal gives a rectified current that varies in amplitude at the beat frequency, heterodyne action gives a means of changing the frequency of an alternating current.

The procedure for changing the frequency of an unmodulated wave by heterodyne action is to superimpose upon this signal a local oscillation having a frequency that differs from that of the signal by the desired frequency. The local oscillation may have either a higher or lower frequency than does the signal, since it is only the difference that is important. This heterodyne signal is then applied to a detector, and the desired beat frequency will be contained in the rectified output. If the wave that is to have its frequency changed is modulated, the amplitude of the beats that are produced by the superposition of the local oscillation will vary in accordance with the amplitude of the modulated

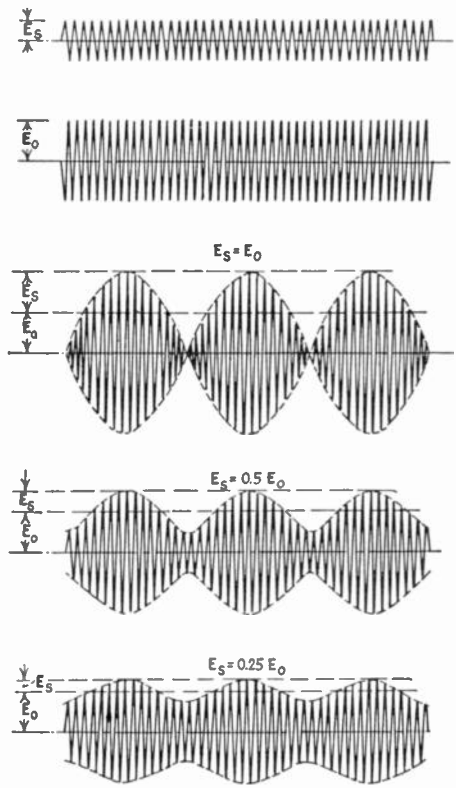


FIG. 149.—Typical heterodyne waves, showing how the combining of two waves of slightly different frequencies results in a wave which pulsates in amplitude at the difference frequency of the component waves, and how the wave shape of the envelope of the resultant wave depends upon the relative amplitudes of the two components.

wave, and the final result of the heterodyne operation is to change the frequency of the carrier wave to the new beat frequency without disturbing the character of the modulation.

The heterodyne principle of frequency changing has a number of important applications in radio communication. It can be used to change the frequency of a code radio signal to an audible frequency such as 1000 cycles, which can be used to actuate a telephone receiver. This result is accomplished by making the difference between the signal and local oscillation frequencies a suitable audio frequency, and is known as heterodyne code reception. Another application of the heterodyne principle is in the super-heterodyne type of radio-frequency amplification, in which the heterodyne principle is used to change the carrier frequency of the radio signal to a predetermined and readily amplifiable radio frequency at which the amplification takes place. In this way the frequency of the signal is changed to fit the amplifier, rather than adjusting the amplifier to fit the signal. Heterodyne action can also be employed to separate frequencies that differ from each other by a relatively small percentage. Thus it would practically be impossible to separate currents having frequencies of 1,000,000 and 1,001,000 cyles by the use of tuned circuits, but this can be readily accomplished by the use of heterodyne action. Thus if the local oscillation has a frequency of 999,000 cycles the heterodyne action will change the original frequencies to 1000 and 2000 cycles, respectively, which can be separated with ease.

Analysis of Heterodyne Detection.—The object of heterodyne detection is to produce sinusoidally varying currents of the difference frequency. When this result is accomplished the detection is distortionless, while, if the detector also produces harmonics of the beat frequency, distortion is introduced. The characteristics required for distortionless detection can be determined by considering the equation of the envelope of a heterodyne signal. When two sine waves of amplitudes E_o and E_s and having a difference frequency $\omega/2\pi$ are superimposed, the shape of the resultant envelope depends upon the relative amplitudes of the two waves, as shown in Fig. 149, and has the equation[1]

$$\text{Instantaneous amplitude of envelope} = \sqrt{E_s^2 + E_o^2 + 2E_sE_o \sin \omega t} \tag{119}$$

The shape of the resulting envelope is identical with the output wave of a full-wave rectifier when the superimposed voltages are of equal amplitude, and approaches a sine-wave variation only in the limit when one of the components is extremely small compared with the other. *The shape of the envelope is such that square-law detection, that is, rectification in which*

[1] In its most general form the equation of the envelope of a heterodyne signal can be written as

$$e = E_o \sin \omega t + E_1 \sin [(\omega + \delta_1)t + \phi_1] + E_2 \sin [(\omega + \delta_2)t + \phi_2] + \cdots \tag{120}$$

where ϕ_1, ϕ_2, etc., are phase angle constants, and δ_1, δ_2, etc., are 2π times the frequency by which their respective terms differ from the frequency of the E_o term. The envelope of Eq. (120) can be found by solving Eq. (120) for zero, equating to zero the *partial derivative with respect to* ω of this transposed equation, and simultaneously solving the resultant equation with Eq. (120) to eliminate ω. Carrying out these operations gives the result

$$\begin{aligned}\text{Envelope} = [&E_o^2 + E_1^2 + E_2^2 + \cdots \\ &+ 2E_oE_1 \cos (\delta_1 t + \phi_1) + 2E_oE_2 \cos (\delta_2 t + \phi_2) + \cdots \\ &+ 2E_1E_2 \cos [(\delta_1 - \delta_2)t + (\phi_1 - \phi_2) + \cdots]^{1/2}\end{aligned} \tag{121}$$

When only two components are present and ϕ_1 is taken as zero, this equation reduces to Eq. (119).

the output is proportional to the square of the envelope, gives distortionless heterodyne detection, while linear rectification, by reproducing the envelope of the heterodyne signal, fails to do so.

The character of the distortion that results from linear detection of a heterodyne signal can be obtained from a Fourier analysis of the envelope equation. When this is done the following is obtained:

$$\begin{aligned} e = {} & E_o\sqrt{1 + r^2}(1 - 0.0625k^2 - 0.0146k^4 - 0.0064k^6 - \cdots) \\ & + \frac{E_s \cos \omega t}{\sqrt{1 + r^2}}(1 + 0.0938k^2 + 0.0341k^4 + \cdots) \\ & - \frac{E_s r \cos 2\omega t}{4(1 + r^2)^{3/2}}(1 + 0.313k^2 + 0.1535k^4 + \cdots) \\ & + \frac{E_s r^2 \cos 3\omega t}{8(1 + r^2)^{5/2}}(1 + 0.548k^2 + \cdots) \\ & - \cdots \end{aligned} \tag{122}$$

where

E_s = weaker component of signal voltage
E_o = larger component of signal voltage
$r = E_s/E_o$ = ratio of weak to strong signal components
$k = 2r/(1 + r^2)$
$\omega = 2\pi$ times the difference frequency of E_s and E_o.

In studying these equations it is of assistance to note that k depends on the ratio E_s/E_o, and reaches a maximum value of unity when the two superimposed waves are of equal amplitude. For this condition the equation of the envelope simplifies to that of the output wave of a full-wave rectifier, which is

$$e = E_s(1.274 + 0.851 \cos \omega t - 0.170 \cos 2\omega t + 0.0729 \cos 3\omega t \cdots) \tag{123}$$

Equations (122) and (123) incorporate the fundamental features involved in the linear detection of heterodyne signals. They show that the difference-frequency output is largely independent of the amplitude of the stronger signal component E_o and is nearly proportional to the strength of the weaker component E_s. The magnitude of the deviation from this approximate relation is indicated by the fact that increasing the stronger signal component E_o from equality with E_s to a value many times E_s, while holding the latter constant, increases the difference-frequency output by approximately 18 per cent. The amplitude of the distortion frequencies produced in linear detection of heterodyne signals is greatest when the two signal components are of equal size, under which condition the second harmonic of the beat frequency is 20 per cent.

When the signal that is to have its frequency changed by heterodyne action contains several frequency components, square-law detection of the heterodyne signal produces an output that contains every possible difference frequency that is present in the heterodyne signal, and these various difference-frequency components of the output each have an amplitude proportional to the product of the amplitudes of the two waves producing the difference frequency. Linear detection of such complex heterodyne signals gives results that are relatively difficult to analyze quantitatively because of the complex character of the envelope of the heterodyne signal, but the character of the behavior that can be expected from linear detection can be estimated qualitatively by the action in the simple case of two components. These considerations relating to the relative behavior of square-law and linear detection of heterodyne signals show that square-law detection is to be preferred from the point of view of avoiding distortion, and that if linear detection is employed the superimposed local oscillation should have an amplitude that is much larger than that of the signal which is to have its frequency changed. Under practical circumstances linear detection is usually employed because the amplitude of the local oscillation is relatively large, and it is almost impossible to maintain square-law action when this is the case.

The difference-frequency output developed by the linear detection of a heterodyne signal consisting of a strong local oscillation superimposed upon a weak signal is almost exactly the same output that would be obtained by applying to the grid of the detector a difference-frequency voltage equal to the weak-signal input multiplied by the efficiency of detection, and then considering the detector to be merely an amplifier tube.

The output of a detector that is rectifying a heterodyne signal always contains components other than the difference-frequency current and its harmonics. When the heterodyne signal contains two components, the most important of such new frequencies has a frequency that is the sum of the two frequencies being combined, while harmonics of this sum frequency, and complex combinations of the two frequencies of the two components, are also usually present in small amplitudes. The exact nature of these additional products of rectification depends upon the detector characteristics and is of little importance since these components have not been found to have practical usefulness.

66. Regenerative Detectors.—In examining the action taking place in detectors employing either grid or plate rectification it will be noted that there are signal currents flowing in the plate circuit of the detector, in addition to the products of rectification, as is clearly shown in Figs. 134 and 142. It is possible to obtain regeneration by feeding back a portion of this signal energy to the circuits associated with the detector input. Regeneration produced in this way by utilizing the radio-frequency energy in the detector plate circuit can be more readily controlled than regeneration in amplifiers, and is frequently used to increase the amplification and selectivity of radio receivers.

There are a number of ways by which regenerative action may be obtained in vacuum-tube detectors, some of the more common of which are illustrated in Fig. 150. The arrangement shown at Fig. 150*a* employs a small inductance, often called a tickler coil, in the detector plate circuit to transfer the amplified energy to the resonant circuit supplying the signal input, and controls the amount of regeneration by varying the mutual inductance. The circuit of Fig. 150*b* takes advantage of the fact that an inductive reactance in the plate circuit causes energy to be transferred to the grid circuit through the grid-plate tube capacity, and uses a variable inductance to control the magnitude of regeneration. In Figs. 150*c* and 150*d* the regeneration is obtained through mutual inductance between a plate coupling coil and the tuned input circuit as in Fig. 150*a*, but these circuits differ in the methods used to control the regeneration.

Regeneration in detectors produces exactly the same action as regeneration in radio-frequency amplifiers, since the regenerative detector is

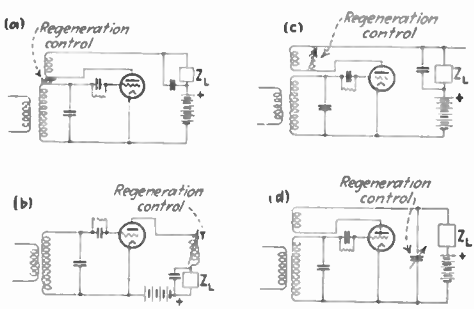


FIG. 150.—Commonly used regenerative and oscillating detector circuits. While grid-leak detection is shown, the use of anode rectification has no effect on the regenerative action.

essentially a radio-frequency amplifier as far as the feed back is concerned. The effect of regeneration, no matter how produced, is equivalent to altering the effective resistance and effective reactance of the input circuit. The change of reactance caused by regeneration alters the resonant frequency slightly, making the resonance point somewhat dependent upon the adjustment of the regenerative control, but this effect is small because the reactance change amounts to only a few ohms and so is small compared with the large reactances in the tuned circuit. The change of resistance resulting from regeneration is much more important because the resistance of the tuned input circuit is so low that a few ohms added or subtracted represents a large percentage variation. When the energy is fed back in the proper phase to reinforce the applied signal the effect is to neutralize a part of the resistance of the tuned input circuit. This raises the effective Q and so increases the resonant rise of voltage (which is equivalent to added amplification), as well as making the selectivity greater.

Analysis of Typical Regenerative Circuit.—The quantitative effect of regeneration can be determined to a first approximation by setting up an equivalent circuit and writing down the equations that give the voltage

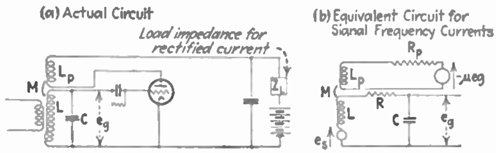


FIG. 151.—Equivalent electrical circuit used in analyzing regenerative action.

and current relations. Such an equivalent circuit for the method of regeneration shown at Fig. 150*a* is given in Fig. 151. Analysis of this circuit leads to the following result[1]

$$\text{Voltage developed across tuned input circuit} = -je_s\frac{X_c}{Z}\frac{1}{1-\dfrac{X_c}{Z}\dfrac{\mu\omega M}{R_p+j\omega L_p}} \quad (124a)$$

where

$X_c = 1/2\pi fC$ = reactance of condenser C
$Z = R + j(\omega L - 1/\omega C)$ = series impedance of tuned circuit
μ = amplification factor of tube
R_p = plate resistance of tube
e_s = signal voltage acting in input circuit
M = mutual inductance between tickler coil and coil of input circuit

[1] This analysis follows: The voltage developed across the input circuit is a result of the combined action of the signal e_s and the voltage $j\omega Mi_p$, which is induced in the input circuit by the tickler coil L_p. The total voltage acting in the input circuit is hence $(e_s + j\omega Mi_p)$, so that

$$e_g = -\frac{(e_s + j\omega Mi_p)jX_c}{Z}$$

But

$$i_p = \frac{-\mu e_g}{R_p + j\omega L_p} = \frac{-\mu e_g(R_p - j\omega L_p)}{R_p^2 + (\omega L_p)^2}$$

Substituting this value for i_p in the above expression for e_g, and solving for e_g results in Eq. (124*a*).

At the resonant frequency of the tuned circuit, $X_c/Z = \omega L/R$, and Eq. (124*a*) can be written as

Voltage across tuned circuit at its resonant frequency =

$$-je_s\frac{\omega L}{R - \dfrac{\omega L\mu(\omega M)R_p}{R_p^2 + (\omega L_p)^2} + j\omega L_p\dfrac{\omega L\mu(\omega M)}{R_p^2 + (\omega L_p)^2}} \quad (124b)$$

Regeneration has therefore changed the effective resistance of the circuit from R to

$$\left(R - \omega L\mu(\omega M)\frac{R_p}{R_p^2 + (\omega L_p)^2}\right)$$

and has had the effect of adding a reactance $\dfrac{j\omega L_p(\omega L)\mu(\omega M)}{R_p^2 + (\omega L_p)^2}$ in series with the tuned circuit.

L_p = inductance of tickler coil
$\omega = 2\pi$ times frequency.

Equation (124a) is approximate in that it assumes that the plate resistance of the tube is constant irrespective of the amplitude of the signal voltage. Actually the curvature of the plate-voltage plate-current characteristic is such that the effective dynamic plate resistance increases slightly as the signal amplitude increases. While this change amounts to only a very small percentage, and is unimportant under ordinary circumstances, it cannot be neglected when the regenerative action is great enough very nearly to neutralize the actual resistance of the tuned input circuit, because then even a slight error in the assumed plate resistance will cause a very large error in the resultant calculated regeneration. Thus while Eq. (124a) holds when the regeneration is not large, it fails when the regeneration is increased to the point where the resultant circuit resistance approaches zero. It can be shown both experimentally and theoretically that the amount of amplification obtainable with critical regeneration [*i.e.*, when the regeneration is such that the effective resistance as given by Eq. (124a) is exactly zero, which is the maximum regeneration possible without causing oscillations] is inversely proportional to the two-thirds power of the signal voltage being amplified and is directly proportional to the response obtained when no regeneration is present.[1] The result is that the same adjustment that gives an enormous regenerative amplification of an extremely small signal will give only one-hundredth as much amplification with a signal 1000 times as large. Furthermore the response with critical regeneration varies with the actual Q of the circuit in the same way as does the response without regeneration, so that a high circuit Q is important if maximum results are to be obtained. When the regeneration is appreciably less than the critical value the amplification will not be so dependent upon the signal amplitude and will approach the values calculated from equations of the type of Eq. (124a).

While representing an inexpensive means of obtaining radio-frequency amplification, regeneration has several disadvantages. In the first place regenerative amplification is obtained by lowering the effective resistance of a tuned circuit, and, since this also greatly increases the selectivity, regeneration tends to suppress the higher side-band frequencies contained in the signal. In the second place the adjustments required to give satisfactory regenerative amplification also depend upon the frequency of the signal, as is apparent from examination of Eq. (124a), so that it is necessary to readjust the regeneration controls for every new signal. Furthermore the adjustments required to give appreciable regenerative action are rather critical, and a certain amount of skill is required to

[1] See Balth. van der Pol, The Effect of Regeneration on the Received Signal Strength, *Proc. I.R.E.*, vol. 17, p. 339, February, 1929.

carry them out properly. Finally, when the regenerative action is carried to the point where the circuit resistance is completely neutralized and becomes negative, as will inevitably occur from time to time as the result of accidental improper adjustments, oscillations will be set up which will heterodyne with any signal that may be present and produce annoying squeals. In many types of receivers such oscillations reach the antenna and cause the receiver to act as a small transmitter having a range that may be as great as 10 miles. These disadvantages of regenerative amplification are so great that it is generally considered better practice to obtain radio-frequency amplification by the use of tuned radio-frequency amplifiers rather than by regenerative action.

67. Oscillating Detectors.—When regeneration is increased to the point where the resistance of the resonant circuit is completely neutralized there will be set up oscillations which will heterodyne with any signal currents in the resonant circuit. The resultant beats are rectified by the detector and cause difference-frequency currents to appear in the detector output. The oscillating detector therefore acts as a heterodyne detector in which the detector tube generates the heterodyne oscillations, as well as functioning as a rectifier. The circuits used in oscillating detectors are the same as those employed for regenerative detectors, the only difference being that the regeneration is increased to the point where oscillations are produced. Grid rectification is always employed in oscillating detectors because of its high sensitivity, and because it automatically supplies the proper grid bias for the oscillations. The heterodyne oscillation that is applied to the grid of the oscillating detector when a signal is combined with the local oscillations, is rectified in the usual manner by the grid leak and grid condenser, with the result that the plate current varies in amplitude at the beat frequency and so represents the original signal modulated upon the beat frequency. The grid leak and condenser should be proportioned according to the usual principles of grid-leak power detection, as is also the plate load impedance that utilizes the rectified output.

The oscillating detector is superior in a number of respects to a heterodyne detector using a separate oscillator. The most important of these are the much greater sensitivity of the oscillating detector, its great stability, and the saving of a separate oscillator and the resulting separate adjustments. When the desired beat frequency is low, as is nearly always the case, the resonant circuit that determines the frequency of the oscillations is automatically brought substantially in resonance with the incoming signal when the local oscillation is of the required frequency, while with a separate heterodyne oscillator it is necessary to tune in the signal and adjust the frequency of the local oscillations with independent circuits. The use of the same resonant circuit to produce the oscillations and tune in the signal means that the resonant circuit is slightly detuned to the signal, but unless the beat frequency is at least several per cent of the carrier frequency the loss from detuning does not outweigh the other advantages of the oscillating detector as compared with a separate heterodyne oscillator.

The great sensitivity of the oscillating detector is a result of the large regenerative amplification which the signal undergoes before being rectified. When no signal is present the oscillations have an amplitude such that the effective plate resistance of the tube has a value for which the regeneration exactly neutralizes the resistance of the resonant circuit. When this condition exists the regenerative amplification to a superimposed oscillation is very great because the situation is much the same as that which exists in an ordinary regenerative detector adjusted to give the maximum possible regeneration. The difference in the two cases however is that with the regenerative detector this adjustment is very critical and impossible to maintain, whereas in the oscillating detector the oscillations automatically assume an amplitude that picks out this critical condition and maintains it with complete stability.

The regenerative amplification which the signal undergoes in an oscillating detector can be analyzed by considering that the signal represents a voltage that is induced in the resonant circuit in addition to the voltage induced by the feed back from the plate circuit. The phase of the signal with respect to the oscillation changes from aiding to opposition at a rate corresponding to the difference frequency, as in the case of any heterodyne signal. To a first-order approximation the variation in the amplitude of the resultant wave applied to the grid of the tube acts as though there were no signal voltage applied to the detector, but rather that the regeneration of the oscillating detector was alternately decreased and increased from its actual value at a rate corresponding to the beat frequency. The oscillating detector hence has its greatest sensitivity when a small change in the regeneration will produce a large change in the amplitude of the generated oscillations. This condition is always realized when the regeneration has the smallest value at which oscillations will exist, and when the resonant circuit has the highest possible Q (*i.e.*, lowest possible actual resistance).

Threshold (Fringe) Howl in Oscillating Detectors.—Many oscillating detectors produce a sustained audio-frequency sound when the adjustment is such that oscillations are just barely maintained. This is known as threshold or "fringe" howl,[1] and is to be avoided since it occurs under conditions for which the oscillating detector is most sensitive. The most common cause of threshold howl is a result of the following action. When the regeneration does not greatly exceed the minimum value at which oscillations can be maintained, it will be found that the direct-

[1] See L. S. B. Alder, Threshold Howl in Reaction Receivers, *Exp. Wireless and Wireless Eng.*, vol. 7, p. 197, April, 1930.

current plate current will decrease as the plate voltage is increased. This is because the amplitude of the generated oscillations increases faster than the plate voltage when the amplitude is small, and the extra amplitude increases the grid bias produced by the oscillations faster than the plate voltage increases. If there is an inductive load in the plate circuit, such as might be supplied by the primary of an audio-frequency transformer, any effect that momentarily increases the amplitude of the oscillations will tend to reduce the plate current. This causes a voltage to be induced across the plate load inductance in such a direction as to increase the voltage applied to the plate of the tube, which further increases the amplitude of the oscillations, causing still more tendency for the plate current to decrease, and so on. The result is that the oscillations build up to an abnormally large amplitude that can be maintained only by an induced voltage across the plate load inductance. When this induced voltage begins to fall off, as it will when the plate current begins to decrease less rapidly, the oscillations start to die out. As their amplitude decreases the plate current tends to increase, inducing a voltage across the plate load inductance that reduces the voltage applied to the plate of the tube below that of the plate-supply source, which still further decreases the amplitude of oscillations, and so on. It is thus seen that the inductive load impedance supplies energy to the oscillations during the period that they are building up and abstracts it from them during the time when their amplitude is decreasing, and if the amount of energy that is thus transferred is sufficiently large the process will be continuous and will result in the production of oscillations modulated at a rate determined by the time constant of the plate circuit.

Threshold howl can be eliminated by the use of a plate load impedance that does not return energy to the oscillations in the proper phase to produce howl. In the case of grid-leak detection this means that inductive load impedances must be avoided, while with anode detection, plate loads with capacitive reactance can be expected to give trouble. When threshold howl occurs in transformer-coupled detectors using grid leak and grid condenser, the usual remedy is to make the load impedance less inductive by shunting a resistance across the transformer terminals. When across the primary the proper resistance will approximate the plate resistance of the tube, and when placed between secondary terminals will approximate the plate resistance multiplied by the square of the turn ratio.

Threshold howl is sometimes caused by the generation of interrupted oscillations of the type described in Sec. 52 rather than by the production of modulated oscillations as has just been described. The cause of the howl in a particular case can be readily traced down by taking advantage of the fact that when intermittent oscillations are generated, the frequency of interruption depends upon the size of grid leak and grid condenser employed, whereas when the howl is the result of modulated oscillations produced by reactance in the plate circuit of the tube, the frequency is determined by the time constant of the plate circuit and is independent of the grid-leak and grid-condenser proportions. When trouble from interrupted oscillations is experienced, it can usually be eliminated by the use of a lower resistance grid leak, or a smaller grid-condenser capacity, or both.

68. Superregeneration.—Superregeneration is a form of regenerative amplification obtained by varying the effective resistance of a tuned circuit from positive to negative values at a low radio-frequency rate, while maintaining the average resistance positive. A simple circuit giving superregeneration is shown in Fig. 152. This consists of a tube arranged to produce regeneration in the manner shown in Fig. 150*a* but supplied with a plate voltage that is a low radio frequency, such as 25 kc. During the half-cycles when the plate is positive, regeneration takes place and makes the effective resistance of the tuned circuit negative, while during the half-cycles that the plate-supply voltage is negative the tube is inoperative and the effective resistance of the tuned circuit is its actual resistance. The result is that the arrangement alternately tends to build up and suppress oscillations in the tuned circuit, but the effective resistance of the circuit is negative for such a short time that the oscillations that tend to build up during this period are suppressed before reaching their full size and as a result attain an amplitude that is proportional to the amplitude of whatever voltage is present in the circuit to supply the initial impulse that starts the building-up process. The oscillations that are thus built up must be completely suppressed during the interval when the circuit resistance is positive. Otherwise the voltage remaining at the end of the positive resistance interval would act as the initial impulse controlling the size of oscillations that would be obtained during the succeeding negative resistance period, with the final result that the amplitude of the oscillations would increase indefinitely. It is therefore necessary that the effective resistance during the dying-out period be somewhat greater than the effective resistance during the building-up period, *i.e.*, the circuit must have an average effective resistance that is positive.

If a modulated radio-signal voltage is acting in the tuned input circuit the initial impulse that starts the building up during the negative resistance interval will be the amplitude of the signal at that instant, so that the maximum voltage which the oscillations reach during the building-up period will also be proportional to the signal. During the time when the circuit resistance is positive these oscillations are suppressed, to be built up again to an amplitude proportional to the voltage produced by the signal currents. The result is that the voltage developed across the tuned circuit is an oscillation modulated at a frequency corresponding to the rate at which the resistance of the tuned circuit is varied, and this modulated wave is also modulated according to the modulation of the original radio-frequency signal. The detailed mechanism of this action can be understood by a study of Fig. 152, which shows the signal voltage acting in the tuned circuit, the effective resistance of the tuned circuit,

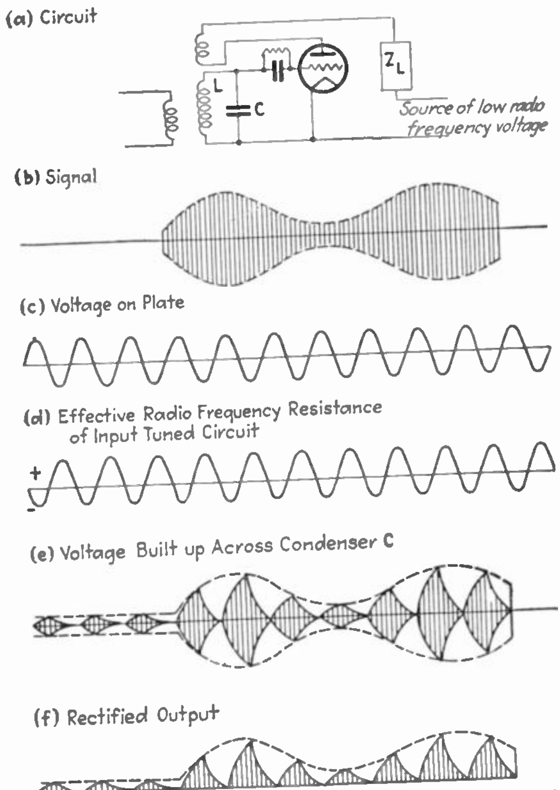


FIG. 152.—Simple superregenerative circuit, together with details showing the mechanism by which superregenerative amplification is obtained.

and the voltage that is built up across the tuned circuit as a result of superregeneration.

By properly proportioning the circuit, and in particular by using circuits that are more suitable than the very simple arrangement shown in Fig. 152, it is possible to obtain enormous amplifications by the use of superregeneration. The adjustments that are required to obtain these amplifications are rather critical, however, and the amplification is not selective, *i.e.*, does not discriminate against signals of different frequency. The result is that superregeneration is not used except in radio receivers operating at extremely high frequencies, such as 100,000 kc or more, where amplification of any other type is practically impossible to obtain.[1]

69. Miscellaneous Features of Detection. *Plate By-pass Condenser.* The condenser that is shunted across the detector plate load impedance for the purpose of by-passing the radio-frequency signal currents in the plate circuit represents a capacitive plate load reactance to these currents. It was shown in Sec. 43 that such a plate load reactance causes the tube to have a positive input resistance because of energy transferred between grid and plate circuits through the grid-plate tube capacity. Since this input resistance is shunted across the resonant circuit that develops the signal voltage applied to the detector grid, its value must be kept high if the effective resistance of the tuned input circuit is not to be raised appreciably, and this requires a plate by-pass condenser capacity that is not too small.

The magnitude of the input resistance produced as a result of the capacitive plate load reactance furnished to the signal by the by-pass condenser can be calculated by Eq. (91). Since the reactance of the by-pass condenser to the signal frequencies is low compared with the plate resistance of the tube, the ratio which the voltage developed across the by-pass condenser bears to the voltage applied to the grid, *i.e.*, the tube amplification A in Eq. (91), is $(\mu/R_p)(1/\omega C)$, while the voltage developed across the condenser lags the equivalent voltage acting in the plate circuit by substantially 90°. With these simplifications, Eq. (91) becomes

$$\text{Input resistance caused by plate by-pass condenser} = \frac{1}{g_m}\frac{C}{C_{gp}} \qquad (125)$$

where

$g_m = \mu/R_p$ = mutual conductance of tube
C = capacity of plate by-pass condenser
C_{gp} = grid plate tube capacity.

Equation (125) shows that the plate by-pass condenser is the cause of an input resistance that is independent of frequency but directly proportional to the ratio of by-pass to tube grid-plate capacity, and inversely proportional to the mutual conductance of the tube. In order to keep the input resistance high, the by-pass condenser must be much larger than is necessary for mere by-passing of high-frequency currents. Thus with a tube having a grid-plate capacity of 10 $\mu\mu$f and a mutual conductance of 1000 μmhos, the by-pass condenser must have a capacity of at least 5000 $\mu\mu$f to keep the input resistance in excess of 500,000 ohms. The plate by-pass condenser should however be no larger than necessary

[1] For further information about superregeneration, and in particular for diagrams of suitable circuits see: Edwin H. Armstrong, Some Recent Developments of Regenerative Circuits, *Proc. I.R.E.*, vol. 10, p. 244, August, 1922.

to keep the tube input resistance up to the required value, because the by-pass condenser shunts the plate load impedance that utilizes the rectified current and thereby tends to reduce the response at the higher output frequencies.

Apparent Selectivity Obtained by Linear Detection.[1]—When two modulated signals of widely different amplitudes, and of carrier frequencies that produce a beat frequency so high as to be inaudible, are simultaneously applied to the input of a linear rectifier, it is found that the weaker of the two signals is not rectified. The reason for this behavior lies in the character of the envelope that results when two modulated waves of different frequencies are superimposed. The resultant envelope has an inaudible variation that represents the difference frequency between the two carriers and which is modulated at the frequency of modulation of the weaker signal, as shown in Fig. 153; while the major envelope is modulated in accordance with the stronger signal. When such a wave is rectified the result is as shown at *d* in Fig. 153, and is seen to contain no component that varies at the modulation frequency of the weaker signal. The weak signal appears only as a modulation of the inaudible beat frequency between carriers, so that the effect of the weaker signal on the output is suppressed, giving what is equivalent to an increase in the effective selectivity. In order that this suppression of the weaker signal may be complete it is necessary that the strong signal be considerably larger than the weaker, and that the detector be exactly linear. If either of these requirements is not entirely met the suppression is only partial. With ordinary rectifiers the suppression is

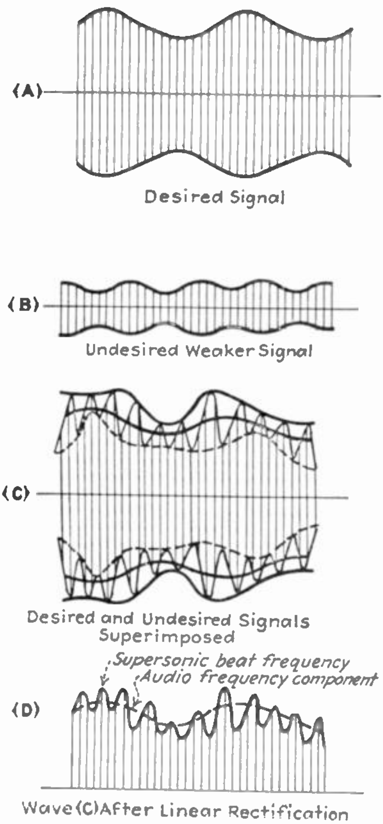


FIG. 153.—Wave forms obtained in the linear detection of a signal consisting of a weak modulated wave superimposed upon a strong modulated wave. The rectified output of the linear detector is seen to contain no component varying at the modulation frequency of the weaker signal.

[1] See R. T. Beatty, Apparent Demodulation of a Weak Signal by a Stronger One, *Exp. Wireless and Wireless Eng.*, vol. 5, p. 300, June, 1928; S. Butterworth, Note on the Apparent Demodulation of a Weak Signal by a Stronger One, *Exp. Wireless and Wireless Eng.*, vol. 6, p. 619, November, 1929.

very great when the ratio of signal amplitudes is at least 2:1, and this action is a valuable property of linear detectors.

Gaseous Detectors.—When a small amount of gas is allowed to remain in a detector tube the characteristics are altered to a considerable extent by the ionization which the gas molecules suffer as a result of collisions with electrons that are traveling toward the anode. One effect of this ionization is to cause the characteristic curves of the tube to have sharp but small "kinks" that represent sudden changes of curvature. Such kinks will give unusually efficient rectification of small signals, but as the location of the kink is very sensitive to grid, plate, and filament potentials, very critical circuit adjustments are required to locate the operating point so as to take advantage of this sudden bend. Gaseous detectors containing an alkali vapor, such as caesium, can be constructed in such a way as to use positive ion drift to produce rectification. Such tubes are somewhat more sensitive rectifiers than those of the high-vacuum type and are not critical as to electrode voltages, but have the disadvantage of a high noise level and so have not been commercially successful.

Diode Rectifiers.—In the early days of radio, two-electrode vacuum tubes consisting of a cathode and a plate were used for detection. These were called Fleming valves after their inventor, Prof. J. S. Fleming, and are used to a limited extent in radio receivers employing automatic volume control. As a detector of small radio-frequency signals the diode or Fleming valve is distinctly inferior to the three-electrode tube because the output energy is obtained directly from the radio-frequency-signal energy by rectification without any amplification.

CHAPTER IX

SPECIAL TYPES OF VACUUM TUBES

70. Screen-grid Tubes.—A screen-grid tube is essentially a three-electrode vacuum tube to which there has been added a second grid that completely surrounds the plate, as in Figs. 154 and 155. This extra electrode is called the screen grid to distinguish it from the usual, or control, grid and serves as an electrostatic shield that eliminates substantially all direct capacity between the inner grid and the plate and so prevents the transfer of energy between output and input circuits through

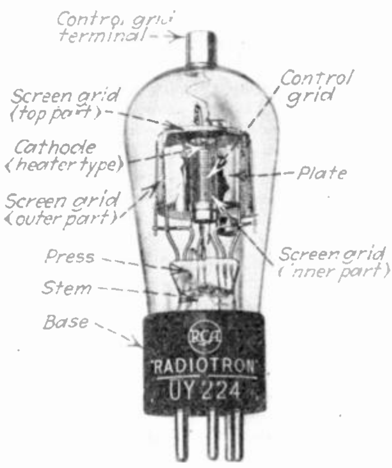


FIG. 154.—Screen-grid tube cut away to show constructional details. This particular tube has a heater cathode.

the tube grid-plate capacity. The screen grid is operated at a positive potential, but most of the electrons which it attracts pass through its meshes and go to the plate, which is also positive, so that while the screen grid serves as an almost perfect electrostatic shield it intercepts only a relatively small proportion of the total electron flow.

The most important use of the screen-grid tube is in the amplification of radio-frequency voltages, for which purpose it is particularly well suited because the screen grid eliminates practically all energy transfer through the grid-plate tube capacity. Screen-grid tubes are also used to some extent in the amplification of audio frequencies, as detectors, and to perform a number of other functions.

The properties of screen-grid tubes are incorporated in characteristic curves similar in nature to the characteristic curves of the three-electrode tube, but with the difference that, since there are two anodes instead of one, a larger number of curves are required to give a complete picture of the behavior of the tube.

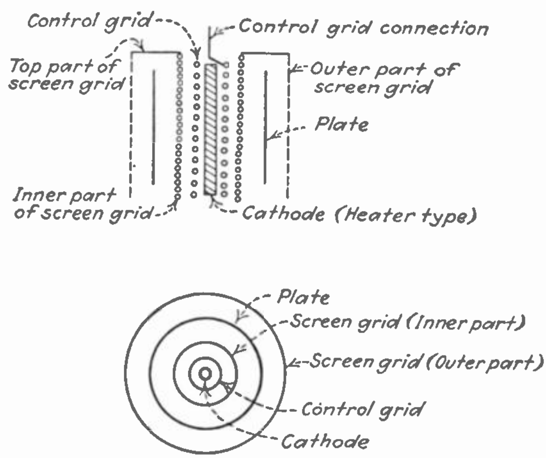


FIG. 155.—Section taken through a screen-grid tube to show the electrode arrangement.

Factors Determining Total Space Current.—The total space current of a screen-grid tube, *i.e.*, the sum of the screen-grid and plate currents, is determined by the electrostatic field in the immediate vicinity of the cathode, exactly as is the case in three-electrode tubes. In the case of the screen-grid tube this electrostatic field is proportional to the quantity $\left(E_g + \frac{E_{sg}}{\mu_{sg}} + \frac{E_p}{\mu_p}\right)$, where μ_{sg} and μ_p are constants determined by the tube construction, and E_g, E_{sg}, and E_p are potentials of the control grid, screen grid, and plate, respectively. The space current in the screen grid is therefore the same function of $\left(E_g + \frac{E_{sg}}{\mu_{sg}} + \frac{E_p}{\mu_p}\right)$ as the plate current of the three-electrode tube is of $\left(E_g + \frac{E_p}{\mu}\right)$. With a heater-type cathode it is therefore possible to write

$$\text{Total space current} = i_p + i_{sg} = K\left(E_g + \frac{E_{sg}}{\mu_{sg}} + \frac{E_p}{\mu_p}\right)^{3/2} \qquad (126)$$

where K is a constant determined by the tube construction, and i_p and i_{sg} represent the plate and screen-grid currents, respectively.

The tube constants μ_{sg} and μ_p that appear in Eq. (126) are determined by the relative effects which potentials applied to the screen grid and the plate, respectively, have on the electrostatic field at the surface of the cathode as compared with the effect which potentials applied to the control grid have. The quantity μ_{sg} therefore represents the amplification factor that would be obtained in the tube if the screen grid were the only anode electrode, and is determined primarily by the construction of the control grid, being higher as the control grid becomes a more perfect shield. The quantity μ_p is a measure of the extent to which the combined action of the screen grid and control grid shield the cathode from the plate potential, and is numerically equal to the product of μ_{sg} and a factor that represents the relative effect which screen-grid and plate voltages have in producing electrostatic field at the surface of the cathode.

When there is a voltage drop in the cathode Eq. (126) still holds when applied to an elementary length of the filament and can be used to derive modified equations for the space current, exactly as was done in the case of the three-electrode tube. The effect of a voltage drop in the cathode is to cause the space current to vary with a power of $\left(E_g + \frac{E_{sg}}{\mu_{sg}} + \frac{E_p}{\mu_p}\right)$ that is somewhere between $\frac{5}{2}$ and $\frac{3}{2}$, all potentials being measured with respect to the negative side of the cathode.

Since the purpose of the screen grid is to place an electrostatic shield between the plate and the control grid, the tube construction is always such that the constant μ_p (which is a measure of the completeness of shielding) will be very large. The plate voltage in ordinary screen-grid tubes therefore has a negligible effect on the total space current, so that for all practical purposes Eq. (126) can be reduced to the much simpler approximate relation

$$\text{Total space current} = i_p + i_{sg} = K\left(E_g + \frac{E_{sg}}{\mu_{sg}}\right)^{3/2} \tag{127}$$

where

i_p = plate current
i_{sg} = screen-grid current
E_g = control-grid voltage
E_{sg} = screen-grid voltage
K = tube constant depending upon tube construction
μ_{sg} = amplification factor, which takes into account relative effects of control-grid and screen-grid potentials in producing electrostatic field at the cathode, and is determined primarily by the control grid construction.

When there is a voltage drop in the cathode Eq. (127) must be modified to take this into account, just as it was necessary to modify Eq. (59) in the case of three-electrode tubes having a voltage drop in the cathode.

A comparison of Eqs. (59) and (127) shows that the total space current in the screen-grid tube is the same function of control-grid and screen-grid potentials as the space current of a three-electrode tube is of control-grid and plate potentials, and the effective amplification factor of the screen-grid tube, as far as the total space current is concerned, is μ_{sg}. The extent to which this is true is illustrated by the curves of Fig. 156, which show the relationship between total space current and control-grid voltage for a number of screen-grid potentials. These curves are seen to be identical in shape with the corresponding curves for the three-electrode tube that are given in Fig. 53. The dotted and solid curves of Fig. 156 are for widely different values of plate voltage and show that the plate voltage has a negligible effect on the total space current.

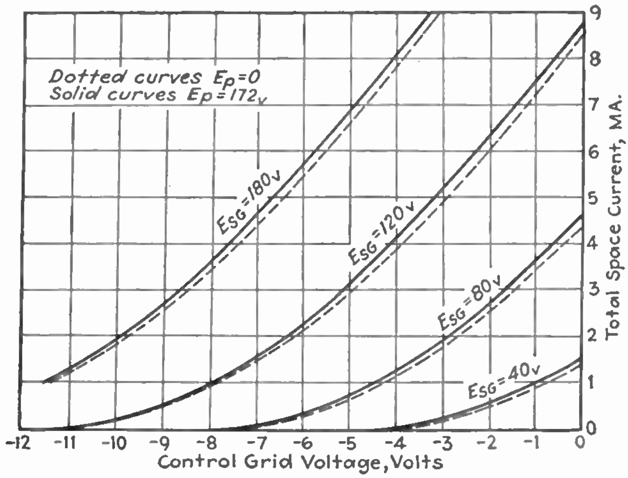


FIG. 156.—Typical curves showing the effect of electrode voltages on the total space current ($i_p + i_{sg}$) of a screen-grid tube. This total current is relatively independent of the plate potential and varies with control- and screen-grid potentials in exactly the same way as the plate current of a triode varies with grid and plate voltages, respectively.

Division of Total Space Current between Plate and Screen Grid.—The way in which the total space current divides between the screen-grid and plate electrodes is complicated by the fact that the electrons emitted from the cathode, *i.e.*, the primary electrons, produce secondary electrons in striking the screen grid and plate. These secondary electrons are then attracted to the electrode having the highest positive potential, with the result that there is an electron flow between screen grid and plate which is superimposed upon the flow of electrons to these electrodes from the cathode. The more positive electrode consequently receives a larger, and the less positive electrode a smaller, current than would be the case if there were no secondary emission, and this effect can be so great as to cause the less positive electrode to lose more electrons by secondary emission than it receives from the cathode. The effect of secondary electron emission in determining the distribution of current between the screen grid and plate is shown in Fig. 157, which shows the way in which the total space current divides between the plate and screen grid as the plate voltage is varied while maintaining the control-grid and screen-grid potentials constant. When the plate potential is less positive than the screen grid the latter receives secondary electrons from the former and so draws a large fraction of the total space current, whereas when the plate is more positive than the screen grid the opposite situation exists. It will be observed that for a range of plate voltages the plate current is negative. This occurs because for each primary electron reaching the plate more than one secondary electron is lost to the screen grid.

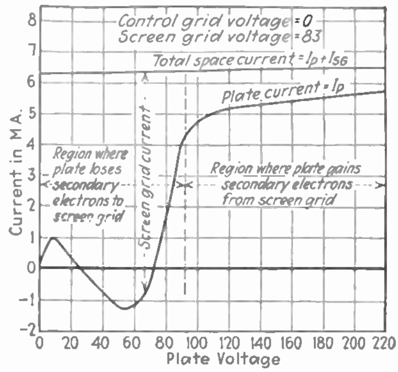


FIG. 157.—Curve showing the way in which the total space current divides between plate and screen grid as the plate voltage is varied. The total current is substantially independent of the plate voltage, but the secondary electrons cause the division of this total space current between the plate and screen grid to depend primarily upon the plate voltage.

Characteristic Curves of Screen-grid Tube.—The fundamental characteristics of screen-grid tubes are incorporated in the characteristic curves of Figs. 159 to 164, and while these curves appear rather complicated they depend upon a relatively small number of factors, as is made clear by the following discussion. The primary electrons travel in substantially straight lines away from the cathode and therefore divide between the screen grid and plate in proportion to the projected areas of the screen-grid and plate structures. The result is that for any particular tube each electrode will always receive the same fraction of the primary electrons irrespective of the electrode voltages or the magnitude of the total space current. The number of primary electrons received by an electrode is hence proportional to the total space current.

The number of secondary electrons that are produced at the screen grid, and at the plate, by the impinging primary electrons is proportional to the number of primary electrons received by the electrode, and hence to the total space current, and is also roughly proportional to the electrode potential. The flow of these secondary electrons is then determined by the screen-grid and plate potentials. When the screen grid is appreciably more positive than the plate practically all the secondary electrons produced at the plate will be drawn to the screen grid, whereas when the plate is more positive than the screen grid only a part of the secondary electrons produced at the screen grid are drawn to the plate. This is because most of the screen-grid secondary electrons are produced on the side of the screen grid that is away from the plate and so are partially shielded from the attraction of the plate by the screen grid. When the plate is the most positive electrode it is therefore able to attract only a portion of the secondary electrons from the screen grid, this proportion becoming larger as the potential of the plate with respect to the screen grid is increased. The result of these factors is that with given screen-grid and plate potentials the ratio which the current received by either electrode bears to the total space current is roughly constant irrespective of the magnitude of the total space current, but that the fraction of the total space current received in any particular case depends very greatly upon the relative (and also the absolute) electrode potentials. This is

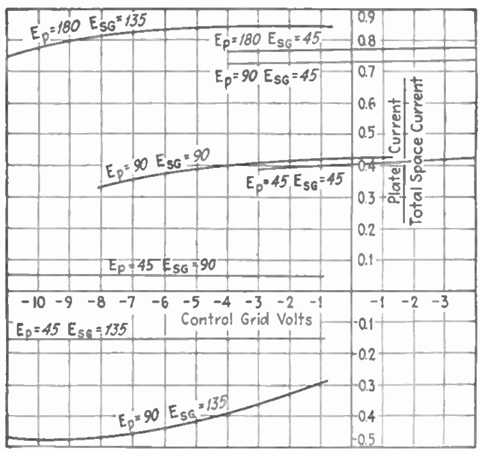


FIG. 158.—Fraction of total space current flowing to plate, as a function of control-grid voltage for various combinations of plate and screen-grid voltages. For given plate and screen-grid voltages the ratio is roughly independent of the control-grid voltage.

brought out by Fig. 158, which shows that to a first approximation the division of currents between the screen grid and plate is independent of the total space current when the potentials of these electrodes are kept constant, but that the division depends upon the value of the electrode voltages.

The way in which these principles actually work out in practice is shown by Figs. 159 to 164, which give the voltage and current relations existing in a typical screen-grid tube. Examination of these characteristic curves indicates that their principal properties may be summed up as follows:

1. The total space current, *i.e.*, $i_p + i_{sg}$, is substantially independent of plate voltage, and varies with screen-grid and control-grid potentials according to Eq. (127), which is the same way that the plate current of a triode varies with plate and grid voltages.

2. The way in which the total space current divides between screen grid and plate depends to a first approximation only on the potentials

of these electrodes, and is substantially independent of the control-grid potential. Varying the control-grid potential therefore causes the current received by the screen grid and by the plate to vary according to the same law that the plate current of a triode varies with grid voltage. Furthermore the only effect which changing the grid voltage has on Figs. 159 and 163 is to vary the magnitude without altering the shape.

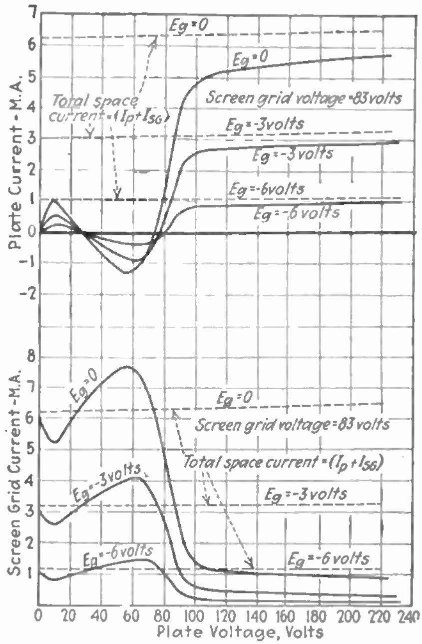


FIG. 159.

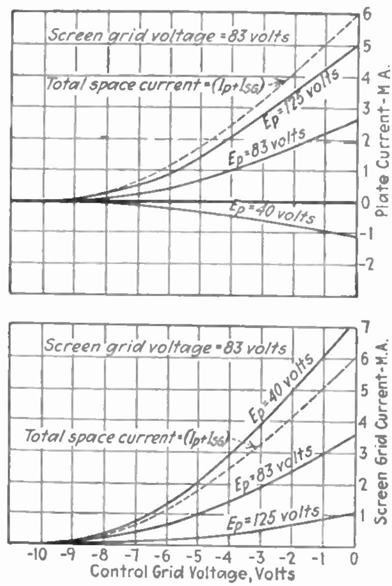


FIG. 160.

FIG. 159.—Variation of plate and screen-grid currents and of total space current, with plate voltage (screen-grid voltage constant). It will be noted that changing the control-grid voltage alters the magnitude of the curves without changing their shape (*i.e.*, the control-grid potential affects the total space current but does not alter its division between the plate and screen grid).

FIG. 160.—Variation of plate and screen-grid currents and total space current with control-grid voltage for several values of plate voltage (screen-grid voltage constant). These curves show in another form the same information that is contained in the curves of Fig. 159.

3. When the plate is appreciably more positive than the screen grid the plate receives secondary electrons from the screen grid, with the result that the net screen-grid current is very low (although seldom negative) and the plate current is large, usually only slightly less than the total space current.

4. When the screen grid is appreciably more positive than the plate, the screen grid receives secondary electrons from the plate, with the result that the net plate current is either very small or negative while the screen grid current is correspondingly large. When the plate potential is sufficiently high (in excess of 27 volts for the tube of Figs. 159 to 164)

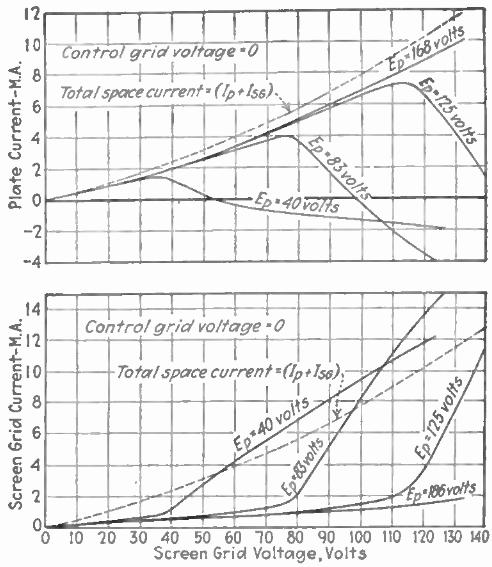


FIG. 161.—Variation of plate and screen-grid currents and total space current with screen-grid voltage for several values of plate voltage (control-grid voltage constant). The total space current is substantially independent of plate voltage, but the part of this current that goes to the plate is determined by the relative plate and screen-grid potentials. The sudden bend in the curves takes place where the potentials of these two electrodes are approximately the same and is caused by the reversal in direction of flow of secondary electrons which takes place under this condition.

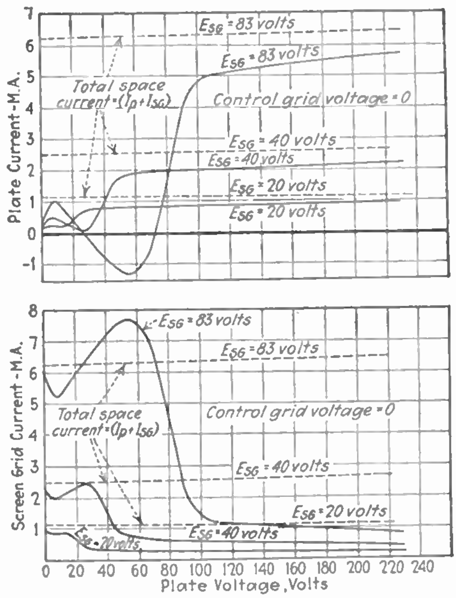


FIG. 162.—Variation of plate and screen-grid currents and total space current with plate voltage for several values of screen-grid voltage (control-grid voltage constant). The total space current is nearly independent of plate voltage, but the way in which this current divides between screen grid and plate depends upon the potentials of these electrodes, the one which is the most positive receiving the major fraction of the current.

each primary electron arriving at the plate produces on the average more than one secondary electron, so that with the screen grid more positive than the plate, the plate current under these conditions will be negative and the screen-grid current will exceed the total space current.

71. The Screen-grid Amplifier.—When used as an amplifier, the screen-grid tube is operated with the plate at a somewhat higher positive

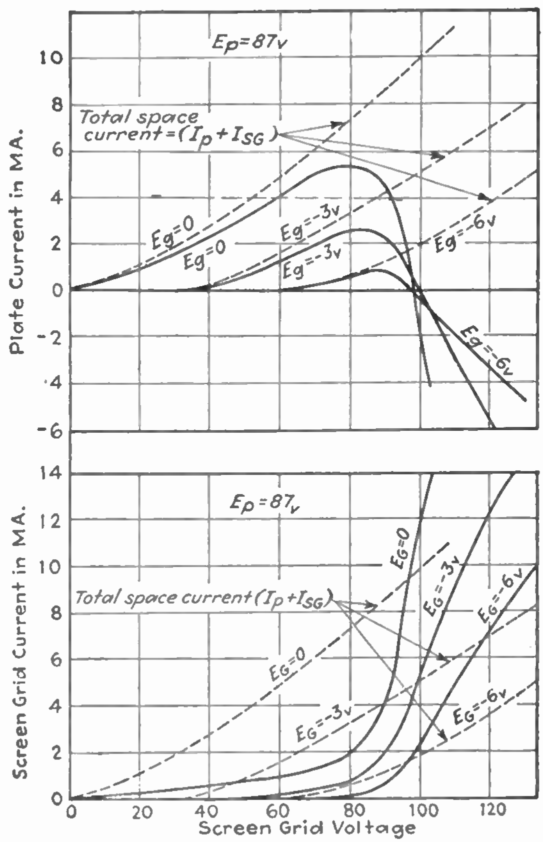


FIG. 163.—Variation of plate and screen-grid currents and total space current with screen-grid voltage for several values of control-grid voltage. The total space current varies with control-grid and screen-grid potentials in much the same way as the plate current in a triode varies with grid and plate voltages, and the major fraction of the total space current goes to the most positive electrode.

potential than the screen grid. The voltage to be amplified is applied to the control (inner) grid which is provided with a negative bias such that the instantaneous grid potential never becomes positive, while the load impedance to which the amplified output is delivered is in series with the plate of the tube. The actual circuit is arranged as shown in Fig. 165*a*.

The effect which a signal voltage applied to the control grid of a screen-grid tube has on the plate current is exactly the same effect that would be produced on the plate current by a somewhat larger voltage acting in a circuit consisting of the dynamic plate resistance of the tube in series with the load impedance. This leads to the equivalent circuit of the screen-grid amplifier shown in Fig. 165b, which is identical with the equivalent circuit of the triode amplifier shown in Fig. 60. The only

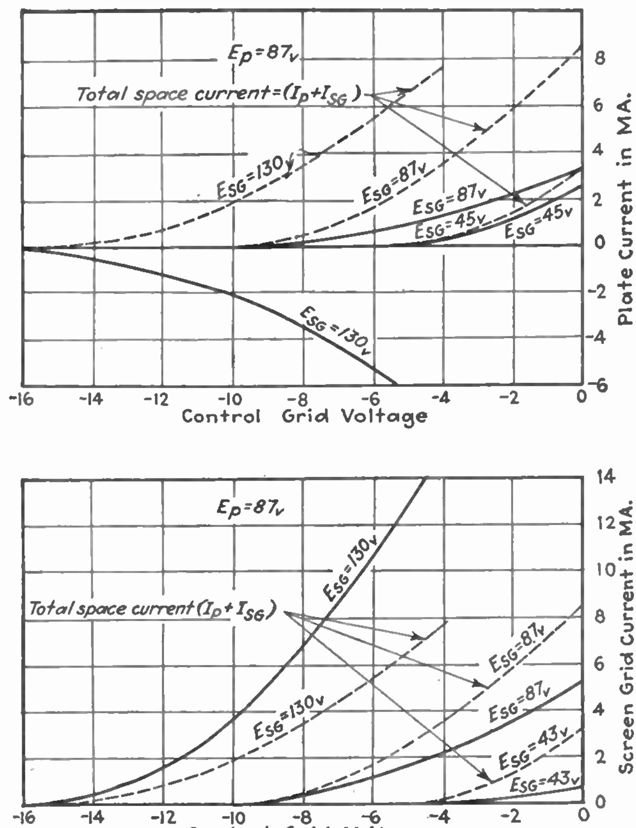


FIG. 164.—Variation of plate and screen-grid currents and total space current with control-grid voltage for several values of screen-grid voltage. These curves show in another way the same information as given in Fig. 163, and bring out how for a given plate and screen-grid voltage the current received by each of these electrodes varies with control-grid voltage in much the same way as the plate current of a triode varies with grid voltage.

difference between the equivalent plate circuits of screen-grid and triode amplifiers is in the numerical values of the plate resistance and amplification factor. The dynamic plate resistance of the screen-grid tube normally has a value in the order of hundreds of thousands of ohms, while the screen-grid amplification factor ordinarily exceeds 100 and depends very considerably upon the electrode voltages. The ratio of amplification factor to plate resistance of the screen-grid tube, that is the mutual conductance, is of the same order of magnitude as in ordinary triodes.

Plate Resistance.—The plate resistance of the screen-grid amplifier is defined by the relation

$$\text{Plate resistance} = R_p = \frac{\partial E_p}{\partial I_p}_{E_g \text{ and } E_{sg} \text{ constant}} \tag{128}$$

The numerical value of the plate resistance is high in screen-grid tubes because when the plate voltage exceeds the screen-grid potential the plate current changes very little with variations in the plate voltage, as is apparent from Figs. 159 and 162. The actual value depends upon the total space current, as in triodes, but is also affected to a considerable extent by the presence of secondary electron emission at the screen grid. Thus consider the increment of plate current that results from a given increment of plate voltage. If there were no secondary electron emission at the screen grid, the plate current would increase in proportion to the increase in total space current resulting from the plate-voltage increment, and would be small because the space current is very little affected by

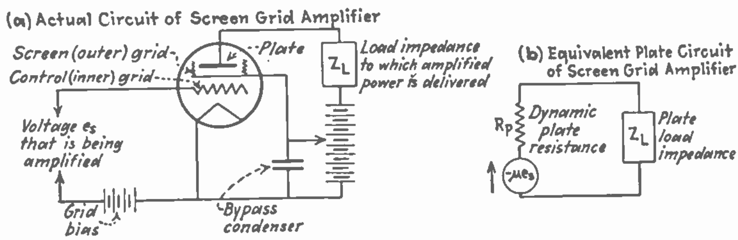


FIG. 165.—Circuit of screen-grid amplifier, together with equivalent plate circuit used in analyzing the behavior of the amplifier.

the plate voltage. In an actual tube secondary electrons are produced at the screen grid at a rate proportional to the total space current and depending upon the screen-grid potential. Adding an increment to the plate voltage increases the proportion of these secondary electrons that are attracted to the plate, and also increases the number of secondary electrons produced because the added plate voltage makes the total space current greater. Both of these factors increase the increment of plate current that results from an increment of plate voltage, and so act to lower the plate resistance below the value that would be obtained with no secondary electrons. The result is that the plate resistance becomes less as the total space current is increased by making the control grid less negative, as shown in Fig. 166. With constant total space current (*i.e.*, control- and screen-grid potentials constant) the plate resistance will be higher the more the plate potential exceeds the screen-grid voltage, as is apparent from Fig. 167. This is because a plate potential greatly in excess of the screen-grid voltage enables the plate to attract nearly all of the secondary electrons produced at the screen grid, and increments in plate voltage have very little effect in increasing the fraction of secondary electrons that can be drawn to the plate. Increasing the screen-grid voltage while maintaining both control-grid and plate potentials constant lowers the plate resistance, as shown in Fig. 168, because this increases the total space current and also reduces the amount by which the plate potential exceeds the screen-grid voltage, and both of these effects lower the plate resistance.

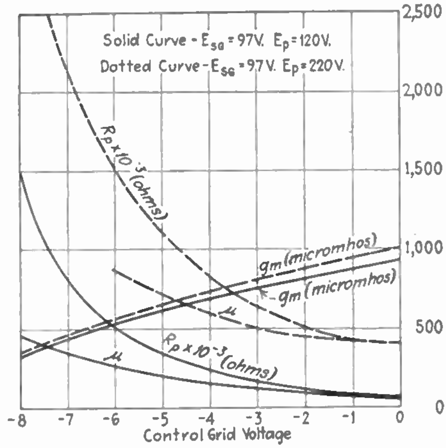


FIG. 166.—Variation of plate resistance, amplification factor, and mutual conductance with control-grid voltage in a typical screen-grid amplifier.

Amplification Factor.—The amplification factor of the screen-grid amplifier represents the relative effect which control-grid and plate potentials have on the plate current, *i.e.*,

$$\text{Amplification factor} = \mu = \left.\frac{\partial E_p}{\partial E_g}\right|_{I_p \text{ and } E_{sg} \text{ constant}} \tag{129}$$

The numerical values of amplification factor developed by screen-grid amplifiers are high because the plate voltage exerts very little influence on the plate current, whereas the control-grid potential has a very large effect. Unlike triodes the amplification factor of screen-grid amplifiers depends to a considerable extent upon the electrode potentials because the fraction of the secondary electrons produced at the screen grid that reaches the plate depends upon the relative potentials of the plate and screen grid. Hence an increase in plate voltage not only increases the total space current but also increases the fraction of this current which is attracted to the plate, whereas the control-grid potential only affects the total space current. The result is that the amplification factor of the screen-grid amplifier approaches the value μ_p (*i.e.*, the factor representing the relative effectiveness of grid and plate potentials in producing electrostatic field at the surface of the cathode) only when the screen-grid potential is so low as to prevent the production of appreciable secondary electrons at the screen grid, or when the plate potential exceeds the screen-grid voltage by such a large amount that practically all screen-grid secondary electrons are drawn to the plate. Under all other conditions (*i.e.*, where the fraction of the total space current that is received by the plate is affected by the relative plate and screen-grid voltages) the actual amplification factor will be less than μ_p by an amount depending upon the number of secondary electrons produced at the screen grid, and upon the relative electrode voltages. This is brought out by the curves in Figs. 166, 167, and 168, which show that the amplification factor varies considerably with different conditions and, in particular, falls off as the screen-grid potential approaches the plate voltage.

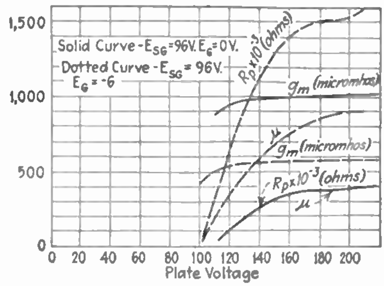


FIG. 167.—Variation of plate resistance, amplification factor, and mutual conductance with plate voltage in a typical screen-grid amplifier.

Mutual Conductance.—The ratio μ/R_p represents the mutual conductance (or as it is sometimes called, the transconductance) of the screen grid amplifier, which can also be defined by the relation

Mutual conductance =

$$g_m = \left.\frac{\partial I_p}{\partial E_g}\right|_{E_{sg} \text{ and } E_p \text{ constant}} \tag{130}$$

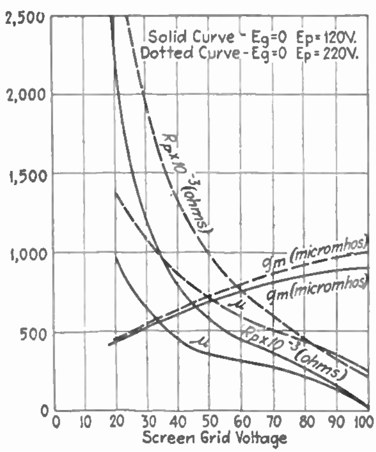


FIG. 168.—Variation of plate resistance, amplification factor, and mutual conductance with screen-grid voltage in a typical screen-grid amplifier.

Since the mutual conductance represents the rate of change of plate current with change of grid voltage when the plate and screen-grid potentials are maintained constant, the value of g_m is fixed primarily by the total space current and only secondarily by the relative plate and screen-grid potentials. This is because when the plate is more positive than the screen grid, the plate receives nearly all of the total space current, and the rate of change of this total space current with control-grid voltage increases as the total space current becomes greater, *i.e.*, as the quantity $\left(E_g + \frac{E_{sg}}{\mu_{sg}}\right)$ is increased. Increasing the ratio of plate to screen-grid voltage increases the mutual conductance somewhat by enabling a greater proportion of the screen-grid secondary electrons to reach the plate. The secondary electron emission at the screen grid materially increases the mutual conductance because the flow of these secondary electrons to the plate is equivalent to increasing the fraction of the total space current that goes to the plate. These various effects are shown in Figs. 166, 167 and 168, where it is seen that the mutual conductance increases with total space current (*i.e.*, with a higher screen-grid potential or a less negative grid) and with plate potentials that exceed the screen-grid voltage by increasing amounts.

Amplification Formulas.—The equivalent plate circuit of the screen-grid amplifier is handled in exactly the same manner as is the equivalent plate circuit of the triode amplifier. Thus the voltage that is developed across a plate load impedance Z_L by a signal voltage e_s applied to the grid is

$$\text{Amplified voltage developed across load impedance} = E = \mu e_s \frac{Z_L}{Z_L + R_p} \tag{131}$$

where

Z_L = load impedance in the plate circuit
R_p = dynamic plate resistance of the tube
μ = amplification factor of the tube
e_s = voltage applied to the control grid of the amplifier.

Since the plate resistance is usually very much greater than the load impedance it is convenient to rewrite Eq. (131) as follows:

$$\text{Voltage amplification} = \frac{E}{e_s} = \frac{\mu}{R_p}\frac{Z_L R_p}{R_p + Z_L} = g_m Z_{eq} \tag{132}$$

where

$g_m = \mu/R_p$ = mutual conductance of the tube
$Z_{eq} = \dfrac{Z_L R_p}{R_p + Z_L}$ = impedance formed by the plate resistance R_p in parallel with the load impedance Z_L.

Equation (132) *states that the ratio of voltage developed across the load impedance to the voltage applied to the grid of the tube,* i.e., *the tube amplification, is equal to the mutual conductance of the tube multiplied by the impedance of the parallel combination formed by the plate resistance and load in parallel.* Since the plate resistance is usually very much larger than the load impedance it is permissible when making rough calculations to assume that the amplification will approximate the product of mutual conductance and load impedance. It is therefore apparent that the mutual conductance is the fundamental tube constant of the screen-grid amplifier, and that the amplification factor and the plate resistance are of secondary importance.

The screen grid of the screen-grid amplifier should always be connected to the cathode through a by-pass condenser of sufficient capacity to have a low impedance to the frequency being amplified. This is necessary in order that the screen grid will be at ground potential and act as a grounded electrostatic shield in spite of any internal impedance that the source of screen-grid voltage may have.

72. Radio-frequency Amplification with Screen-grid Tubes.—The most important application of the screen-grid tube is in the amplification of radio-frequency voltages. Such amplifiers use a load impedance in which the essential element is a resonant circuit tuned to the frequency to be amplified and either directly connected in the plate circuit as shown at Fig. 169*a* or coupled by means of a transformer as shown at Fig. 169*b*. These two circuits, which are the ones most commonly employed in screen-grid radio-frequency amplifiers, are similar in form to the direct-coupled and transformer-coupled radio-frequency amplifier circuits used with triodes and shown in Fig. 91. The only special feature of the screen-grid amplifier is that the plate resistance of the tube is so high that it is seldom possible to obtain the optimum load impedance, *i.e.*, impedance equal in magnitude to the plate resistance of the tube. Since under these conditions the higher the load impedance the greater will be the amplification, the usual screen-grid radio-frequency amplifier employs either direct coupling, in which the full parallel-resonant impedance of the tuned circuit (shunted by the grid-leak resistance) forms the load impedance, or transformer coupling in which a high load impedance is obtained by employing a large primary (preferably having at least as many turns as the secondary), closely coupled to the secondary.

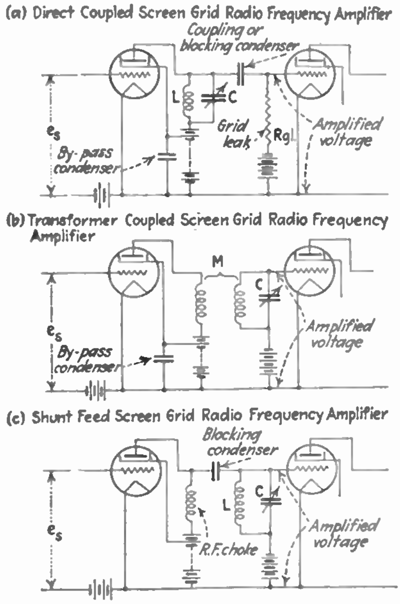


FIG. 169.—Circuits of typical tuned radio-frequency screen-grid amplifiers.

The amplification of a screen-grid tuned radio-frequency amplifier is most satisfactorily determined by first calculating the amplification at the resonant frequency of the tuned circuit and then obtaining the amplification at other frequencies from the effective Q of the amplifier.

In the case of direct coupling the amplification at resonance can be obtained directly from Eq. (131) or (132), using a load impedance Z_L consisting of the grid-leak resistance R_{gl} in parallel with the parallel-resonant resistance $\omega_o LQ = (\omega_o L)^2/R$ of the tuned circuit. The equation that results is

$$\text{Amplification at resonance with direct coupling} = g_m(Q\omega_o L)\frac{R_{eq}}{R_{eq} + Q\omega_o L} \tag{133}$$

where

g_m = mutual conductance of tube
R_{gl} = grid-leak resistance
$\omega_o L$ = inductive reactance of tuned circuit at resonant frequency
Q = $\omega_o L/R$ of tuned circuit at the resonant frequency.

The effective Q of the amplification curve is given by the relation

$$\frac{\text{Effective } Q \text{ of amplification curve}}{\text{Actual } Q \text{ of tuned circuit}} = \frac{R_{eq}}{R_{eq} + \omega_o LQ} \tag{134}$$

where R_{eq} is the equivalent resistance $(R_{gl}R_p)/(R_{gl} + R_p)$ formed by the grid leak in parallel with the plate resistance of the tube, and Q is the actual Q of the resonant circuit. With transformer coupling the amplification at resonance can be calculated from Eq. (84), which when applied to screen-grid tubes may be conveniently rearranged as:

$$\text{Amplification at resonance with transformer coupling} = g_m \frac{(\omega M)Q_s}{1 + \dfrac{(\omega M)^2/R_s}{R_p}} \tag{135}$$

where

g_m = mutual conductance of tube
M = mutual inductance between primary and secondary
Q_s = $\omega L/R$ of secondary circuit at resonance
R_s = series resistance of secondary circuit at resonance
R_p = plate resistance of the tube.

The effective Q of the amplification curve with transformer coupling depends upon the ratio which the coupled impedance $(\omega M)^2/R_s$ bears to the tube plate resistance R_p, and can be evaluated either from Eq. (89*a*) or from Fig. 94.

The screen-grid tube has a number of advantages over the three-electrode tube in the amplification of radio-frequency voltages. In the first place greater amplification is obtainable with screen-grid tubes because, while the mutual conductance is about the same as in triodes, the higher amplification factor of a screen-grid tube results in increased amplification, even though the full possibilities of this large value of μ are not fully realized. Secondly the selectivity of a screen-grid amplifier is only slightly less than the selectivity of the resonant circuit when taken alone because the plate resistance of a screen-grid tube is much higher than the load impedance. Finally, the shielding effect of the screen-grid electrode eliminates substantially all direct capacity between the grid and plate, avoiding the necessity of employing some neutralizing system, none of which are perfect in their operation. These considerations show that the screen-grid radio-frequency amplifier is more selective, develops more amplification, and gives less trouble from low input impedance than does the corresponding triode amplifier, and for these reasons is generally preferred to the latter.

Input Impedance of Screen-grid Amplifiers.—The input impedance of a screen-grid amplifier, while high, is not infinite, since there is always some residual direct capacity between the plate and grid circuits. In practical screen-grid tubes this capacity is kept low by extending the screen grid to cover both sides of the plate and by bringing the plate and control-grid connections out through opposite ends of the tube, as shown in Fig. 154, with the result that in receiving tubes the residual grid-plate capacity is in the order of 0.01 to 0.02 $\mu\mu$f. Although this capacity is small it does permit appreciable regeneration at high frequencies because the tube amplification of the screen-grid amplifier is usually very large.

The minimum value of input resistance that can be obtained with a resonant load impedance is equal to the reactance of the residual grid-plate capacity divided by half the amplification computed from Eq. (132), as is explained in Sec. 43 and stated in Eq. (93a). When reasonable numerical values are substituted in this equation, it will be found that an input resistance as low as 100,000 ohms can be obtained at frequencies in the order of 1,000,000 cycles, and even lower values will be obtained as the frequency is increased. Thus, when the amplification as computed from Eq. (132) is one hundred, and the residual grid-plate tube capacity is 0.02 $\mu\mu$f, the minimum input resistance will be $(1/0.02 \times 10^{-12} \times 2\pi \times 10^6)/(100/2) = 159{,}000$ ohms at 1000 kc. At frequencies slightly below resonance this will be a negative resistance and will produce oscillations when shunted across a tuned input circuit having a parallel-resonant impedance greater than 159,000 ohms. In order to control this residual feed back that takes place in screen-grid tubes it is necessary either to use circuits with considerable losses or to insert a resistance in series with the grid lead for the purpose of compensating for the feed back, as explained in Sec. 44.

73. Audio-frequency Amplification with Screen-grid Tubes.—While the principal use of screen-grid tubes is in the amplification of radio frequencies, such tubes may also be used to amplify audio frequencies, using resistance, impedance, or transformer coupling. The circuits for these various types of amplification are shown in Fig. 170 and are seen to be similar to those employed with three-electrode tubes.

The high amplification factor of screen-grid tubes makes it theoretically possible to obtain enormous amplifications with resistance coupling, but such large amplifications are not realizable in practice because a coupling resistance in the order of megohms would be required, and the voltage consumed by the direct-current plate current flowing through such a resistance would be several thousand volts. With reasonable values of plate-supply voltage it is possible, however, to obtain amplifications that are at least equal to, and usually more than, those given by resistance-coupled amplifiers employing three-electrode tubes. The factors that determine the design of screen-grid resistance-

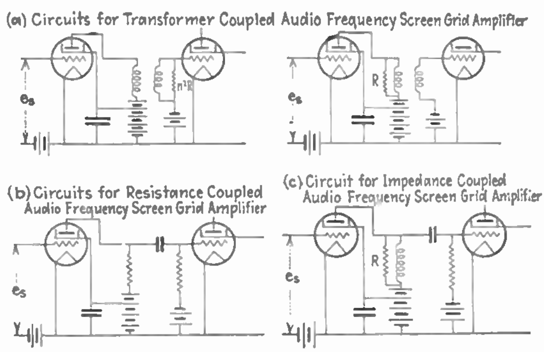


FIG. 170.—Circuits for transformer-, resistance-, and impedance-coupled audio-frequency screen-grid amplification.

coupled amplifiers are in general the same as those discussed in Sec. 35 and need not be repeated. The only difference is that the voltage drop in the coupling resistance must be somewhat less than the amount by which the plate-supply voltage exceeds the screen-grid potential, for otherwise the plate would have a lower potential than the screen grid. This requirement means that the coupling resistance that can be employed with a fixed plate-supply voltage can be made large only by lowering the screen-grid potential, and this reduces the mutual conductance of the tube and partially counteracts the increased amplification that would otherwise be obtained from the higher resistance. It is found that the most satisfactory amplification is obtained with a relatively low screen-grid potential combined with the highest plate-supply voltage that is available, and a coupling resistance of at least several hundred thousand ohms.

The amplification obtainable from a screen-grid resistance-coupled amplifier can be determined by substituting in Eq. (132), and is found to be

$$\left.\begin{array}{l}\text{Maximum amplification of resistance-}\\ \text{coupled screen-grid amplifier}\end{array}\right\} = g_m R_{eq} \qquad (136)$$

where g_m is the mutual conductance of the tube at the plate, screen-grid, and control-grid voltages employed, and R_{eq} is the equivalent resistance formed by the plate resistance, the coupling resistance, and the grid-leak resistance, all in parallel. In the usual case R_{eq} does not differ greatly from the coupling resistance, so that the amplification is roughly proportional to the product of mutual conductance and coupling resistance. The amplification given by Eq. (136) represents the maximum amplification obtainable from the amplifier and occurs at moderate frequencies, as in triode resistance-coupled amplifiers. At low frequencies the amplification falls off as a result of the voltage drop in the coupling condenser, exactly as described in Sec. 35, and it also falls off at high frequencies because the various capacities that shunt the coupling resistance lower the equivalent load impedance. The extent of the loss of amplification at high frequencies depends upon the ratio of the reactance of the shunting condenser to the equivalent resistance formed by the plate, coupling, and grid-leak resistances in parallel, and can be evaluated in terms of this ratio by means of Fig. 63.

In the transformer-coupled audio-frequency screen-grid amplifier it is necessary to place a resistance across either the primary or the secondary of the transformer if excessive frequency distortion is to be avoided. Without such a resistance the amplification is extremely great at the frequency for which the primary inductance of the transformer is in parallel resonance with the shunting capacity, and is relatively much less at other frequencies. The shunting resistance avoids this frequency distortion by making the impedance across the terminals of the transformer substantially constant over a wide range of frequencies. *The proper value of resistance to connect across the primary terminals of the transformer can be determined from the fact that the frequency distortion in a screen-grid transformer-coupled amplifier is exactly the same as the frequency distortion that would be present in a triode amplifier employing the same transformer and having a plate resistance equal to the equivalence resistance formed by the transformer shunting resistance in parallel with the screen-grid plate resistance.*[1] *Since the plate resistance of the screen-grid tube is very large*

[1] This is shown by the following demonstration: In the triode amplifier having a plate resistance R_p, amplification factor μ, and mutual conductance g_m, the voltage developed across the primary terminals of a transformer having a primary impedance Z_L is

$$\text{Amplified voltage of triode} = \frac{\mu Z_L}{R_p + Z_L} e_s = g_m \frac{R_p Z_L}{R_p + Z_L} e_s$$

In the case of the screen-grid amplifier having a plate resistance R_p' and a mutual conductance g_m' the voltage developed across the primary of a transformer having a primary impedance Z_L and shunted by a resistance R is found by substituting in Eq. (132) and has the value

$$\text{Amplified voltage of screen-grid tube} = g_m Z_{eq} e_s = g_m \frac{Z_L R_{eq}}{Z_L + R_{eq}} e_s$$

this means that the resistance shunted across the primary should approximate the plate resistance that would be used with the same transformer in a triode amplifier. When the shunting resistance is placed across the secondary terminals of the transformer the value should be the square of the step-up ratio times the proper resistance for use across the primary terminals.

In the impedance-coupled audio-frequency screen-grid amplifier it is necessary to shunt the coupling inductance with a suitable resistance in order to prevent excessive frequency distortion. The reason for this is exactly the same as in the case of transformer coupling, and the value of resistance is determined by the same considerations, *i.e.*, the frequency distortion of an impedance-coupled screen-grid amplifier is exactly the same as the frequency distortion that results when the same coupling inductance is employed in a triode amplifier having a plate resistance equal to the resistance of the shunting resistance and the screen-grid plate resistance in parallel.

Merits of Screen-grid Audio-frequency Amplifier.—Audio-frequency amplification with screen-grid tubes is at least as satisfactory as audio-frequency amplification obtained with triodes but has the disadvantage of requiring a more complicated and hence more expensive tube, and a tube requiring an extra supply voltage for the added electrode. The screen-grid resistance-coupled amplifier using reasonable plate voltages will give slightly more amplification than can be obtained with resistance-coupled triode amplifiers, and if an exceptionally high plate-voltage source is available the screen-grid amplification will be very much higher than when a triode is used. A comparison of the amplification obtainable with screen-grid and triode tubes using the same transformer or same coupling inductance shows that when the frequency distortion in the two cases is identical the voltage amplifications developed by the two types of tubes are exactly in the ratio of the mutual conductances of the tubes. Since screen-grid tubes at normal plate voltages have mutual conductances at least equal to and often higher than those of small triodes, the screen-grid tube will ordinarily give at least as much amplification as the corresponding triode.

When several stages of amplification are involved screen-grid tubes will give a somewhat better performance at the high audio frequencies than do triodes because the input capacity of a screen-grid tube is not increased by the load impedance in the plate circuit as is that of three-

---

where Z_{eq} is the equivalent impedance formed by Z_L, R, and R_p' all in parallel, and R_{eq} is the equivalent resistance formed by the screen-grid plate resistance R_p' in parallel with the resistance R that is shunted across the transformer primary. If R_{eq} is equal to the plate resistance R_p of the triode amplifier then the amplification expressions for the screen-grid and triode amplification are identical, and the frequency distortion will be the same in the two cases, while the amplifications obtained will be in the ratio of the mutual conductances.

electrode tubes. The result is that a stage of amplification which delivers its output voltage to a screen-grid tube will give better response at the high audio frequencies than will the same amplifier when it delivers its output to a triode.

When compared with the triode audio-frequency amplifier the screen-grid audio-frequency amplifier is seen to have some advantages and some disadvantages. This has resulted in the screen-grid tube finding a

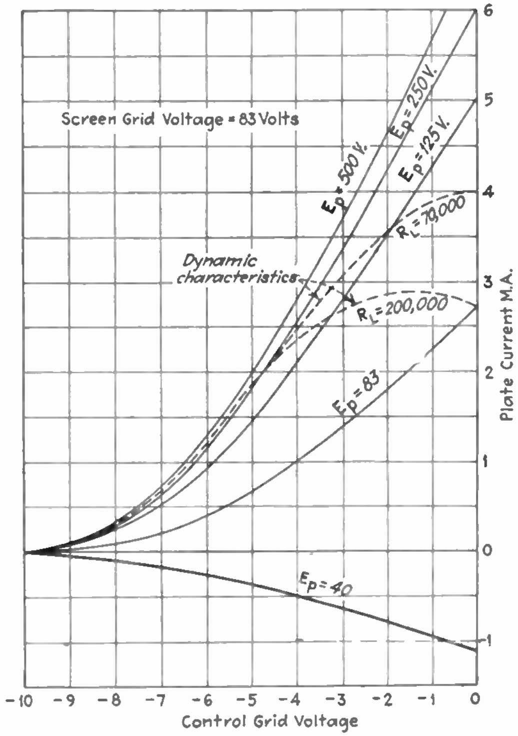


FIG. 171.—Dynamic characteristic of screen-grid amplifier for two values of load resistance. The useful operating range is in the region where the instantaneous plate voltage is more positive than the screen grid, and the dynamic characteristic is far from straight even over this range.

limited field of usefulness as an audio-frequency amplifier, but it shows no tendency to displace the triode tube under ordinary circumstances.

74. Power Amplification with Screen-grid Tubes.—The screen-grid tube is inherently unsuitable for the development of large quantities of undistorted power output by reason of the fact that the dynamic characteristic of the screen-grid amplifier is not only relatively curved but also makes an abrupt bend at the point where the plate voltage equals the screen-grid potential. The dynamic characteristic of a typical screen-grid amplifier is shown in Fig. 171 and is relatively independent of load resistance as long as the plate voltage is appreciably greater than the screen-grid voltage, because the plate current is then largely independent of plate voltage. As soon as the instantaneous plate voltage of the dynamic characteristic begins to approach the screen-grid potential, however, there is a sharp bend in the characteristic. In order to approximate distortionless amplification it is accordingly necessary to restrict the operating range to the region where the instantaneous plate potential always exceeds the screen-grid voltage (*i.e.*, amplified voltage developed across load impedance must not exceed the difference between plate-supply voltage and screen-grid potential), which makes the power output low, and even then the curvature of the dynamic characteristic is such that the distortion is considerably greater than in power amplifiers using triodes.

In linear (*i.e.*, Class B) and high efficiency (*i.e.*, Class C) amplifiers amplitude distortion is not important because the tuned load impedance rejects the harmonics. Screen-grid tubes are consequently widely used in amplifiers of these classes when very high frequencies are to be handled, because the screen grid eliminates energy transfer between plate and grid circuits through the tube capacity much more satisfactorily than any neutralizing scheme.

The linear screen-grid amplifier is operated with a grid bias that approximates cut-off (*i.e.*, a negative grid bias voltage equal to E_{sg}/μ_{sg}) and with a plate load impedance such that the voltage developed across the load impedance by the application to the grid of an alternating voltage having a crest value equal to the grid bias is slightly less than the difference between the plate-supply voltage and the screen-grid potential. Under these conditions the plate current flows in impulses that approximate half sine waves, and the entire situation is exactly the same as in the case of the linear triode amplifier. The load impedance of the screen-grid linear amplifier must be such that the voltage developed across the load is less than the difference between the direct-current voltages applied to screen grid and plate. Otherwise when the instantaneous plate potential becomes less positive than the screen-grid the plate current wave is no longer a half sine wave, and the linear relation between input and output voltages is destroyed. The required value of load impedance can be readily obtained by determining the resistance that will have to be inserted in the plate circuit to make the direct-current plate and screen-grid potentials equal when the grid bias on the tube is zero and there is no grid excitation voltage, and then employing a resonant circuit to supply a plate load impedance that has twice this value.

The Class C screen-grid amplifier is similar to the screen-grid linear amplifier except that the grid bias is greater than the cut-off value, *i.e.*, $E_c > E_{sg}/\mu_{sg}$, and the load impedance used in the plate circuit is greater than that employed in linear amplification, with the exact value depending upon the fraction of the cycle during which the plate current flows. The grid of the Class C amplifier is ordinarily allowed to go somewhat positive at the positive crest of the applied voltage wave but should never become more positive than the screen grid.

Screen-grid power amplifiers are made in sizes up to about one kilowatt rated output. The construction of one of these large tubes is shown in Fig. 172 and is seen to differ from the construction of the small screen-grid tube of Fig. 154 in that the plate is not entirely surrounded by the screen grid. The high voltages employed in power tubes would prohibit such a method of construction, but since the screen grid is very long it

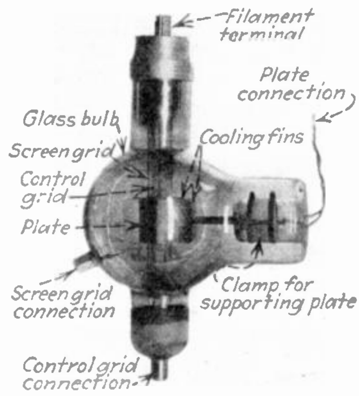


FIG. 172.—Photograph of screen-grid power tube having a rating of 500 watts output when operating as a Class B amplifier.

effectively shields the plate and control grid from each other and keeps the grid-plate capacity to a very low value. The large screen-grid power tubes usually are merely ordinary power tubes with the addition of a screen grid and perhaps with a modified mesh in the control-grid structure. The amplification factor μ_{sg} of the control grid with respect to the screen grid is made relatively low in screen-grid power tubes in order that a high mutual conductance may be obtained with a low screen-grid potential. The screen grid is operated at a relatively low potential, while the plate-supply voltage is very high, thus permitting a large voltage to be developed across the load impedance.

75. Space-charge Grid Tubes.—The space-charge grid tube is essentially a three-electrode vacuum tube that is provided with an auxiliary (or space-charge) grid located between the cathode and control grid and operated at a moderate positive potential, as shown in Fig. 173. The effect of the space-charge grid is to increase the number of electrons drawn out of the space charge near the cathode. Some of these electrons are immediately attracted by the space-charge grid, but many of them pass through the meshes into the space near the control grid, where they are slowed down by a retarding field and come to rest near the control grid where a second space charge is formed. This latter space charge forms a virtual cathode as shown in Fig. 173 and from which electrons are drawn to the plate as a result of the combined action of the control-grid and plate potentials. The amplification factor and plate resistance of a space-charge grid tube are determined by considering the virtual cathode as an actual cathode, and because this virtual cathode has a very large cross-sectional area and is located extremely close to the control grid, the plate resistance of a space-charge grid tube is considerably lower than the plate resistance of a comparable triode having the same amplification factor.

The plate current in the space-charge grid tube varies with control-grid and plate potentials in almost exactly the same way as in the three-

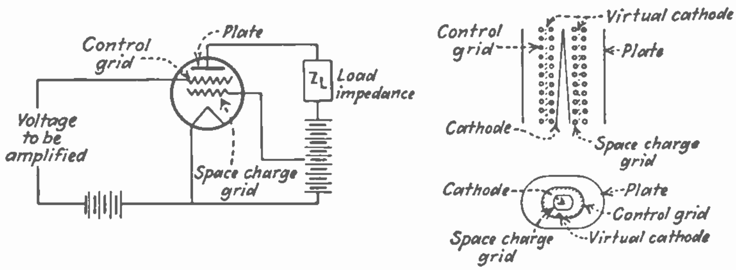


Fig. 173.—Circuit diagram and constructional details of space-charge grid tube. The inner grid is the space-charge grid and is operated at a moderate positive potential.

electrode tube. This is clearly brought out in Fig. 174, which gives characteristic curves of a typical space-charge grid tube. The space-charge grid current is usually somewhat greater than the plate current and represents an addition to the space current that must be supplied by the electron emission from the cathode. The magnitude of the space-charge grid current increases rapidly with space-charge grid potential, and varies with control-grid and plate voltages in such a way that the sum of the space-charge grid and plate currents is determined almost solely by the potential of the space-charge grid and is substantially independent of control-grid and plate voltages. The amplification factor of space-charge grid tubes is determined primarily by the grid construction but varies somewhat with electrode voltages. The plate resistance depends upon the control-grid and plate potentials, as in an ordinary triode, and in addition decreases as the space-charge grid is made more positive.

The ordinary screen-grid tube can be used as a space-charge grid tube by utilizing the inner grid as the space-charge grid and the outer (or shielding) grid as the control grid, but such a tube seldom has characteristics as desirable as those of a tube specially designed for space-charge grid purposes. This is because the screen grid to plate capacity of a screen-grid tube is higher than necessary for space-charge grid applications as a result of the fact that the shielding grid completely encloses the plate. The amplification factor of the usual screen-grid tube when used in the space-charge connection is also too high for most purposes.

The space-charge grid tube is fundamentally an amplifier capable of giving very high amplification per stage while being limited to a low power output. The amplification is great because of the low plate resistance in comparison with the amplification factor, while the power capacity is small because the characteristic curves of the space-charge grid tube have a greater curvature than the corresponding triode characteristics and so cause amplitude distortion to be introduced if large signal amplitudes are applied to the tube. The space-charge grid tube is capable of giving excellent amplification with anode voltages as low as 6 to 10 volts. It has been used extensively in Europe for some years but has not as yet been employed in commercial equipment in this country.

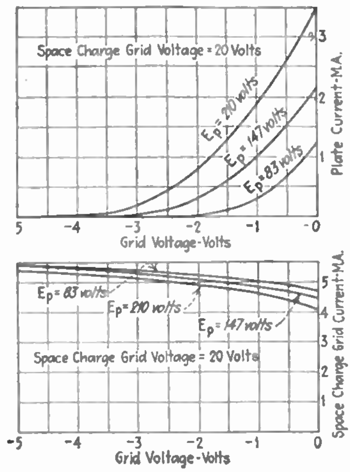


FIG. 174.—Characteristic curves of space-charge grid tube. The plate current varies with control-grid and plate voltages in much the same way as in a triode, while the space-charge grid current is much larger than the plate current and decreases as the plate current increases.

76. Co-planar Grid Tubes.—The co-planar grid tube can be thought of as an ordinary triode to which there has been added a second grid wound between the meshes of the usual grid. Such a double grid tube obviously introduces a number of possibilities not present in simpler tubes, and, while the co-planar grid tube is still in the experimental state, it shows promise as a grid-leak power detector and as a power tube.

When used as a grid-leak power detector the two grids of the co-planar tube are connected to the opposite ends of the coil across which the radio-frequency voltage to be rectified is developed. The cathode of the tube is then connected to the midpoint of this coil through a grid-leak grid-condenser combination. The input circuit of such a tube is similar in all respects to the center-tapped full-wave rectifier circuit used in supplying the anode power for radio receivers, with the grid-leak grid-condenser combination serving as the load impedance to which the rectified current is delivered. The voltage developed across the grid-leak and grid-condenser by the rectified current is applied in the same phase to the two grids, which then operate together to control the plate current. Compared with the ordinary grid-leak power detector using triode tubes, the co-planar grid rectifier has several advantages, of which the most important is that there is no radio-frequency current flowing in the plate circuit because the radio-frequency voltage is applied to the two grids in opposite phase and so balances out as far as the plate circuit is concerned. This eliminates the factor that fixes the load limit in the ordinary power grid detector, and makes the co-planar tube capable of delivering approximately twice as much output voltage without overloading as can the comparable triode detector.

In the power type of co-planar grid tube one of the co-planar grids is operated at a relatively high positive voltage while the other grid serves

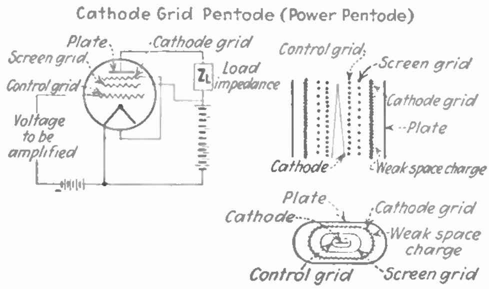


FIG. 175.—Circuit diagram and electrode arrangement of screen-grid (or power) pentode. This tube is a screen-grid tube with the addition of a grid located between screen grid and plate and connected to the cathode.

as the ordinary control grid.[1] The positive grid creates a strong electrostatic field at the surface of the cathode and hence causes a large space current to be drawn, but the magnitude of this current can be controlled by the potential of the second grid. The result is that the co-planar power tube has characteristics which, as far as the plate current is concerned, are very similar to the screen-grid tube, but with the difference that there is very little tendency for secondary electrons produced at the plate to be drawn to the positive grid even when this grid is much more positive than the plate. This is because of the shielding action of the negatively biased control grid. The result is a performance somewhat the same as that of the pentode as far as power amplification is concerned.

77. Pentodes.—Of the many possible five-electrode tubes the only one that is used to an appreciable extent is the cathode-grid (or suppressor-grid) pentode. This is essentially a screen-grid tube to which there has been added an auxiliary grid that is located between plate and screen grid and is permanently connected to the mid-point of the filament, as shown in Fig. 175. This additional grid, called the cathode or suppressor grid, prevents the secondary electrons produced at the plate from being able to reach the screen grid when the plate potential is less than

[1] See H. A. Pidgeon and J. O. McNally, A Study of the Output Power Obtained from Vacuum Tubes of Different Types, *Proc. I.R.E.*, vol. 18, p. 266, February, 1930.

the screen-grid voltage, while still permitting high-velocity primary electrons to reach the plate. The cathode grid thus eliminates the most important factor that prevents the ordinary screen-grid tube from being able to deliver large power outputs, and is what might be called a power type of screen-grid tube in which the secondary electrons produced at the plate cannot cause distortion.

The characteristics of the cathode-grid pentode are similar in many respects to those of the screen-grid tube. Thus the plate current is

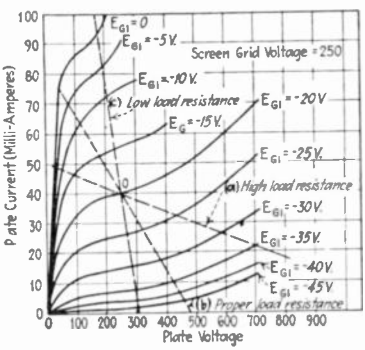


Fig. 176.

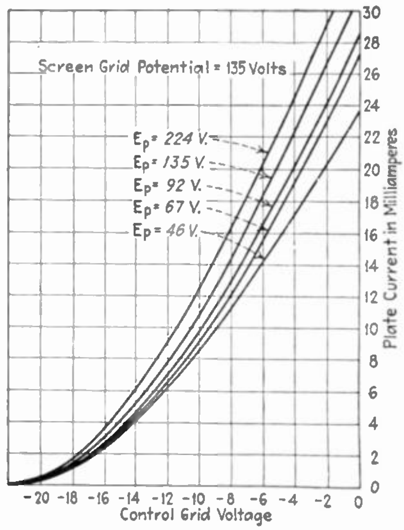


Fig. 177.

Fig. 176.—Characteristic pentode curves showing the relation between plate voltage and plate current for various constant values of control-grid voltage. Note that the plate current is relatively independent of plate voltage except at small plate voltages.

Fig. 177.—Characteristic curves of a typical power pentode. Comparison with the corresponding curves of a screen-grid tube shown in Fig. 171 show that the power pentode differs mainly in that its plate current does not suddenly drop and become negative when the plate voltage becomes less than the screen-grid potential.

determined primarily by the control-grid and screen-grid potentials and is relatively independent of the plate voltage provided the plate potential is not so low as to permit the formation of an appreciable space charge in the vicinity of the cathode grid. The plate resistance and amplification factor are consequently high, while the mutual conductance has a normal value. Characteristic curves of typical pentodes are given in Figs. 176 and 177.

When the pentode is used to develop large quantities of undistorted output considerable care must be taken in choosing the load impedance. If this impedance is very high, as shown by line *a* of Fig. 176, the minimum plate potential that is reached becomes low enough to permit the formation of a space charge in the vicinity of the cathode grid. This prevents the plate current from varying with the control-grid potential and so introduces distortion. On the other hand if the load impedance is too small the possible output drops off much more than is the case with a triode amplifier operating with a low load impedance.

In order to avoid distortion the crest alternating-current voltage developed across the load should not exceed 80 to 85 per cent of the plate-supply potential, and the load impedance should be such that a crest alternating-current current approximately 80 to 85 per cent of the direct-current plate current will develop this potential in flowing through the load impedance. As a result, the proper load impedance is very nearly equal to the ratio of plate-supply voltage to direct-current plate current at the operating point. The plate efficiency with this value of load impedance is in the order of 30 to 35 per cent, which is to say, the maximum possible undistorted power output obtainable is approximately one-third of the direct-current plate power supplied the tube.[1]

When compared with the triode power tube the cathode-grid pentode type of power amplifier has the advantage of giving somewhat more undistorted output for the same total space current and for the same plate potential. The commercial forms of this tube also require only about one-third as much signal voltage to develop a given output as does the ordinary triode tube and are so designed that the screen-grid and plate potentials are the same, thus simplifying the power supply problem. At the same time the power pentode is more expensive and has a somewhat greater harmonic output even under the most favorable conditions and in general requires more careful control of the operating conditions in order to obtain a satisfactory performance.

The cathode-grid type of pentode has certain advantages when used as a voltage amplifier, particularly when used as a radio-frequency voltage amplifier. Compared with the ordinary screen-grid tube, the pentode gives even more effective shielding between the input and output circuits and furthermore can be operated with both plate and screen grid at the same potential. This is often of importance since it permits the use of a low plate potential, whereas the screen grid tube requires a plate potential of about twice the screen-grid voltage if the secondary electrons produced at the screen grid are not to cause a serious reduction in the mutual conductance.

78. Miscellaneous Uses of Screen-grid, Space-charge Grid, Co-planar Grid, and Five-electrode Tubes.—While these tube types have been developed primarily for amplification purposes, they can also be used as plate rectifiers, since the relationship between control-grid voltage and

[1] This results from the fact that the direct-current plate power is E_bI_b while the alternating current in the load is in the order of $0.85I_b$ and the crest alternating-current voltage across the load approximately $0.85E_b$. The alternating-current power is hence roughly $0.85^2E_bI_b/2$.

plate current for these tubes is similar to the corresponding triode characteristic. The detector action is therefore similar in character to that which takes place in triodes and can be analyzed in exactly the same way. The efficiency of rectification is about the same, and the relative merits of the different types of tubes from the point of view of plate rectification are essentially the relative merits of the different tubes from the point of view of amplifier action.

Experiments show that the relationship between control-grid voltage and control-grid current in a vacuum tube is not affected by the number of electrodes, so that screen-grid, space-charge grid, and co-planar grid tubes, as well as pentodes, give exactly the same grid rectification as do triodes. These four- and five-electrode tubes can therefore be used for grid rectification of either weak or strong signals, their merits as grid rectifiers compared with triode-grid rectification being exactly the relative merits of the tubes considered from the point of view of audio-frequency amplification.[1]

Since the four- and five-electrode tubes that have been discussed are all able to function as amplifiers they can also serve as oscillator tubes. The only situation where oscillators with four- and five-electrode tubes are preferred to oscillators employing the simpler and less expensive triode is where it is desirable to avoid energy transfer between the input and output circuits of the tube through the grid-plate tube capacity. In triode tubes the effective grid-plate tube capacity depends on the alternating voltage acting between the plate and cathode and is much larger than the geometrical tube capacity, as explained in Sec. 43. The result is that the frequency of oscillation is affected by the adjustment in the plate circuit, and also the highest frequency that can be obtained is lower than if this high grid-cathode capacity were not present. Since screen-grid tubes have an input capacity that does not depend on the alternating voltage in the plate circuit and which is no greater than the geometrical tube capacity, such tubes can sometimes be used to advantage in oscillators that must meet special requirements. Screen-grid tubes are also useful in oscillators which obtain a high frequency stability by passing the amplified output of the oscillator tube through a number of tuned circuits before being applied to the grid of the tube for excitation purposes. In such oscillators the direct transfer of energy through the tube capacity has the effect of by-passing the tuned circuits and thus lowers the frequency stability.[2]

[1] For example, see F. E. Terman and Birney Dysart, Detection Characteristics of Screen-grid and Space-charge Grid Tubes, *Proc. I.R.E.*, vol. 17, p. 830, May, 1929.

[2] For a description of such an oscillator see: Ross Gunn, A New Frequency Stabilized Oscillator System, *Proc. I.R.E.*, vol. 18, p. 1560, September, 1930. The Gunn oscillator is essentially two stages of tuned radio-frequency amplification with the output of the second stage supplying the excitation for the first stage.

In addition to their use in oscillators, detectors, and amplifiers, tubes with four and five electrodes can be employed as multiple-function tubes which simultaneously perform two or more operations. Thus it is possible to devise circuits by which one tube can serve simultaneously as a one-stage radio-frequency amplifier, a detector, and a one-stage audio-frequency amplifier. The number of such combinations possible is almost unlimited, but in practically every case the same operations can be performed more satisfactorily by using several tubes of simpler construction. The result is that these multiple-function uses of tetrodes and pentodes have not met with general favor.

79. The Dynatron.—The dynatron is a vacuum tube in which secondary electron emission is used to give a negative resistance characteristic. The usual dynatron consists of a screen-grid tube operated with a plate voltage that is appreciably less than the screen-grid potential. Under these conditions there is an appreciable range of plate voltage in which the plate current decreases with increasing plate voltage, as can be seen from Fig. 159. A characteristic of this type represents a negative dynamic resistance since a positive increment (*i.e.*, an increase) in plate voltage causes a negative increment (*i.e.*, a decrease) of plate current, and *vice versa*.

The negative resistance characteristic of the dynatron results from the secondary emission of electrons at the plate. The number of primary electrons (*i.e.*, electrons emitted from the cathode) which arrive at the plate of a screen-grid tube is determined primarily by the screen- and control-grid potentials and the projected area of the screen grid, and is substantially independent of the plate potential. The number of secondary electrons that are produced at the plate however increases as the plate potential is made greater because a higher plate potential increases the velocity with which the primary electrons strike the plate. Since all of the secondary electrons produced at the plate are drawn to the more positive screen grid, the net plate current (*i.e.*, the difference between the primary electrons received from the cathode and the secondary electrons lost to the screen grid) decreases as the plate becomes more positive, with the result that the plate circuit possesses a negative dynamic resistance.

The magnitude of the negative resistance of the dynatron plate circuit varies inversely with the number of primary electrons which strike the plate, and this in turn is determined by the control- and screen-grid potentials. The control grid forms a convenient means of controlling the negative resistance, since by maintaining this electrode at a negative potential with respect to the cathode the control can be obtained without the consumption of energy. The effect of different control-grid voltages on the negative resistance can be seen from the plate-current curves of Fig. 159 and is to change the slope of the negative resistance part of the characteristic without otherwise altering the general shape. The curves for different control-grid voltages all pass through zero plate current at the same plate voltage. The plate voltage at which this occurs is the potential causing the primary electrons to strike the plate with such a velocity that on the average each primary electron dislodges one secondary electron. Under these conditions the net plate current is zero irrespective of the control-grid potential, which merely varies the number of primary electrons that strike the plate without affecting the velocity of impact.

The negative resistance of the dynatron can be used in many ways. Thus Fig. 178*a* shows how a positive resistance *R* may be neutralized in such a way as to give voltage amplification. This is accomplished

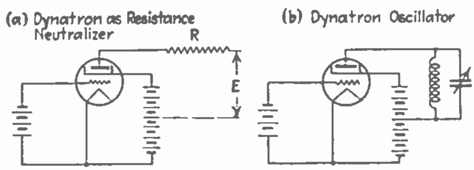


FIG. 178.—Circuits showing how the negative plate-cathode resistance of a dynatron can be used to (*a*) neutralize a positive resistance *R*, and (*b*) develop oscillations in a tuned circuit.

by connecting *R* in series with the plate-cathode resistance of a dynatron adjusted to give a negative resistance approximately equal to *R*. The net resistance which the circuit offers to the applied voltage *E* is then very low, and the current in the circuit will therefore be large. The voltage developed across the resistance *R* will hence be large, and with careful adjustment voltage amplifications as great as 1000 can be obtained. In Fig. 178*b* an arrangement is shown by which the negative resistance of the dynatron may be utilized to produce oscillations in a resonant circuit. When the negative resistance shunted across the tuned circuit has a magnitude that is numerically less than the parallel-resonant impedance of the circuit, the net resistance of the combination is negative, and oscillations will be developed.[1]

80. Inverted Vacuum Tubes.[2]—The inverted vacuum tube is a three-element vacuum tube that has been made to operate as a voltage step-down device by interchanging the grid and plate functions in the ordinary tube circuit. The basic circuit of the inverted vacuum tube is shown in Fig. 179. The essential features of this circuit are:

[1] For further information about the dynatron and its uses see: Albert W. Hull, The Dynatron, a Vacuum Tube Possessing a Negative Resistance, *Proc. I.R.E.*, vol. 6, p. 5, February, 1918.

[2] See F. E. Terman, The Inverted Vacuum Tube, a Voltage-reducing Power Amplifier, *Proc. I.R.E.*, vol. 16, p. 447, April, 1928.

1. The grid is operated at a positive potential and so draws considerable current.
2. The plate is operated at a negative potential and so draws no current.
3. The plate circuit is the input circuit, while the grid circuit is the output circuit.

The inverted vacuum tube is merely an ordinary vacuum tube with the grid doing what the plate usually does, and with the plate carrying on the functions usually performed by the grid.

The inverted vacuum tube acts as a voltage reducer or voltage step-down device, because a voltage applied at the input, *i.e.*, to the plate, has the same effect on the grid current as though a similar but much smaller voltage had instead been added to the grid battery. If the vacuum tube used has a voltage amplification factor of μ, the step-down ratio will be approximately μ. The inverted vacuum tube acts as a power amplifier, for with the plate negative practically no energy is consumed in the plate circuit when a voltage is applied to the input, while at the same time this input voltage controls an appreciable grid current thus controlling considerable energy in the grid, or output, circuit.

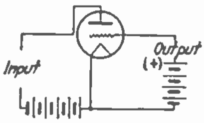


FIG. 179.—Circuit of inverted vacuum tube. The tube is an ordinary triode in which the plate is the control electrode and is operated at a negative potential, while the grid acts as the positive anode.

The operation of the inverted vacuum tube rests on the fundamental fact that the space current flowing to the anode, which in this case is the positive grid, depends almost solely upon the electros tatic field in the vicinity of the cathode and is substantially independent of how this field is produced. Since both plate and grid potentials affect the intensity of this electrostatic field it is possible to use a negative plate as a control electrode to serve the same purpose as the negative grid in the usual triode. The only differences in the result are that the amplification factor is low, being approximately $1/\mu$, where μ is the amplification factor of the tube operated in the normal manner, and that the dynamic anode resistance is much lower than in the corresponding tube operated in the normal manner because changes in grid, *i.e.*, anode, voltage produce large changes in the electrostatic field near the cathode and hence large changes in anode, *i.e.*, grid, current.

The inverted vacuum tube was originally developed as a means for taking voltage oscillograms in vacuum-tube circuits where the voltage was high and where circuit operation would be affected by even a small consumption of current in the measuring device. The voltage wave to be photographed is applied directly to the plate using a suitable plate-bias battery, and by the reducing action of the tube is transformed into a smaller potential acting in the low-impedance grid circuit, developing current variations that can be registered by the oscillograph. The inverted vacuum tube can also perform all the functions of triodes, such as the generation, amplification, and detection of alternating currents, and the measurement of voltages.

81. The Magnetron.—The magnetron is a vacuum tube in which the flow of current from cathode to anode is controlled by a magnetic

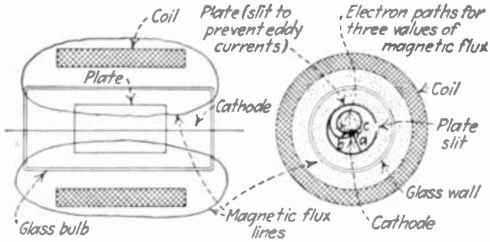


FIG. 180.—Magnetron tube, together with sketch of approximate electron paths for (*a*) zero magnetic field, (*b*) moderate-strength magnetic field, and (*c*) strong field.

field instead of the electrostatic field in the vicinity of the cathode. A typical magnetron consists of a cylindrical anode along the axis of which is located a filament-type cathode, with the entire tube in a magnetic field that is parallel to the filament, as shown in Fig. 180. The anode is prevented from acting as a short-circuited secondary turn to the magnetic field by the slit that is shown in the figure.

The magnetron acts as a valve controlled by the intensity of the magnetic flux. The flow of electrons to the positive plate is substantially unaffected by the presence of magnetic fields having a strength less than a certain critical value, but is completely stopped if the field strength exceeds this critical value even by a small amount. This behavior is a result of the path followed by electrons when being acted upon simultaneously by both electrostatic and magnetic fields, and can be worked out from the principles discussed in Sec. 24. With zero magnetic field the electrons travel directly to the plate as in an ordinary tube and follow a straight-line radial path such as shown at *a* in Fig. 180. When a moderate current is passed through the coil the resulting magnetic flux deflects the path of the electron as shown at *b* because the effect of such a magnetic field is to deflect the electron at right angles to the direction of motion, but the plate current is not affected by the magnetic field because the electrons all reach the plate in spite of their longer paths. When the magnetic field is very intense, however, the electrons travel along a path such as shown at *c* in Fig. 180 and are turned back toward

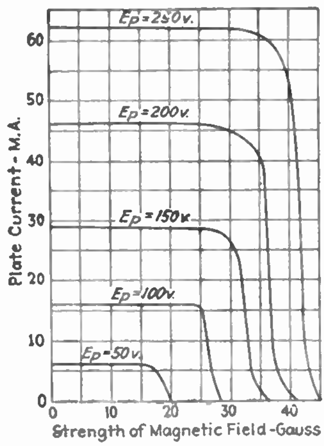


FIG. 181.—Relationship between plate current and magnetic-field strength for various plate voltages with magnetron of type shown in Fig. 180. When the field strength exceeds a certain critical value the plate current is suddenly reduced to zero.

the cathode by the magnetic field without ever reaching the plate even when the plate is at a high positive voltage. This cut-off of plate current is very critical with respect to the magnetic-flux density, so that for all magnetic fields below a critical value the plate current will be substantially unaffected by the presence of the field, while for all greater field strengths the plate current will be nearly zero, as is clearly shown in Fig. 181.

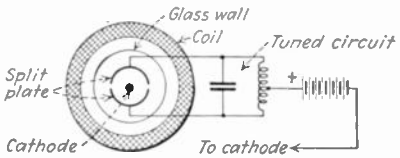


Fig. 182.—Split-anode type of magnetron-oscillator circuit. When the field strength is just greater than the critical value oscillations are developed which have the frequency of the resonant circuit.

The valve action possessed by magnetrons gives them a number of useful properties. Since the energy that must be supplied to the field coil to turn the plate current on and off is less than the energy thus controlled in the plate circuit the magnetron can be used as an amplifier, and as a result of this amplifying property is also capable of functioning as an oscillator. Successful radio-frequency amplifiers of the magnetron type, and magnetron oscillators developing 10 kw. radio-frequency power, have been built.[1] The usefulness of magnetrons in which the plate current is controlled by the strength of the magnetic field is confined to audio and low radio frequencies because of the difficulty of obtaining strong magnetic fields with the small coils that must be employed at high radio frequencies.

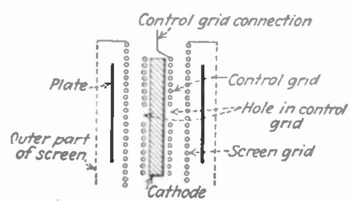


Fig. 183.—Cross section of a variable-mu tube in which the varible-mu characteristic is obtained by a hole in the control-grid structure of a screen-grid tube.

It has been found that the magnetron is capable of generating very high frequency oscillations with good efficiency by using the arrangement shown in Fig. 182, in which the anode is split into two halves and associated with a tuned circuit as shown in the diagram. The axial magnetic field is produced by a d-c current and should have a strength slightly greater than the cut-off value. Magnetron oscillators of the split-anode type are capable of generating relatively large quantities of power at frequencies in the order of 100,000 kc, and in many respects are superior to triode oscillators for such high frequencies.[2]

Magnetrons also operate very satisfactorily as oscillators of the Barkhausen type, in which the frequency is determined by the time of

[1] For additional information regarding magnetrons and their applications see: Albert W. Hull, The Magnetron, *Jour. A.I.E.E.*, vol. 40, p. 715, Sept., 1921; The Axially Controlled Magnetron, *Trans. A.I.E.E.*, vol. 42, p. 915, 1923; Frank R. Elder, The Magnetron Amplifier and Power Oscillator, *Proc. I.R.E.*, vol. 13, p. 159, April, 1925.

[2] See W. C. White, Producing Very High Frequencies by Means of the Magnetron, *Electronics*, vol. 1, p. 34, April, 1930.

passage of the electrons. In the generation of such oscillations best results are obtained when the magnetic flux is slightly greater than the cut-off value, and the frequency generated is determined primarily by the tube dimensions and plate voltage. The highest frequency ever generated by a vacuum-tube oscillator was obtained from a magnetron type of Barkhausen oscillator.[1]

82. Variable-mu Tubes.[2]—The variable-mu tube differs from an ordinary vacuum tube in that the design has been modified in such a way as to cause the plate-current control-grid characteristic curves of the tube to tail off at high grid biases rather than to have a well-defined cut-off point. A characteristic of this type is obtained by using non-uniform tube structure in which the amplification factor (or in the case of screen-

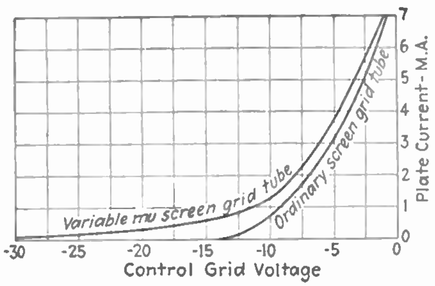


FIG. 184.—Characteristic curve of a typical variable-mu tube compared with the characteristic curve of a corresponding screen-grid tube of ordinary construction.

grid tubes the mutual conductance) is much lower for one part of the tube than for the remainder. Such a construction is illustrated in Fig. 183, in which several turns of wire in the control-grid structure have been omitted. This forms a hole where the control-grid potential exerts relatively little effect on the space current with the result that the tube allows a small amount of space current to flow even at very high negative control-grid voltages, and this small current varies only slowly with grid potential. This is the reason for the characteristic curve illustrated in Fig. 184, in which the behavior is substantially the same as that of an ordinary tube under the usual electrode potentials but differs radically as the grid bias is made highly negative.

Variable-mu tubes are used when it is desired to control the amplification by varying the grid-bias potential of the tube. The characteristic

[1] See Hidetsugu Yagi, Beam Transmission of Ultra-short Waves, *Proc. I.R.E.*, vol. 16, p. 715, June, 1928; Kinjiro Okabe, On the Short Wave Limit of Magnetron Oscillations, *Proc. I.R.E.*, vol. 17, p. 652, April, 1929. The highest frequency that has been reported at the date of this writing is 5,350,000,000 cycles per second, corresponding to a wave length of 5.6 cm.

[2] See Stuart Ballantine and H. A. Snow, Reduction of Distortion and Cross-talk in Radio Receivers by Means of Variable-mu Tetrodes, *Proc. I.R.E.*, vol. 18, p. 2102, December, 1930.

curves of ordinary triode and screen-grid tubes possess considerable curvature where the mutual conductance is low, whereas the characteristic curves of a variable-mu tube are only slightly curved even at extremely low mutual conductances. This is well brought out by comparing the shape of the curves in Fig. 184 for points at which the slope is the same and is very small. The low curvature of the variable-mu characteristic at low mutual conductances minimizes cross-talk interference that would otherwise result, and for this reason variable-mu tubes are used extensively in broadcast receivers.

CHAPTER X

MODULATION

83. Waves with Amplitude Modulation.—In all the commonly used systems of radio communication the intelligence is transmitted by varying the amplitude of the radiated waves, as explained in Sec. 3 and illustrated in Figs. 4 and 5. Communication carried on in this way is said to take place by means of *amplitude modulation*, and the equation of the envelope of the modulated wave represents the equation of the intelligence actually transmitted. If the envelope variations of the modulated wave exactly reproduce the original signal,[1] *i.e.*, sound pressure, light intensity, etc., this signal is transmitted without distortion; otherwise distortion is introduced and the intelligence contained in the modulation envelope is not exactly the same as that represented by the original signal.

When the equation of the wave envelope does not contain the different frequency components of the original signal in their correct relative magnitudes the modulation possesses frequency distortion. If on the other hand the equation of the modulation envelope includes frequency components that were not present in the original signal the modulation possesses amplitude or non-linear distortion. Finally if the phase relations of the different frequency components of the wave envelope are not the same as in the original signal, phase distortion results. These three types of distortion thus have the same significance when applied to modulation as they do in the case of amplification.

The extent of the amplitude variations in a modulated wave is expressed in terms of the degree of modulation, which is defined by the relation

$$\text{Degree of modulation} = m = \frac{\text{average envelope amplitude} - \text{minimum envelope amplitude}}{\text{average envelope amplitude}} \tag{137}$$

When the degree of modulation is 1.0 the amplitude variations carry the envelope amplitude to zero during the troughs of the modulation cycle, and the modulation is complete, *i.e.*, the envelope is varied through the maximum range that is possible without amplitude distortion.

Analysis of Modulated Wave.—It was shown in Sec. 5 that a modulated wave consists of a carrier wave and a number of side-band components

[1] The term "signal" when used in connection with modulation refers to the original intelligence. Thus signal frequency represents the frequency which is modulated upon the radio wave.

the exact nature of which depends upon the equation of the wave envelope. This envelope equation can always be written in the form of an average value plus one or more alternating-current terms that take into account the envelope variations, *i.e.*,

$$\text{Equation of wave envelope} = E_0 + E_1 \sin (2\pi f_1 t + \phi_1) + E_2 \sin (2\pi f_2 t + \phi_2) + E_3 \sin (2\pi f_3 t + \phi_3) + \cdots \quad (138)$$

where E_0 is the average amplitude of the envelope (which is also the amplitude when the wave is unmodulated), and the remaining terms represent the components of the envelope variations having frequencies f_1, f_2, f_3, etc., crest amplitudes E_1, E_2, E_3, etc., and phases ϕ_1, ϕ_2, and ϕ_3. The term E_0 in Eq. (138) represents the average amplitude of the modulation envelope and is the crest amplitude of the carrier wave, while each of the sinusoidal components in the wave envelope gives rise to a pair of side-band frequencies. The two side-band components arising from each sinusoidal component of the wave envelope have frequencies that are respectively greater and less than the carrier frequency by the corresponding envelope frequency, and each has an amplitude one-half of the corresponding envelope component. Thus a modulated wave having an envelope given by the equation.

$$\text{Envelope amplitude} = 100 + 50 \sin 2\pi f_1 + 20 \sin 2\pi f_2$$

consists of a carrier wave having a crest value of 100 volts, a side band having a frequency that is f_1 cycles greater than the carrier frequency and a crest amplitude of 25 volts, a companion side band of the same amplitude but of frequency f_1 cycles less than the carrier frequency, and a second pair of side-band components each of 10 volts amplitude and having frequencies f_2 cycles more and f_2 cycles less than the carrier frequency.

The carrier wave is the same, irrespective of the presence or absence of modulation. Varying the amplitude of the carrier wave merely has the effect of generating the side bands, which are the part of the modulated wave that conveys the intelligence being transmitted. The carrier wave carries no information because it is not affected by the modulation.

The energy contained in a modulated wave is the sum of the energies of the separate frequency components and is therefore increased during modulation because of the energy contained in the side bands. When the carrier wave is completely modulated by a sinusoidal variation of the envelope amplitude, *i.e.*, when in Eq. (138) $E_1 = E_0$, and E_2, E_3, etc., are all zero, there are two side-band components each having an amplitude half that of the carrier and hence each containing one-fourth as much power as does the carrier. The two side bands together thus make the power of the completely modulated wave ($m = 1.0$) 50 per cent greater than the carrier power, and one-third of the total energy of the wave is in the side bands, while two-thirds is in the carrier. With degrees of modulation other than 1.0 the side-band power will be proportional to m^2, so that the fraction of the total wave energy that is contained in the intelligence-bearing side bands rapidly decreases as the degree of modulation is reduced. For this reason the highest possible degree of modulation should always be used.

The frequencies contained in the modulation envelope depend upon the character of the amplitude variations that are impressed on the wave, and in general are higher the more rapidly the amplitude of the wave envelope is varied. In particular when the amplitude changes with extreme rapidity, as would be the case in a code transmitter if the waves could be turned on and off instantly, the envelope equation contains very high frequency components, and the side-band frequencies extend over a wide frequency band.

In the actual transmission of intelligence by means of modulated waves it is generally unnecessary to modulate upon the carrier wave all of the frequencies contained in the equation of the original signal in order to make the intelligence understandable at the point of reception. Furthermore the modulation by a wider frequency band than is absolutely necessary to carry the desired information is to be avoided because unessential frequencies in the modulation envelope mean high-frequency side-band components, and this utilizes frequencies in the transmitting medium that could otherwise be employed for additional communication.

Side Bands Required in Telegraph, Telephone, and Picture Transmission.—In the transmission of telegraph signals by the Continental Morse Code it is possible to operate telegraph relays provided each side band has a width of 0.131 cycle per letter transmitted per minute.[1] Thus transmission at the rate of 100 letters per minute can be carried on using side bands which extend only 13.1 cycles on each side of the carrier frequency.

In the transmission of speech and music of high quality, side-band components extending at least 5000 cycles on each side of the carrier frequency must be employed. Such a band provides for the transmission of audio-frequency sounds having pitches up to 5000 cycles, and while the human voice and music contain frequencies up to approximately 15,000 cycles these higher pitch sounds are not absolutely essential for reasonably satisfactory results. Understandable speech requires the

[1] This figure represents the minimum side band that can be used, and unless refinements are employed a somewhat wider band is desirable. With the five-element two-valued code employed in printing telegraph systems the side band need be only about three-fourths as wide as with the Continental Morse Code. By employing a synchronous vibrating relay to restore the shape of the received signals transmitted by the printing telegraph code it is possible to cut the frequency band in half. See F. E. Terman, Some Possibilities of Intelligence Transmission when Using a Limited Band of Frequencies, *Proc. I.R.E.*, vol. 18, p. 167, January, 1930.

reproduction of all frequencies from about 250 to 2700 cycles, or side-band frequencies ranging from 250 to 2700 cycles above and below the carrier frequency.

The side band required in picture transmission when using amplitude modulation is proportional to the area of picture and number of elements per square inch. The minimum side band that will carry the picture is one-half of the number of picture elements transmitted per second, and this is sufficient only when all phase and amplitude distortion is corrected. It therefore requires a side band at least 1800 cycles wide to transmit 1 sq. in. per second of 60-line picture under the most favorable conditions. Television is merely picture transmission speeded up, but, as each individual picture need not have a quality equal to that required of a still-picture, television transmission can use fewer lines per inch in the picture and can use a lower frequency band in proportion to the number of elements. A good-quality 50-line picture 1 in. square, when repeated sixteen times per second, has been found to require a side band of at least 15,000 cycles per second when all phase and amplitude distortion has been corrected.[1]

Methods of Amplitude Modulation.—While the methods by which amplitude modulation of a carrier wave can be obtained appear to be almost without number, nearly all of them come under one of the following four classifications:

1. Modulated oscillators.
2. Modulated amplifiers.
3. Modulation by means of non-linear circuit elements.
4. Modulation by means of variable circuit elements.

Modulated oscillators include those arrangements in which the amplitude of the generated oscillations is controlled by the intelligence that is to be transmitted. In the modulated-amplifier type of modulator the unmodulated carrier wave is applied to the amplifier grid, and the amplification is then varied in accordance with the intelligence to be transmitted. Modulators which employ non-linear circuit elements make use of the fact that the current which flows through a non-linear circuit is not proportional to the applied voltage, so that when the radio-frequency carrier voltage and low frequency signal voltage are superimposed and applied to a non-linear circuit, the amplitude of the radio-frequency current that flows is modulated by the signal. In modulators employing variable circuit elements the impedance which is offered to the radio-frequency carrier voltage is varied in accordance with the intelligence that is to be transmitted, with the result that the radio-frequency current is modulated as desired. The following paragraphs describe the methods

[1] See: Frank Gray, J. W. Horton, and R. C. Mathes, The Production and Utilization of Television Signals, *Bell System Tech. Jour.*, vol. 6, p. 560, October, 1927.

of amplitude modulation which are most widely used in telephony and in picture transmission. The means used to turn code transmitters on and off in accordance with the telegraph characters are considered in Sec. 104.

84. The Plate-modulated Oscillator.—This method of modulation, which was invented by R. A. Heising, makes use of an oscillator in which the amplitude of oscillations and the plate current drawn by the tube are both proportional to the applied plate voltage. Such a characteristic can be readily approximated in practice, as is apparent from the curves of Fig. 185. The modulated wave is produced by superimposing upon

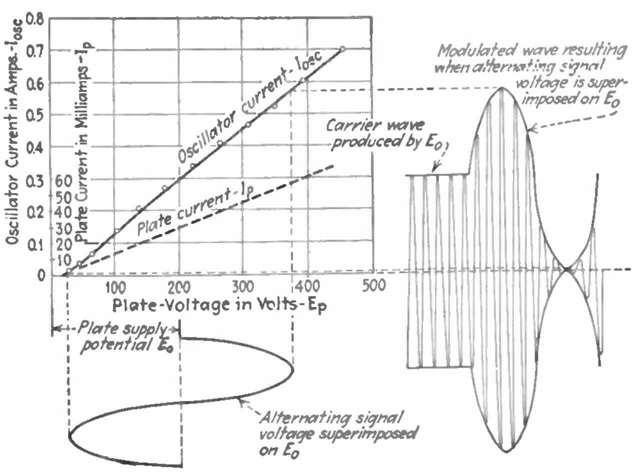


FIG. 185.—Plate and resonant-circuit currents as a function of plate-supply voltage in a typical oscillator having grid leak and condenser, and adjusted for efficient operation. The oscillating current is seen to be almost exactly proportional to the plate voltage, so that the oscillations can be modulated by varying the plate-supply voltage about the average value E_o as shown in the figure.

the direct-current plate voltage an alternating voltage that varies in accordance with the signal that is to be transmitted, causing the actual voltage applied to the tube to vary about the average or direct-current value. Since the amplitude of oscillations is proportional to the plate voltage the result is that the oscillator generates the desired modulated wave, as is made apparent in Fig. 185.

The alternating voltage that varies with the signal being transmitted and which is superimposed upon the direct-current plate potential is normally obtained from the output of a vacuum-tube amplifier having sufficient power capacity and amplification to deliver the power that is needed to vary the plate potential of the oscillator through the required range. The power tube in such an amplifier is often spoken of as a modulator tube. In Fig. 186 are shown the methods commonly employed to couple the amplifier (or modulator) output to the plate circuit of the oscillator. It will be noted that with each arrangement the oscillator plate circuit represents the load impedance to which the amplified signal power is delivered. This load impedance is equivalent to a resistance having a value equal to the ratio of oscillator plate voltage to oscillator plate current (*i.e.*, the reciprocal of the slope of the $E_p - I_p$ curve of Fig. 185). In Fig. 186*a* the load resistance as represented by the oscillator plate circuit is directly in series with the plate circuit of the amplifier,

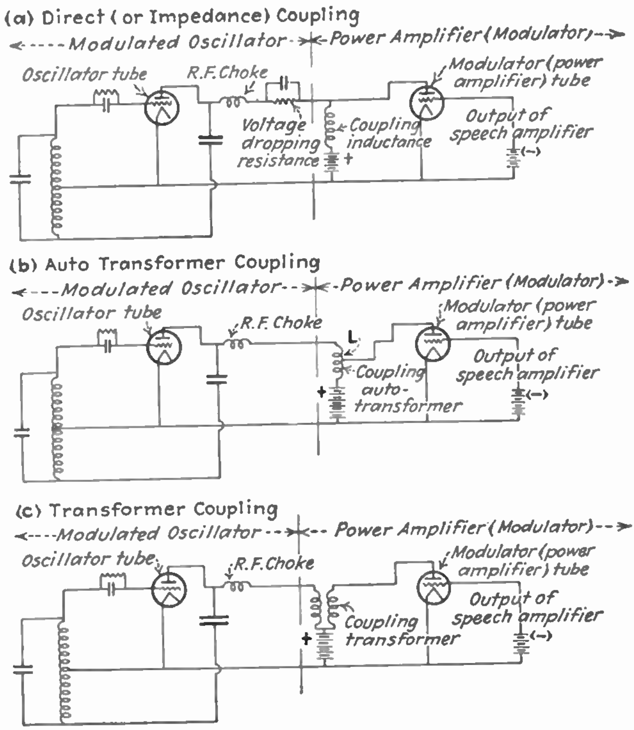


FIG. 186.—Circuits for the plate-modulated oscillator. In each case the oscillator plate circuit represents the load impedance to which the amplifier power output is delivered. A shunt-feed Hartley oscillator circuit is shown in each case.

the coupling inductance merely providing a means of applying the direct-current plate-supply voltage to both oscillator and amplifier.[1] Figure 186*b* is an autotransformer method of coupling the oscillator and amplifier, and when a step-up transformation ratio is used permits a better matching of the load impedance to the plate resistance of the amplifier than is possible with the direct-coupled circuit of Fig. 186*a*. The transformer-

[1] The reactance which this choke offers to the amplifier output should be so high as to prevent the coil from carrying appreciable alternating current. Under these conditions the current through the coil is substantially constant and for this reason the plate-modulated oscillator is sometimes called the constant-current system of modulation.

coupled arrangement of Fig. 186*c* also permits proper matching of load and plate resistance.

The degree of modulation that is obtainable from a plate-modulated oscillator depends upon the amount of undistorted power which the signal amplifier, *i.e.*, modulator, is able to deliver to the plate circuit of the oscillator. When sufficient undistorted power is available to superimpose upon the direct-current oscillator plate voltage an alternating signal voltage that swings the instantaneous plate potential from zero to twice the direct-current plate voltage, as shown in Fig. 185, the degree of modulation is 100 per cent, while smaller amplifier outputs will give correspondingly lower amounts of modulation.

The linear relation between amplitude of oscillations and plate voltage that is required in the modulated oscillator can be obtained in practically any type of oscillator circuit provided the grid bias is obtained by the use of a grid leak and grid condenser, and provided the oscillator is adjusted to operate with good efficiency. The grid leak and condenser are required in order that the grid bias may vary with the plate voltage in such a way as to insure efficient operating conditions at all times. The linear relation between oscillating current and plate potential that exists in an efficient oscillator employing a grid leak and condenser results from the fact that the amplitude of oscillations under these conditions is always such that the alternating plate-cathode voltage has a crest value only slightly less than the plate-supply potential, as explained in Sec. 50.

Satisfactory operation of the modulated oscillator can be obtained only when the oscillator plate current is approximately proportional to the plate-supply potential. This is because a curved $E_p - I_p$ oscillator characteristic causes the oscillator plate circuit to offer a variable resistance to the amplifier and so results in amplitude distortion. The characteristics of ordinary oscillators are fortunately such that no difficulty is encountered on this score.

Power Relations in Modulated Oscillator.—The power relations that exist in the modulated oscillator can be determined by assuming that the curves of Fig. 185 are straight lines which pass through the origin. When the oscillations are unmodulated (*i.e.*, no alternating potential superimposed upon the direct-current plate voltage) the power P_0 which is supplied to the oscillator from the direct-current plate-supply voltage is

$$\left.\begin{array}{l}\text{Power supplied oscillator from}\\ \text{direct-current voltage source}\end{array}\right\} = P_0 = E_0 I_0 \qquad (139)$$

where E_0 is the applied direct-current plate voltage, and I_0 is the plate current which is drawn with this voltage (see Fig. 185). The power of the oscillations generated under these conditions, *i.e.*, the carrier power, is equal to nE_0I_0, where n is the plate efficiency of the oscillator. When the amplitude of the oscillator is modulated the amplifier must deliver to the oscillator plate circuit the power which is required to vary the plate voltage. With complete modulation (*i.e.*, $m = 1.0$) the alternating voltage which the amplifier output must develop has a crest value equal to the direct-current plate potential, and this requires an amplifier output equal to exactly one-half of the power drawn by the unmodulated oscillator from the source of direct-current voltage. With other degrees of modulation the power output which the amplifier must deliver will be proportional to m^2. Thus if the direct-current power drawn by an oscillator when unmodulated is 1000 watts the undistorted power output which the amplifier must deliver for complete modulation is 500 watts, while for $m = 0.20$ the amplifier output need be only 20 watts, but the power contained in the intelligence-bearing part of the modulated wave will be correspondingly less. *The power relations that exist in a plate modulated oscillator can be summarized by stating that the power required to generate the carrier wave is supplied from the direct-current plate-supply voltage, while the power required to generate the side-band components of the modulated wave must be supplied to the oscillator from the output of the amplifier (modulator).*

An oscillator which is to have its output completely modulated must be operated at a somewhat lower direct-current plate-supply voltage, and must have a lower plate loss with this plate voltage than is permissible when the oscillator will not be modulated. A low direct-current plate voltage is required because with complete modulation there are intervals when the total voltage applied to the plate is twice that of the direct-current source. The power supplied to the plate of the tube averages 50 per cent greater during intervals of complete modulation than when only a carrier wave is being generated, and the oscillator must be operated to take this into account. Since the plate losses are also 50 per cent greater when the modulation is complete than when there is no modulation, the carrier wave that can be generated by a completely modulated oscillator contains only two-thirds as much energy as could be generated with the same plate loss if the oscillations were of constant amplitude.

As a consequence of the power relations that exist in a plate-modulated oscillator, and also because of the low plate efficiency of the distortionless power amplifier, it takes three to five amplifier tubes to completely modulate the oscillations generated by a similar tube operated as an oscillator. The direct-current plate power P_a which must be handled by an amplifier having a plate efficiency n_a to modulate completely the output of an oscillator tube having an allowable plate loss of P and a plate efficiency n_0 is readily shown to be

$$\text{Amplifier plate power} = P_a = \frac{P}{3n_a(1 - n_0)} \qquad (140)$$

The amount of completely modulated carrier power that can be developed by an oscillator of allowable plate loss P and plate efficiency n_0 is

$$\text{Carrier power} = \frac{2n_0P/3}{1 - n_0} \tag{141}$$

Thus if the oscillator has an allowable plate loss P of 100 watts and a plate efficiency n_0 of 60 per cent, while the amplifier plate efficiency is 20 per cent, the required amplifier plate power is 417 watts, *i.e.*, it takes at least four tubes acting as amplifiers to modulate completely a similar tube functioning as an oscillator, and generating a carrier power of $\frac{2}{3} \times 0.6 \times 100/(1 - 0.6) = 100$ watts.

Circuit Requirements of Oscillator.—In order that the amplitude of oscillations may change as rapidly as does the plate potential, the effective Q of the oscillator resonant circuit and the size of grid leak and condenser must be properly proportioned. Since the side-band frequencies differ slightly from the resonant frequency of the oscillator tuned circuit they tend to be suppressed to an extent that is dependent on the effective Q of the circuit and the ratio of modulation, *i.e.*, signal, frequency to the carrier frequency. The problem of avoiding frequency distortion therefore becomes more difficult the lower the carrier frequency and the higher the frequency that is modulated upon the carrier. The negative bias which the grid leak and condenser combination applies to the grid of the oscillator during modulation must vary as rapidly as does the plate potential if amplitude and frequency distortion are to be avoided. This requires that the grid leak and grid condenser be chosen in relation to the degree of modulation and highest signal frequency that is to be modulated upon the wave in such a way as to satisfy Eq. (109), which is

$$\frac{X}{R} \geqslant \frac{m}{\sqrt{1 - m^2}}$$

where m is the degree of modulation, R is the grid-leak resistance, and X is the reactance of the effective grid-condenser capacity at the modulation frequency involved.

Methods of Coupling Modulator to Oscillator Plate Circuit.—When the oscillator plate circuit is directly coupled to the amplifier output as in Fig. 186*a*, the load impedance which the oscillator offers to the amplifier is such that complete modulation cannot be obtained unless the oscillator is operated at a lower direct-current plate potential than is the amplifier. The usual way of meeting this requirement is to place in series with the oscillator plate lead a suitable voltage-dropping resistance that is bypassed to currents of the modulation frequency by a large condenser, as shown in Fig. 186*a*. The circuit of Fig. 186*a* has the disadvantage that the load impedance into which the amplifier tube operates is determined by the characteristics of the oscillator and cannot easily be adjusted to the most desirable value. This limitation is overcome by the circuit shown in Fig. 186*b*, in which the inductance coil L acts as an autotransformer for the amplifier output, as well as furnishing a means of applying the direct-current plate voltage to the tubes. By connecting the amplifier plate to the proper point on the inductance, the load impedance into which the amplifier operates can be adjusted to the optimum value. With autotransformer coupling, complete modulation can be obtained without the use of a resistance in the plate circuit of the oscillator. In Fig. 186*c* the plate circuit of the oscillator is coupled to the amplifier by means of a transformer, and by varying the turn ratio it is possible to make the effective load impedance into which the amplifier operates have any desired value. The transformer-coupled arrangement also permits oscillator and amplifier tubes requiring widely different plate voltages to be operated together, since separate direct-current supply sources can be used.

The coupling inductances used in the circuits of Fig. 186*a* and 186*b* and the transformer of Fig. 186*c* are designed exactly as in the case of any power amplifier. This means that if frequency distortion is to be avoided the impedance of the coupling element to the signal currents must be high as compared with the resistance formed by the amplifier plate resistance in parallel with the equivalent load resistance reduced to a unity step-up ratio. In every case this ratio of impedance to resistance drops off at low signal frequencies because of the low reactance which an inductance has at low frequencies, and drops off at high frequencies because of distributed capacities that shunt the coil. The result is a tendency to discriminate against both low and high modulation frequencies. The coupling inductance is called upon to carry relatively large d-c currents and must be designed to avoid direct-current core saturation. In the transformer-coupled circuit of Fig. 186*c* the two windings can be so polarized that the magnetizing effect of the direct-current plate current drawn by the amplifier tubes opposes the magnetization of the d-c current flowing to the oscillator plate, and if the turn ratio and tube characteristics are properly chosen the resultant core magnetization can be made negligible. When this condition exists the air gap in the iron core can be small, thus giving a large inductance with light weight and small volume.

The plate-modulated oscillator can be easily adjusted to give satisfactory results, has no limit to the amount of power that can be modulated, and permits a high degree of modulation without introducing amplitude distortion. The chief limitation of this type of modulator arises from the fact that the frequency generated by ordinary oscillators depends somewhat upon the plate-supply voltage. The carrier frequency generated by a plate-modulated oscillator therefore tends to vary with the modulation envelope, introducing frequency modulation (see Sec. 89), which produces undesirable consequences. As a result the plate-modulated oscillator, which was once the universally used method of modulation, is now employed only where frequency stability is unimportant.

85. Plate-modulated Class C Amplifiers.—A plate-modulated Class C amplifier is an ordinary Class C amplifier (*i.e.*, an amplifier which is biased beyond cut-off as explained in Sec. 52) to the grid of which there is applied the unmodulated carrier voltage. The modulation is accomplished by superimposing upon the direct-current plate voltage an alternating potential that varies in accordance with the intelligence to be transmitted exactly as in the plate-modulating oscillator of Sec. 84.

The details of the action that takes place in a modulated Class C amplifier are shown in Fig. 187, which is for a case where the grid is biased to twice cut-off, and where complete modulation is obtained. The signal voltage is shown at Fig. 187*b*, and when superimposed upon the plate-supply voltage gives the plate-voltage wave of Fig. 187*c*.

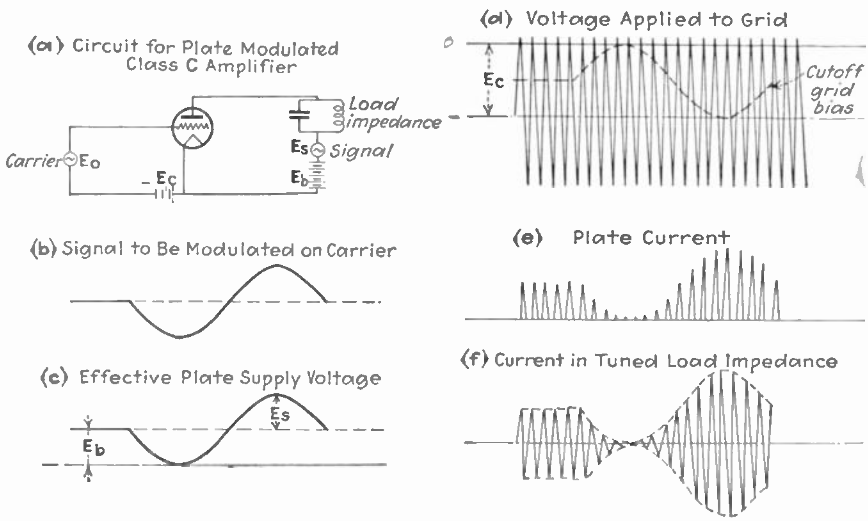


Fig. 187.—Schematic circuit of plate-modulated Class C amplifier, together with oscillograms showing details of the modulator operation.

The degree of modulation depends upon the range through which the plate voltage is varied by the signal and is 100 per cent when the variation is between zero and twice the direct-current supply potential. The voltage acting on the grid of the modulated amplifier is shown at Fig. 187*d* and consists of the carrier voltage superimposed upon a grid bias approximately twice the cut-off value. The carrier wave normally has sufficient amplitude to run the grid slightly positive at the crest of each cycle. The plate current that results from the joint action of the plate supply of Fig. 187*c* and the grid voltage of Fig. 187*d* is shown at Fig. 187*e* and is in the form of impulses of current having an amplitude that varies in accordance with the plate voltage. This is because plate current flows only when the instantaneous grid voltage, as given at Fig. 187*d*, exceeds the cut-off value, while this cut-off grid bias is at the same time proportional to the plate voltage and so varies with the modulation as shown by the dotted lines of Fig. 187*d*. The plate-current impulses of Fig. 187*e* produce in the tuned load circuit the current shown at Fig. 187*f*. This is the desired modulated wave having a modulation which reproduces the original signal of Fig. 187*b*, and a carrier frequency corresponding to the frequency of the alternating-current exciting voltage applied to the amplifier grid.

Analysis of Action.—A quantitative analysis of the action taking place in a modulated Class C amplifier can be carried out by a modification of the method used in analyzing the Class B amplifier (see Sec. 46), but the expressions which result are too complicated to be useful under ordinary circumstances.[1] It is possible, however, to draw a number of general conclusions regarding the behavior of the modulated Class C amplifier without making a complete analysis. Thus it will be observed that when the grid bias is twice the value required to cut-off the direct-

[1] This analysis can be carried out as follows: On the assumption that a voltage acting on the grid produces the same effect on the plate current as a voltage μ times as great added to the plate voltage, the current impulses actually existing in the plate-modulated amplifier will be the same as though the grid potential was zero and the plate voltage was

$$[E_B - \mu E_c + E_{\text{mod.}} + \mu E_0 \cos \omega t - I_{ac} Z_L \cos (\omega t + \theta)]$$

where E_c and E_B are the grid bias and direct-current plate voltages, respectively; $E_{\text{mod.}}$ is the signal voltage that is superimposed on E_B to produce the modulation; $E_0 \cos \omega t$ is the carrier voltage applied to the grid; and $I_{ac}Z_L \cos (\omega t + \theta)$ is the voltage drop across the load impedance. The term I_{ac} in this last expression is the crest amplitude of the modulated wave contained in the plate-current impulses for the value of $E_{\text{mod.}}$ that exists at the instant in question, while Z_L is the magnitude and θ the phase angle of the load impedance which the plate circuit offers to the modulated wave contained in the plate-current impulses.

The current I_{ac} varies with $E_{\text{mod.}}$ and for distortionless modulation must be directly proportional to $E_{\text{mod.}}$. If the dynamic characteristic of the modulated amplifier is linear over the part of the cycle during which current flows, and the ratio between plate voltage and plate current for this linear characteristic is r, then the plate current impulses are equal to

$$\frac{I}{r}[E_B - \mu E_c + E_{\text{mod.}} + \mu E_0 \cos \omega t - I_{ac} Z_L \cos (\omega t + \theta)]$$

with the understanding that negative values of this expression represent zero current. The value of I_{ac} can now be determined by making a Fourier analysis of this expression for plate current. In the usual case where the load impedance Z_L is a resistance (load impedance tuned to resonance), the phase angle θ is zero, and I_{ac} is given by the equation

$$I_{ac} = \frac{2}{\pi}\int_0^T \frac{1}{r}[E_B - \mu E_c + E_{\text{mod.}} + (\mu E_0 - I_{ac} Z_L) \cos \omega t] \cos \omega t d \qquad (142)$$

where T represents the value of time that makes the quantity in the main brackets under the integral sign become zero. Carrying out this integration and solving the resulting equation for I_{ac} gives the modulator output in terms of $E_{\text{mod.}}$ and the other parameters.

current plate-supply voltage, the plate current at the crest of the modulation cycle will flow in impulses having the shape of half sine waves. This is the condition for Class B (linear) amplifier operation and results in the plate efficiency given by Eq. (101). During the remainder of the cycle each impulse of plate current lasts for less than a half cycle, and Class C operation results. Since the plate efficiency of a Class C amplifier is greater than that of a Class B amplifier it is apparent that the

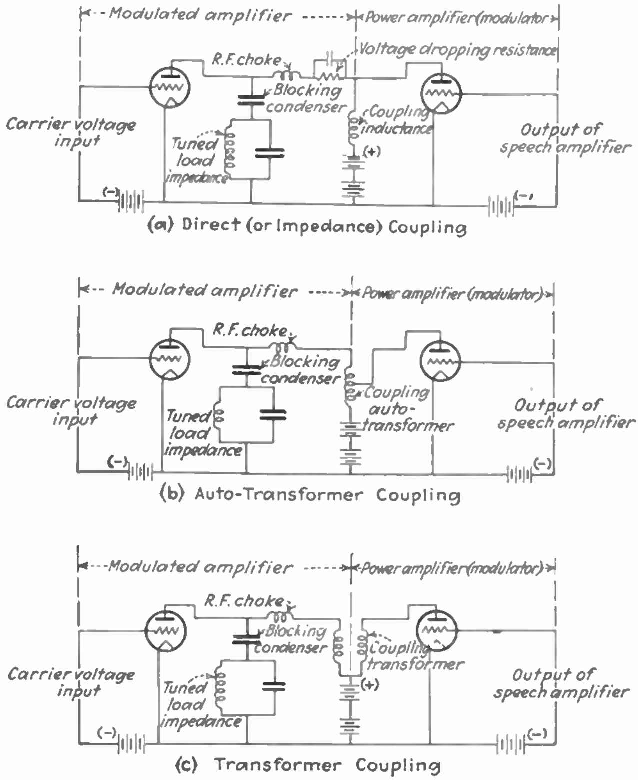


FIG. 188.—Circuit diagrams for plate-modulated Class C amplifiers. Comparison with Fig. 186 shows that the circuits for the plate-modulated oscillator and the plate-modulated amplifier are very similar. In practice it is customary to neutralize the grid-plate capacity of the modulated amplifier.

plate efficiency of a modulated Class C amplifier will vary somewhat during the modulation cycle but will never be less than the value corresponding to that given by Eq. (101) and Fig. 104 for Class B (linear) amplifier operation. Thus the maximum possible plate efficiency of the plate-modulated Class C amplifier lies somewhere between 78 and 100 per cent, and the plate efficiency under practical conditions will normally be not less than 50 per cent but will depend upon the impedance of the load in relation to the plate resistance of the tube in much the same way as does the plate efficiency of a Class B amplifier.

When the plate-modulated Class C amplifier is operated with a grid bias equal to twice the cut-off value, and with complete modulation (the conditions for which Fig. 187 is drawn), the power output can be calculated from the fact that at the crest of the modulation cycle, that is, when the total plate voltage is twice the direct-current value, the modulator is working as a Class B amplifier and has an output which is given by Eq. (98). This output represents the peak power of the modulation cycle, which is four times the carrier power of the completely modulated wave and eight times the total side-band power. The undistorted power output which the modulator tube must supply to the amplifier to accomplish the modulation represents to a first approximation the power from which the side-band components of the modulated wave are generated, and can be calculated with an accuracy sufficient for ordinary purposes in the same way as in the case of a plate-modulated oscillator by using the average plate efficiency of the amplifier.

Circuits of Modulated Class C Amplifiers.—The circuits used to obtain plate-modulated Class C amplification are shown in Fig. 188 and combine an ordinary radio-frequency amplifier circuit with a plate power-supply system similar in all respects to those used in plate-modulated oscillators as shown in Fig. 186. The relative merits of the different arrangements are the same in both types of modulators and have already been taken up. Inasmuch as the plate-modulated amplifier is essentially a radio-frequency amplifier it is always desirable to use neutralization of some type in order to prevent regeneration and self-oscillation.

Factors Determining Degree and Linearity of Modulation.—The degree and linearity of the modulation depend upon the adjustment of the modulated amplifier. The degree of modulation is determined by the range through which the amplifier plate-supply voltage is varied, and will reach 100 per cent when the crest value of the signal voltage which is introduced into the plate circuit is equal to the voltage of the direct-current supply. The linearity of the modulation, that is, the exactness with which the envelope of the modulated output wave varies in accordance with the intelligence to be transmitted, depends upon the tube characteristics and the conditions under which the tube is operated. As a general rule most satisfactory results are obtained when the grid bias is approximately twice the value required to cut-off the direct-current plate potential, and when the grid is allowed to go slightly positive during each cycle. Increasing the ratio of the load impedance to plate resistance of the tube will always make the modulation more nearly linear, and if satisfactory results are to be obtained it is essential that the ratio Z_L/r when defined as in the case of the linear amplifier, be at least 2. By making the load resistance very high compared with the plate resistance of the tube it is possible to obtain substantially distortionless modulation irrespective of the remaining adjustments, although this involves a sacrifice in power output.

Design of Plate-modulated Class C Amplifier.—The procedure to be followed in designing a plate-modulated Class C amplifier is as follows: A tube is first chosen which is satisfactory as a Class B (linear) amplifier. If m is the maximum degree of modulation to be employed this tube must have a power capacity when operated as a linear amplifier that is $(1 + m)^2$ times the desired carrier output, assuming that with linear amplifier operation the plate voltage is $(1 + m)$ times the direct-current plate potential to be actually used in the modulated amplifier. The direct-current plate-supply voltage must be somewhat less than the normal rating for the tube when operated as an amplifier or unmodulated oscillator, exactly as in the case of the modulated oscillator and for the same reason. The grid bias that gives most satisfactory results approximates twice the cut-off value as computed on the basis of the direct-current plate voltage, and in no case should be less than $(1 + m)$ times the cut-off value. The load impedance in the plate circuit of the modulated amplifier should be sufficiently high to make the ratio Z_L/r, as defined for the linear amplifier, have a value of at least 2, and preferably more. The modulator is then made large enough to deliver $m^2P_0/2\eta$ watts of undistorted signal-frequency power to the modulated amplifier, where P_0 is the carrier power that is to be obtained from the modulated amplifier, m is the maximum degree of modulation desired, and η is the estimated plate efficiency, which will normally be in the order of 60 per cent. After setting up the amplifier by the approximate procedure outlined above, the optimum grid bias, plate load impedance, and grid-exciting voltage must then be determined by experiment. The best adjustment is the one for which the relation between direct-current plate voltage and current in the tuned load circuit is most nearly linear when the grid-exciting voltage is kept constant.

The plate-modulated Class C amplifier is the most widely used type of modulator. When properly adjusted it will develop a completely modulated wave free of frequency and phase distortion and with very little amplitude distortion. The plate efficiency of the modulator is also high, and a large power output per tube is developed. The only disadvantage of the plate-modulated Class C amplifier is that it requires a relatively large modulator capacity in order to obtain complete modulation.

86. The van der Bijl Type of Modulated Class A Amplifier.—This type of modulator consists of a Class A, *i.e.*, distortionless, power amplifier, to the grid of which are applied a small radio-frequency carrier voltage and a large signal voltage. Because of the curvature of the plate-current grid-voltage characteristic the amplification of the small carrier voltage depends upon the amplitude of the signal voltage, thus causing the amplified output to be the desired modulated wave. The detailed mechanism by which the modulation is produced is shown in Fig. 189. The degree of modulation depends upon the curvature of the $E_g - I_p$ characteristic and the amplitude of the signal voltage that is applied to the grid. In order to obtain a high degree of modulation the tube must be operated over a curved portion of the $E_g - I_p$ characteristic, and the signal voltage must be large.

An exact theoretical analysis of the behavior of the van der Bijl modulator is extremely difficult because the grid-voltage plate-current tube characteristic does not follow a simple law, and also because the dynamic characteristic of the modulated amplifier is complicated by the load impedance in the plate circuit of the tube offering different impedances to the carrier and signal frequencies. An approximate solution

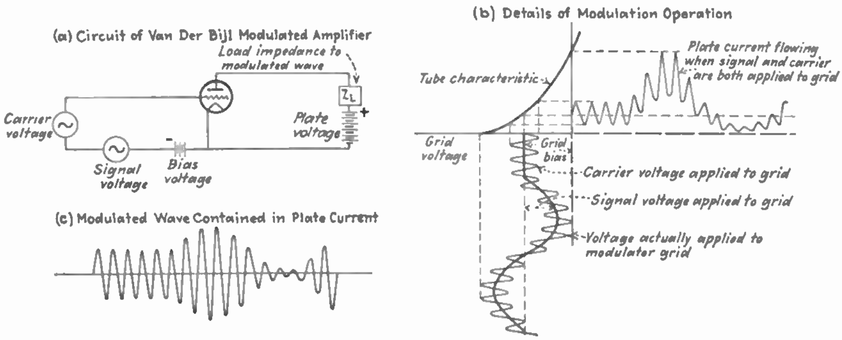


FIG. 189.—Circuit of van der Bijl type of modulated Class A amplifier, together with oscillograms showing details of operation. The curvature of the grid-voltage plate-current tube characteristic causes the amplification of the carrier to depend upon the grid potential, which in turn varies in accordance with the signal.

representing a first-order approximation can however be obtained by assuming that the tube characteristic is a section of a parabola (*i.e.*, follows a square-law characteristic) and that the carrier voltage applied to the input of the amplifier tube is much smaller than the applied signal voltage. Under these conditions the plate-current components that flow as a result of the application of the signal and carrier voltages are exactly the currents that would be produced by a series of suitable generators acting in a circuit consisting of the dynamic plate resistance in series with the plate load impedance. The magnitude of these generator voltages is given by the formula

Equivalent voltage acting in plate circuit =

$$\mu\frac{\left(E_s\dfrac{R_p}{R_p + Z_s} + E_o\dfrac{R_p}{R_p + Z_o}\right)^2}{2\dfrac{G_m}{\partial G_m/\partial E_g}} - \mu E_s - \mu E_o \qquad (143)$$

where

E_s = signal voltage applied to the grid
E_o = carrier voltage applied to the grid
R_p = dynamic plate resistance of tube

μ = amplification factor of tube
G_m = mutual conductance of tube
E_g = grid-bias voltage
Z_o and Z_s = plate load impedance to carrier and signal frequencies, respectively.

In making use of Eq. (143) it is necessary to write E_o and E_s as a function of time before squaring. Upon carrying out the operation indicated by the equation it is found that the equivalent voltage that can be considered as producing the plate current contains a number of components of different frequencies, included among which are terms representing side-band and carrier components.[1]

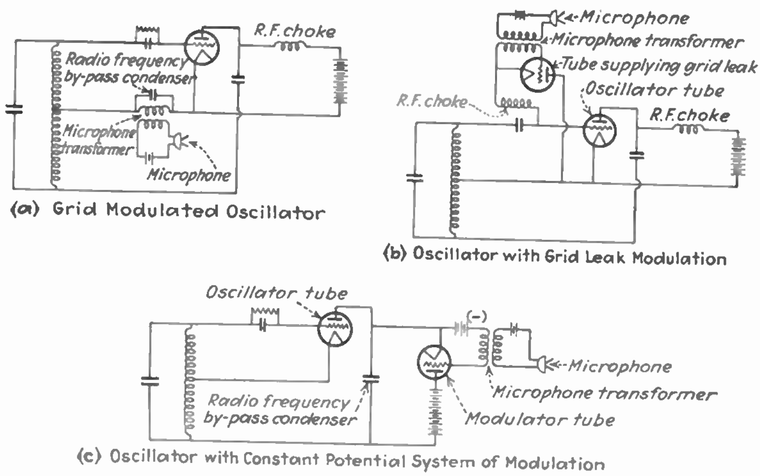


FIG. 190.—Types of modulated oscillators which are seldom used. In each case a conventional Hartley oscillator circuit is shown, and the signal voltage is indicated as derived directly from a microphone, although it is to be understood that the required signal voltage could be obtained from the output of an amplifier.

An analysis based on Eq. (143) shows that the maximum output is obtained when the plate load impedance is negligible to currents of the signal frequency and is one-third the plate resistance to carrier frequency currents. The proper grid bias to use is one-half of the cut-off bias, and in order to obtain complete modulation the applied signal voltage should have a crest value approximately equal to the grid bias.

The van der Bijl type of modulated amplifier is relatively easy to adjust and requires negligible signal power. At the same time, its usefulness is limited by the fact that the plate efficiency, *i.e.*, the ratio of power contained in the side bands to the direct-current plate power, is very low, and that a high degree of distortionless modulation is hard to obtain.

[1] This method of treating the van der Bijl type of modulator was originated by John R. Carson and is described in detail in his paper, The Equivalent Circuit of the Vacuum-tube Modulator, *Proc. I.R.E.*, vol. 9, p. 243, June, 1921.

This system of modulation is used extensively in carrier-current telephone communication, where only small amounts of modulated power are required, and where a high degree of modulation is not necessary because the carrier component is suppressed.

87. Miscellaneous Methods of Modulation.—While the preceding three sections describe the only types of modulators that are now used to any great extent, many other methods have been devised. The most important and representative of these miscellaneous modulation methods are briefly considered in the following paragraphs.

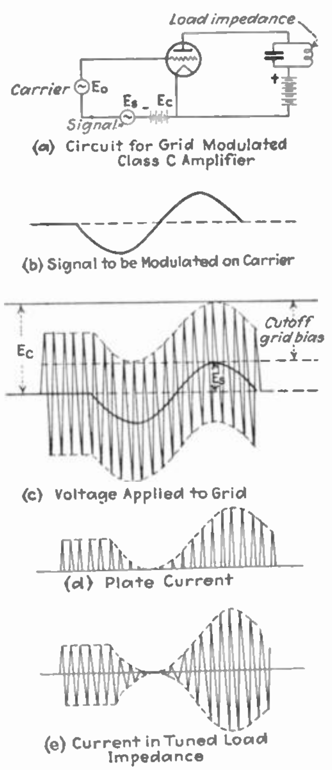


FIG. 191.—Circuit of grid-modulated Class C amplifier, together with oscillograms showing details of operation.

Modulated Oscillators.—Three types of modulated oscillators are shown in Fig. 190. The first of these operates by applying a signal-frequency voltage to the oscillator grid, thereby varying the effective grid bias as is done in the van der Bijl type of modulated amplifier. This varying grid bias modulates the generated oscillations but introduces excessive amplitude distortion unless the degree of modulation is very low, so that such modulators are seldom employed. In the circuit shown at Fig. 190*b* the amplitude of the generated oscillations is controlled by varying the grid-leak resistance in accordance with the signal to be transmitted. This variable-resistance grid leak is supplied by the plate-cathode resistance of a vacuum tube, to the grid of which is applied the signal voltage. Varying the grid-leak resistance alters the effective grid bias in such a way as to control the amplitude of the generated oscillations, and it is possible in this way to obtain fairly satisfactory modulation. The adjustments for a high degree of distortionless modulation are rather delicate, however, and as a result this method of modulation is seldom used although it has been employed in broadcasting.[1] The modulated-oscillator arrangement shown in Fig. 190*c* is similar to the plate-modulated oscillator except that a series instead of a shunt feed is used. This arrangement, which is sometimes called the constant-potential system of modulation, is inferior to the Heising method because complete modulation cannot be obtained

[1] See Charles A. Culver, An Improved System of Modulation in Radio Telephony, *Proc. I.R.E.*, vol. 11, p. 479, October, 1923.

without amplitude distortion, and furthermore the amplifier and oscillator tubes must be operated at different direct-current potentials with respect to ground.

Grid-modulated Class C Amplifier.—The grid-modulated Class C amplifier is similar to the plate-modulated Class C amplifier except that the modulation is accomplished by varying the grid bias of the amplifier instead of the plate voltage. The circuit for such an amplifier together with the details of operation are illustrated in Fig. 191. The tube is operated at a grid bias greater than the cut-off value, and the amplitude of the carrier wave that is applied to the grid is somewhat less than the bias. A signal voltage is then superimposed on the bias and causes the voltage actually applied to the grid to vary as shown at Fig. 191*c*. The result is the plate-current impulses shown at Fig. 191*d*, which are similar to the corresponding plate-current impulses of the plate-modulated amplifier.

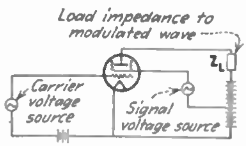


FIG. 192.—Screen-grid modulated amplifier circuit in which the carrier voltage is applied to the control grid and the signal voltage to the screen grid.

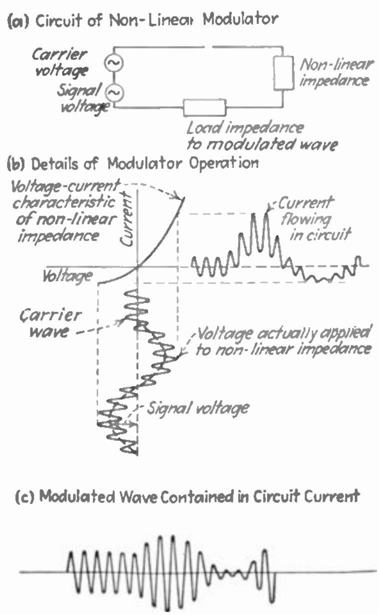


FIG. 193.—Circuit of modulator utilizing a non-linear circuit element, together with details showing mechanism of operation. The signal voltage varies the impedance offered to the carrier voltage and thus causes the current that flows to be a modulated wave.

The grid-modulated Class C amplifier consumes practically no energy from the signal but gives greater amplitude distortion than does plate modulation, while the plate efficiency and hence the power output per tube are also lower than in the case of plate modulation.

Modulated Screen-grid Amplifier.—The only types of modulated amplifiers in addition to the Class C and van der Bijl types that are of importance are those which employ screen-grid tubes. Screen-grid modulated amplifiers of the grid-modulated Class C and van der Bijl types are similar in all respects to the corresponding triode modulators, since the amplification of both the three- and four-electrode tubes depends in the same way upon the control-grid potential. The amplification of a screen-grid tube however also depends upon the screen-grid voltage, so that a modulated wave can be obtained by applying the carrier to the control grid and superimposing the signal voltage upon the direct-current screen-grid potential as shown in

Fig. 192. Such a screen-grid modulated amplifier is occasionally used in laboratory equipment where a high degree of modulation is not required.

Modulators Employing a Non-linear Impedance.—Modulators utilizing non-linear impedances operate by superimposing the signal and carrier voltage upon each other and applying the combination to a circuit element having an impedance that depends upon the applied voltage. The action that takes place is shown in Fig. 193 and in a broad way can be summarized by stating that the carrier frequency current that flows depends upon the amplitude of the superimposed signal-frequency voltage, with the result that a modulated wave is produced. The only essential difference between the non-linear impedance and van der Bijl modulators is that the latter involves amplification. Non-linear modulation is often unintentionally produced in vacuum-tube amplifiers by the curvature of the plate-voltage plate-current characteristic. The only practical application of the non-linear circuit-element method of modulation is in the grid-current modulator shown in Fig. 194,[1] which makes use of the non-linear resistance that exists between grid and cathode of a vacuum tube operated at zero grid bias. The signal and carrier voltages are applied to the grid of the tube, and as a result of the non-linear grid-cathode tube resistance, a modulated carrier wave flows in the grid circuit. This wave is forced to flow through the grid impedance Z_g, across which there is produced a voltage drop that is amplified in the plate circuit of the tube by ordinary amplifier action.

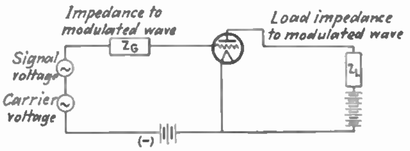


Fig. 194.—Grid-current modulator, in which the non-linear relation between grid current and grid voltage is used to produce a modulated wave of grid current, that in flowing through the series grid impedance produces a voltage drop which is then amplified by the tube.

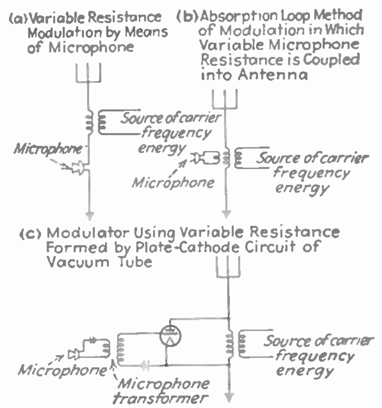


Fig. 195.—Representative modulation methods employing a circuit element having an impedance that depends on the signal amplitude.

Modulation by Means of a Variable Impedance.—A number of methods of modulation which operate by varying an impedance in accordance with the intelligence to be transmitted are shown in Fig. 195. The circuit shown at Fig. 195*a*, which is probably the first modulation system

[1] See Eugene Peterson and Clyde R. Keith, Grid Current Modulation, *Bell System Tech. Jour.*, vol. 7, p. 106, January, 1928.

ever devised, operates by varying the resistance of the antenna circuit in accordance with the resistance of a carbon-type microphone. A modification of this circuit is shown at Fig. 195*b*, in which the microphone is inductively coupled to the antenna system. Both of these arrangements can be used to modulate small amounts of power but at best are rather unsatisfactory. An improvement is shown at Fig. 195*c*, in which the microphone output is used to vary the plate-cathode resistance of a vacuum tube associated with the antenna circuit. Another type of variable-impedance modulator makes use of the fact that the inductance which a coil having a magnetic core offers to an alternating voltage depends upon the direct-current saturation, so that if the saturation is varied in accordance with the signal, the inductance to the superimposed alternating current, and hence the magnitude of the resulting radio-frequency current, will vary likewise. Such modulators can be used only at low frequencies and are inferior to the numerous arrangements employing vacuum tubes. Another type of magnetic modulator makes use of the fact that the core loss, and hence coil resistance, depends upon the direct-current saturation, so that if the saturation is varied in accordance with the signal the resistance which the coil offers to a high-frequency current will vary likewise, and a modulated wave will result. Modulators based on this principle, which is known as the "flutter effect," can be operated with some success on frequencies as high as 1000 kc, but are now obsolete.

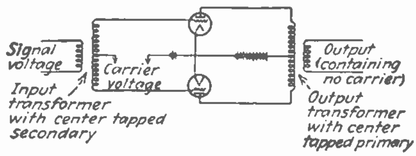


FIG. 196.—Balanced modulator circuit arranged to suppress the carrier wave from the output without altering the side bands.

88. Carrier-suppression and Single Side-band Systems of Communication.—The carrier component of a modulated wave has a frequency, amplitude, and phase that are not affected by the presence or absence of modulation, and so contains no part of the intelligence being transmitted. The carrier wave therefore need not be transmitted, since it can be reproduced at the receiving station by the use of a suitable oscillator. The carrier wave can be readily suppressed by the balanced modulator circuit shown in Fig. 196. The carrier voltage is applied to the two modulator tubes in the same phase, as shown in the figure, while the signal voltage is applied in opposite phase to the two grids by means of the center-tapped transformer. The outputs in the plate circuits of the two tubes are combined through a transformer with a center-tapped primary in such a way that voltages applied to the two grids in the same phase cancel each other in the output, while voltages applied to the two grids 180° out of phase are added in the output. The result is that the carrier voltage, which is applied to the two tubes in the same phase, does not appear in the secondary of the output transformer, whereas the side-band components, which are produced in opposite phase, are added to give an output that is a modulated wave from which the carrier component has been removed. Since two-thirds of the power of a completely modulated wave is contained in the carrier component, the suppression of this part of the wave gives a power saving of two-thirds or more, depending on the degree of modulation. At the same time the problem of replacing the suppressed carrier at the receiver by a local oscillator which has exactly[1] the required frequency is so difficult that carrier-suppression systems of communication are seldom used.

Each side band taken alone contains all of the information present in a modulated wave since for every component of the signal there is in each side band a component having a corresponding amplitude and frequency. It is therefore possible to carry on communication by transmitting only a single side band and suppressing the carrier and other side band at the transmitter. The single side band is obtained by first suppressing the carrier by some such arrangement as shown in Fig. 196 and then passing the resulting side bands through filter circuits which are sufficiently selective to transmit one side band while suppressing the other. At the receiving point the single side band is combined with a locally generated oscillation having a frequency as close as possible to that of the suppressed carrier. This local oscillation heterodynes with the side-band frequencies, developing difference frequencies that reproduce the signal originally modulated on the transmitted wave. Unlike the situation that exists in carrier-suppression transmission, no serious distortion results if the locally generated carrier used in the reception of a single side band does not have exactly the correct frequency.

The single side-band system of communication is able to transmit a given signal with a frequency band only half as wide as that required by a modulated wave consisting of two side bands and a carrier, and also saves at least two-thirds in power because of the suppression of the carrier. Single side-band transmission is extensively used in carrier-current communication over wire lines, but the difficulty of producing large amounts of single side-band power at radio frequencies, and the frequency-stability requirements which the receiver's local oscillator must meet when the suppressed carrier is a high radio frequency have prevented single side-band transmission from being standard practice in radio work.[2]

[1] See R. V. L. Hartley, Relations of Carrier and Side Bands in Radio Transmission, *Proc. I.R.E.*, vol. 11, p. 34, February, 1923.

[2] The only commercial radio circuit using single side-band transmission is the long-wave transatlantic radio telephone which is described in R. A. Heising, Production of Single Side Band for Transatlantic Radio Telephony, *Proc. I.R.E.*, vol. 13, p. 291, June, 1925; A. A. Oswald and J. C. Schelleng, Power Amplifiers in Transatlantic Radio Telephony, *Proc. I.R.E.*, vol. 13, p. 313, June, 1925.

The waves that are sent out by carrier-suppression and single side-band systems of communication differ in appearance from a modulated wave in several respects as is apparent from Fig. 197, which shows the same signal transmitted by amplitude-modulation, carrier-suppression, and single side-band systems. The wave with carrier suppression differs from the modulated wave primarily in having an envelope that varies in amplitude at twice the modulation frequency as a result of heterodyne action between the two side bands. The wave representing a single

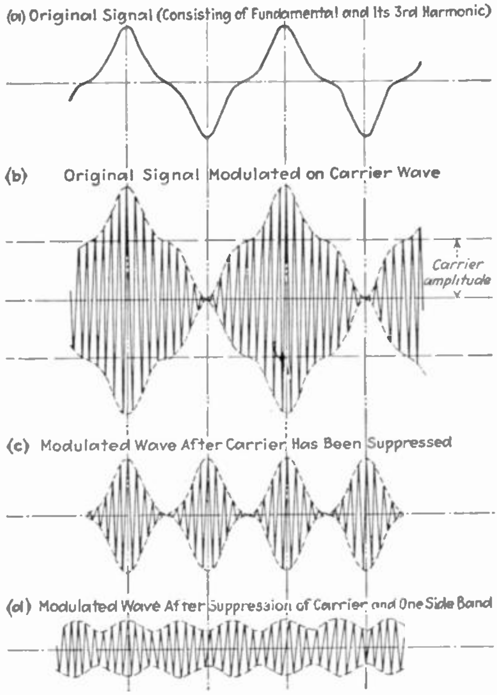


FIG. 197.—Character of waves produced when carrier is suppressed, and when carrier and one side band are both suppressed. When the carrier is removed the envelope varies at a frequency that is twice that when the carrier is present, while when only a single side band is present the envelope varies in accordance with the difference frequencies formed by the components of the original signal, which in the case shown consisted of a fundamental wave and its third harmonic.

side band consists of a number of frequency components, one for each component in the original signal. Each of these components has an amplitude proportional to the amplitude of the corresponding signal component and a frequency differing from that of the carrier by the signal frequency. The result is that in a general way the envelope amplitude of the single side-band signal increases with the degree of modulation and varies in accordance with the difference frequencies formed by the various frequency components of the single side-band heterodyning with each other.

89. Frequency and Phase Modulation.—Instead of transmitting intelligence by varying the amplitude of the radiated wave it is also possible to carry on communication by keeping the amplitude of the wave constant and varying the frequency in accordance with the signal to be transmitted. This is known as frequency modulation and results in a radiated wave having the appearance shown in Fig. 198*b*. The extent of the frequency variation in such a wave is made proportional to the amplitude of the signal, while the rate of frequency variation, *i.e.*, the number of times the frequency is changed between the minimum and maximum values per second, corresponds to the modulation frequency in amplitude modulation. Thus if a 500-cycle sound wave is to be transmitted by frequency modulation of a 1,000,000-cycle carrier wave this

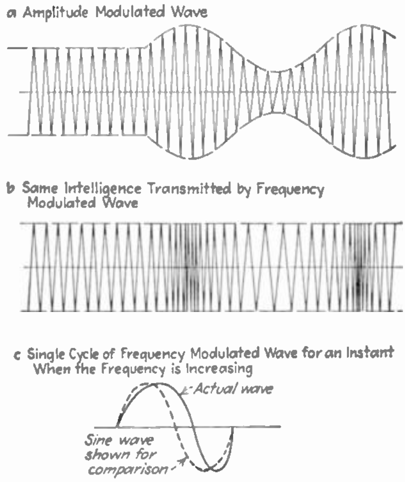


FIG. 198.—Character of waves produced by frequency modulation, together with large-scale reproduction of a single cycle, showing how the wave shapes are not sinusoidal.

could be done by varying the transmitted frequency between 1,000,010 and 999,990 cycles, 500 times a second. If the pitch of the sound wave was increased to 1000 cycles, the carrier frequency would be varied between the same two limits 1000 times a second, while a sound wave of twice the intensity would be transmitted by varying the carrier frequency through twice the frequency range, *i.e.*, from 1,000,020 to 999,980 cycles in the above case.

Frequency-modulated waves can be readily produced by varying the capacity of the oscillator tuned circuit. The simplest way of doing this is to employ a small auxiliary condenser, one plate of which is a thin diaphragm that is vibrated by the voice currents in the same manner as is the diaphragm of a telephone receiver. Reception is accomplished by converting the frequency-modulated wave into an amplitude-modulated wave by detuning the resonant circuits of the receiver so that the received frequency falls slightly to one side of the resonant point. As the frequency of the modulated wave varies, the response of the tuned circuit becomes alternately large and small thus producing amplitude modulation.

Analysis of the Frequency-modulated Wave.—A superficial examination of frequency modulation might lead one to believe that intelligence could be transmitted in this way with an extremely narrow frequency band, since in the case cited above it appears that only 20 cycles band width is required to transmit the 500-cycle sound wave. This is not correct, however, because the variation in the frequency prevents the individual cycles from being exactly sinusoidal in shape. This is illustrated in Fig. 198*c* where it is apparent that since the changing frequency causes the time required to complete one quarter cycle to differ from the time required by the next quarter cycle, the actual wave contains more than a single frequency. In fact, exact analysis shows that the frequency-modulated wave not only contains the same side-band frequencies as does the amplitude-modulated wave but also has higher order side bands that differ from the carrier frequency by integral multiples of the modulation frequency. Thus when a carrier wave of frequency f_o is frequency modulated at a rate of f_s cycles per second, the resultant wave contains components having frequencies of f_o, $f_o + f_s$, $f_o - f_s$, $f_o + 2f_s$, $f_o - 2f_s$, $f_o + 3f_s$, $f_o - 3f_s$, etc.

The exact nature of a frequency-modulated wave can be determined by writing down the equation giving the instantaneous wave amplitude and then determining the frequency components contained in the result. The frequency-modulated wave can be readily shown to be expressible by the following equation[1]

[1] See Hans Roder, Amplitude, Phase, and Frequency Modulation, *Proc. I.R.E.*, vol. 19, p. 2145, December, 1931; Balth. van der Pol, Frequency Modulation, *Proc. I.R.E.*, vol. 18, p. 1194, July, 1930; John R. Carson, Notes on the Theory of Modulation, *Proc. I.R.E.*, vol. 10, p. 57, February, 1922.

The equation of the frequency-modulated wave is derived as follows: If the current is defined by the relation $i = A \sin \phi$, then the frequency at any instant is $(d\phi/dt)/2\pi$, and $\phi = \Phi + \int 2\pi f dt$, where Φ is an arbitrary phase constant. In the case of frequency modulation, the instantaneous frequency f is $f = f_o(1 + k_f \cos vt)$ where k_f is a constant. Substituting this in the integral giving ϕ and denoting $(\omega t + \Phi)$ by $\omega_o t$ results in

$$i = A_o \sin (\omega_o t + m_f \sin vt)t$$

where

$$m_f = \frac{\text{variation in radio frequency away from the mean}}{\text{audio frequency}}$$

This can then be rewritten as

$$\begin{aligned} i = A_o[I_o(m_f) \sin (\omega_o t) + I_1(m_f)[\sin (\omega_o + v)t \\ - \sin (\omega_o - v)t] + I_2(m_f)[-\sin (\omega_o + 2v)t + \sin (\omega_o - 2v)t] + \cdots] \end{aligned}$$

where $I_n(x)$ means the Bessel function of the first kind and the nth order.

$$i = I_m \sin (\omega t + \Phi + m_f \sin vt)$$

where ω and v are 2π times the carrier and audio frequencies, respectively, Φ an arbitrary phase constant, and m_f the ratio

$$\frac{\text{Variation in carrier frequency about average carrier frequency}}{\text{Audio frequency}}$$

By making use of Bessel's functions it can be readily demonstrated that this wave consists of a carrier wave plus a series of side bands such as described above. Since Bessel's functions are not familiar to most engineers the results which they lead to have been plotted in Fig. 199, which shows the amplitude of the first-, second-, and third-order side-band components, *i.e.*, the side bands that differ from the carrier frequency by f_s, $2f_s$ and $3f_s$. These curves show that when the modulation index m_f is less than one, *i.e.*, when the range through which the frequency is varied is less than the audio frequency, the amplitude of the first-order side band is approximately proportional to the modulation index, while the higher order side bands are practically negligible. When the modulation index exceeds unity, *i.e.*, when the range through which the radio frequency is varied is greater than the signal frequency, the second and other higher order components become of importance while the carrier amplitude drops rapidly and may even be zero. When the modulation index exceeds unity there will be side-band components of appreciable magnitude extending on either side of the carrier up to the extreme limits between which the radio frequency is varied.

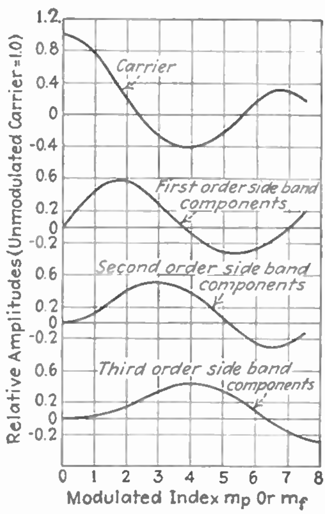


FIG. 199.—Amplitudes of frequency components of a frequency- or phase-modulated wave assuming that the amplitude of the unmodulated wave is 1.0. In the case of the side bands the amplitude shown is the amplitude of the individual side-band component and not of the pair of companion side bands taken together.

Frequency modulation is not particularly satisfactory as a means of transmitting intelligence. The frequency band is at least as great as that employed with amplitude modulation and is in general somewhat greater. Furthermore, the amplitude of the side-band components depends upon the signal frequency, so that the modulation index m_f is inversely proportional to the signal (audio) frequency. Thus, if the amplitude of the audio frequency is kept constant while the frequency is varied, the modulation index and hence the amplitude of the intelligence-carrying side bands will decrease as the signal frequency becomes greater, thus introducing distortion. The only advantage of frequency modulation is the ease with which it can be applied to high-power short-wave transmitters.

Phase Modulation.—In phase modulation the intelligence is transmitted by varying the phase rather than the frequency of the radio wave. While no satisfactory method has been devised for the reception of such waves, phase modulation is of importance because it often occurs as an unwanted by-product of other methods of modulation. The equation of a phase-modulated waye is as follows

$$i = I \sin (\omega t + \Phi + m_p \sin vt)$$

where I is the amplitude of the wave, ω and v are 2π times the radio and audio frequencies, respectively, Φ is an arbitrary constant, and m_p the angle in radians through which the phase is displaced about the average phase. This equation is seen to be identical with that for the frequency-modulated wave, the only difference being in the interpretation of the modulation index m, which in the case of phase modulation depends only on the amplitude of the modulation and is independent of the frequency of the audio signal. It is hence apparent that the phase-modulated wave contains the same side-band components as does the frequency-modulated wave, and if the modulation indexes in the two cases are the same the relative amplitudes of these different components will also be the same. As long as the modulation index is less than unity (*i.e.*, phase shifts less than 57.3°), only the first-order side-band components are of importance, but each additional 57.3° of phase shift will add another pair of important side-band components. The relative amplitude of the carrier and the first three side bands can be obtained from Fig. 199 for any given modulation index.

Combinations of Phase, Amplitude and Frequency Modulation.—Frequency and phase modulation are often combined with amplitude modulation as undesirable by-products. For example in the plate-modulated oscillator the plate-supply voltage of the oscillator tube is varied in accordance with the intelligence being transmitted, and since the generated frequency depends more or less upon the plate voltage the oscillations actually generated possess both frequency and amplitude modulation. Again in modulated amplifiers it is possible for the modulation process to produce a phase shift that varies with the signal voltage, thus producing phase modulation.

When phase modulation is combined with amplitude modulation the principal effect is to add additional side-band components that are not necessary for the transmission of the intelligence and which may produce interference with other communications. If the receiver does not discriminate against these higher order side-band components the envelope of the received signal will not be affected by the phase modulation, and with linear detection no distortion will result. If square-law detection is employed, however, or if some of the higher order side-band components are suppressed, the received signal will no longer be a true reproduction of the original audio signal.

When frequency modulation occurs with amplitude modulation the situation is somewhat less unfavorable than when phase modulation is present. This results from the fact that the modulation index m_f is inversely proportional to the audio frequency, so that at the higher audio frequencies the side-band components produced as a result of frequency modulation have relatively small amplitude. At low signal frequencies, where the modulation index m_f is large, the signal frequency is so low that the second- and third-order side bands lie well within the response band of the receiver and so are not suppressed by the receiver and do not produce interference with adjacent channels.

When phase, amplitude, and frequency modulation are combined the result at high audio frequencies is substantially that of phase and amplitude modulation alone, since the modulation index of frequency modulation is then small. On the other hand at the lower audio frequencies the combined result is very much the same as though there was only amplitude and frequency modulation, since here the modulation index m_f becomes so large as to make the phase-modulation index negligible by comparison.

CHAPTER XI

SOURCES OF POWER FOR OPERATING VACUUM TUBES

90. Cathode Heating Power.—The cathodes of vacuum tubes are heated with commercial alternating-current power whenever possible because of the convenience and low cost of such electrical energy. Alternating current cannot always be employed with filament-type tubes, however, and in some applications it is necessary to operate the filament on direct current or to use heater-type tubes.

Direct-current power for filament heating is usually obtained from either storage or dry batteries. Storage batteries are more economical and so are used whenever possible, while dry batteries find their chief usefulness in portable equipment and where power for recharging storage batteries is not available. Direct-current generators, and direct-current obtained by rectifying and filtering alternating current, are also occasionally used to heat the filaments of vacuum tubes.

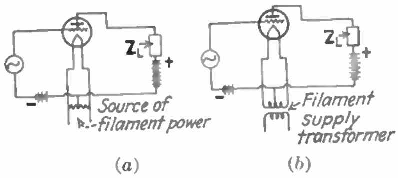


FIG. 200.—Methods of connecting the grid and plate return leads when alternating filament current is used.

When alternating current is supplied to the filament of a vacuum-tube amplifier which has been arranged for operation with direct-current filament power, the effect is to introduce into the amplifier output alternating currents of the frequency of the filament current and of twice this frequency. Furthermore any signal voltage that is being amplified by the tube will also be modulated slightly in accordance with the same two frequencies. The sum total of these effects is referred to as "alternating-current hum" since in radio sets and similar equipment the result is a low-pitched hum in the loud-speaker.

The part of the alternating-current hum that has the same frequency as the filament supply current can be entirely eliminated by bringing the grid and plate return leads to a point having the same potential as the mid-point of the filament. This is accomplished either by connecting the returns to the mid-point of a resistance located across the filament, as shown in Fig. 200*a*, or to a center tap on the filament transformer, as shown in Fig. 200*b*. Alternating-current operation of filaments in this way always calls for a slightly greater grid bias than is required for corresponding direct-current operation because the grid return lead is brought to a point that is positive with respect to the negative end of the filament.

Causes of Residual Alternating-current Hum from Alternating-current Filament Current.—After the fundamental frequency component of the alternating-current hum has been eliminated by one of the arrangements shown in Fig. 200, there remains a residual hum that has a frequency twice that of the filament current. This double frequency hum can arise from the cyclical variation of filament temperature, the effect which the alternating magnetic flux, set up by the filament current, has on the space current of the tube, and the effect which the voltage drop in the filament has on the space current. Since the heat generated in the filament at any instant is proportional to the square of the instantaneous filament current, the filament temperature will pulsate at twice the frequency of the filament current. The heat capacity of filaments used in vacuum tubes is so high, however, that the resulting variation in filament temperature is very small with 60-cycle filament current and produces negligible hum when temperature saturation is present.[1] The magnetic field produced by the filament current deflects the electrons flowing to the anode according to the principles discussed in Sec. 24 and causes the plate current to be slightly larger when the filament current is zero than when the current is at either a positive or a negative maximum. The voltage drop in the filament causes the negative half of the filament to supply more electrons to the anode than does the positive half, and since the number of electrons drawn from the filament is proportional to the three-halves power of the electrostatic field, the current from the negative end of the filament is increased more by the filament drop than the current drawn from the positive end is decreased. The total space current of the tube is hence slightly greater when there is a voltage drop in the filament than when the entire filament is at the same potential.

An exact analysis of the effect of the voltage drop in the filament shows that the resulting hum acts exactly as though it were produced by an alternating voltage applied to the grid of the tube, and that this equivalent alternating voltage that can be considered as being applied to the grid is directly proportional to the square of the alternating filament voltage and inversely proportional to the effective anode voltage $\left(E_g + \frac{E_p}{\mu}\right)$, where the grid and plate voltages E_g and E_p are measured with respect to the center point of the filament. If the hum voltage is to be kept low it is therefore apparent that a low-voltage filament is essential and that the effective anode voltage $\left(E_g + \frac{E_p}{\mu}\right)$ should be high,

[1] A detailed analysis of the factors producing the double frequency hum is given by W. J. Kimmel, The Cause and Prevention of Hum in Receiving Tubes Employing Alternating Current Direct on the Filament, *Proc. I.R.E.*, vol. 16, p. 1089, August, 1928.

i.e., that the plate current should be relatively large. The hum voltage produced by the voltage drop in the filament can be reduced to a considerable extent by using an inverted V-type of filament. This construction places the opposite ends of the filament in close proximity, causing the positive end to attract electrons from the negative leg with the result that the total space current is not increased as much by the voltage drop in the filament as would be the case if all of the electrons emitted from the filament went to the anode.

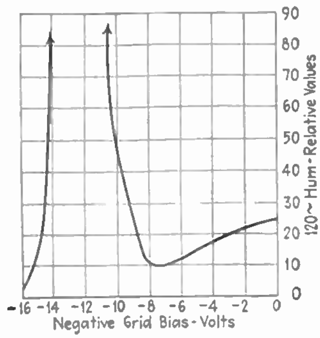


FIG. 201.—Variation of hum with grid bias in tube designed for alternating-current filament heating. At the proper grid bias the effects of the voltage drop in the filament and the magnetic flux of the filament current cancel each other, and the residual hum is very small.

The effects produced by the filament's magnetic field and by the voltage drop of the filament are of opposite phase, since the magnetic flux produced by a large filament current tends to reduce the anode current, while the voltage drop in the filament that accompanies this large current tends to produce an increased anode current. It is possible to take advantage of this canceling effect to give a very low residual hum by using a low-voltage inverted V filament and giving the open end of the V the proper spacing to make the hum produced by the voltage drop in the filament balance out the hum resulting from the magnetic effect. The extent to which the hum can be reduced in this way is illustrated in Fig. 201 which shows that the balance between the two factors depends on the grid and plate potentials and can be made very good if the operating conditions are properly chosen.

Use of Alternating-current Power under Practical Conditions.—The use of alternating-current filament power is most successful in oscillators, power amplifiers, and radio-frequency amplifiers. Alternating filament current has negligible effect on the output of an oscillator and increases the filament life by distributing the direct-current plate current between the two sides of the filament better than when direct-current filament current is used. Power amplifiers operate satisfactorily with alternating-current filament current because the high effective anode voltage of such tubes makes the hum voltage relatively small provided a low-voltage filament is employed, and because the signal voltage applied to the grid of a power tube is always relatively large and so tends to drown out the small hum that may be present. Alternating-current power is also satisfactory for heating the filaments of radio-frequency amplifiers because the only effect of the hum voltage in such cases is to modulate the radio-frequency voltage by the van der Bijl method, and as the hum voltage is small the degree of hum modulation is almost insignificant.

Audio-frequency amplifier and detector tubes operating at low power levels must be of the heater type if alternating current power is to be used in heating the cathode. This is because the low power level makes the ratio of hum to signal large. Even with heater-type cathodes hum is sometimes introduced by the alternating-current heater current as a result of incomplete shielding of the heater. Hum from this source is eliminated either by connecting one side of the heater to the cathode or by making the heater slightly positive with respect to the cathode.

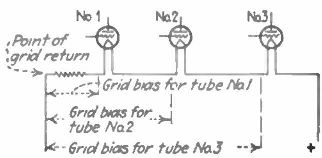


FIG. 202.—Circuit arrangement for obtaining a negative grid-bias voltage from the filament battery by operating the tubes in series.

91. The Grid-bias Voltage.—Dry batteries form the simplest method of obtaining the negative bias voltage required by the grids of vacuum tubes. Since the negative grid draws no electrons, such batteries, which are often called C batteries,[1] need have only a small capacity. The usual grid bias or C battery consists of a number of small dry cells assembled in a moisture-proof carton, and, when properly made, a life of several years can be expected.

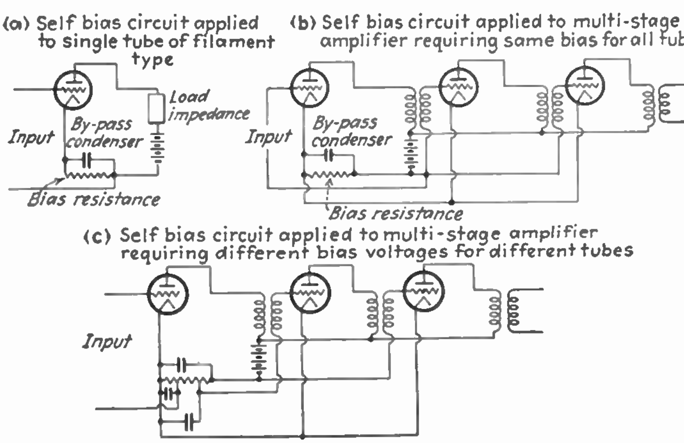


FIG. 203.—Self-bias circuits for obtaining a negative grid bias with filament-type tubes by utilizing the voltage drop produced across a resistance placed in the plate return lead.

When the filaments of several tubes are connected in series and heated by direct current power it is possible to use the voltage drop in the

[1] The terms A, B, and C to indicate filament, plate, and grid-bias batteries, respectively, originated when vacuum tubes first began to be used. The first tube circuits required filament and plate batteries, and in order to differentiate clearly between the two batteries the early instructions designated the filament battery as battery A (*i.e.*, Battery 1) and the plate battery as Battery B (*i.e.*, Battery 2), and these letters ultimately came to be used as the name of the battery. Later when the importance of the grid-bias battery was discovered it was naturally called a C battery (*i.e.*, Battery 3).

filaments and rheostat as a grid bias by bringing the grid return lead to the negative battery terminal as is shown in Fig. 202. In such an arrangement the available bias is greatest for the tube next to the positive side of the filament battery and least for the tube next to the negative side. It is sometimes possible to obtain all necessary grid-bias voltages in this way.

Self-bias Arrangements.—When the cathode of a tube is heated by alternating current, and the plate power is obtained from rectified alternating current, it is desirable to obtain the grid bias by means other than batteries. In filament-type tubes this can be done by placing a resistance in the negative plate-supply lead and connecting the grid return lead as shown in Fig. 203*a*. In this way the grid is made more negative than the cathode by the voltage which the plate current develops across the resistance. When a number of tubes requiring the same grid bias are to be operated together the arrangement shown in Fig. 203*b* is employed, while if different tubes must have different bias voltages these can be obtained by tapping the resistance as shown in Fig. 203*c*.

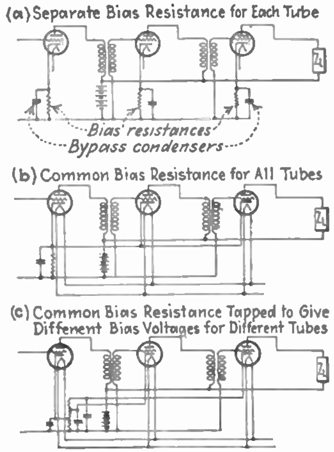


FIG. 204.—Self-bias circuits for obtaining a negative grid bias with heater-type tubes by utilizing the voltage drop across a resistance in series with the cathode to make the cathode positive with respect to the grid return lead.

The grid-bias voltage for heater-type tubes can be obtained by joining the grid and plate return leads, and then connecting the cathode to this junction through a suitable resistance, as shown in Fig. 204. In this way the cathode is made more positive than the grid, *i.e.*, the grid is made negative with respect to the cathode, by a voltage equal to the drop produced by the plate current flowing through the bias resistance. When a number of heater-type tubes are involved it is possible to use a separate bias resistance for each tube, as shown in Fig. 204*a*, or to connect all cathodes and grids in parallel and use a single resistance as in Fig. 204*b*. When the different tubes require different grid-bias voltages these can be obtained either with a separate resistance for each tube or by a single tapped resistance as illustrated in Fig. 204*c*.

When an anode detector is self-biased by the arrangement shown in either Fig. 203*a* or Fig. 204*a*, the detector characteristics are quite different from those obtained with a bias derived from a battery because the direct-current plate current of an anode detector is roughly proportional to the carrier voltage being rectified. The result is that the bias developed by the resistance depends upon the amplitude of the carrier wave being rectified, and this fact must be taken into account in proportioning the circuit.

Regeneration in Self-biased Tubes.—The self-bias arrangements of Figs. 203 and 204 all have a tendency to produce regeneration because the amplified signal currents flowing in the plate circuit produce a voltage drop across the bias resistance, and this drop is applied to the grids of the tubes. In order to minimize this effect the bias resistance is always shunted by a large condenser, as shown in Figs. 203 and 204, in order to short-circuit the bias resistance to alternating currents and prevent an alternating-current voltage from developing across it. This expedient is very effective at the higher audio and at radio frequencies where the capacitive reactance of the condenser is very small, but fails at low audio frequencies where the capacitive reactance is large.

Regeneration resulting from a grid-bias resistance acts exactly as does regeneration produced by a common plate impedance (see Sec. 40) and can be analyzed in the same way. Considering first the case where each tube is biased separately, as in Figs. 203*a* and 204*a*, it is readily demonstrated that[1]

$$\left.\begin{array}{l}\text{Actual amplification taking into}\\ \text{account regeneration from self bias}\end{array}\right\} = \frac{A}{1 + \dfrac{\mu Z_c}{R_p + Z_L + Z_c}} \qquad (144)$$

where

A = vector amplification obtained without regeneration
μ = amplification factor of tube
R_p = plate resistance of tube
Z_c = impedance between terminals of bias resistance (impedance of combination formed by bias resistance and shunting capacity)
Z_L = load impedance in plate circuit of tube.

The regeneration is seen to become greater as the impedance Z_c across

[1] The derivation of Eq. (144) follows: The voltage e_g acting on the grid of the tube is $(e_s + e_r)$, where e_s is the applied signal, and e_r is the voltage developed across the bias impedance Z_c. The voltage e_r can be expressed in terms of e_g by making use of the equivalent plate circuit of the vacuum tube, which leads to the relation

$$e_r = -\mu e_g \frac{Z_c}{Z_L + R_p + Z_c}$$

The voltage e_g actually acting on the grid hence is

$$e_g = e_s + e_r = e_s - \frac{Z_c}{Z_L + R_p + Z_c}\mu e_g$$

Solving this for e_g/e_s gives

$$\frac{e_g}{e_s} = \frac{1}{1 + \dfrac{\mu Z_c}{Z_L + R_p + Z_c}}$$

Equation (144) now follows at once.

assumed R_p >> Z_c
(for pentodes)

the terminals of the bias resistor is increased, and may either decrease or increase the amplification depending on the phase angles of impedances Z_c and Z_L. The effects produced in a typical case of regeneration from a self-bias arrangement are shown in Fig. 205, where it is seen that when the bias resistance is not shunted by a condenser the regeneration is very pronounced at all frequencies, but with a large shunting condenser is noticeable only at low frequencies where the condenser reactance is high.

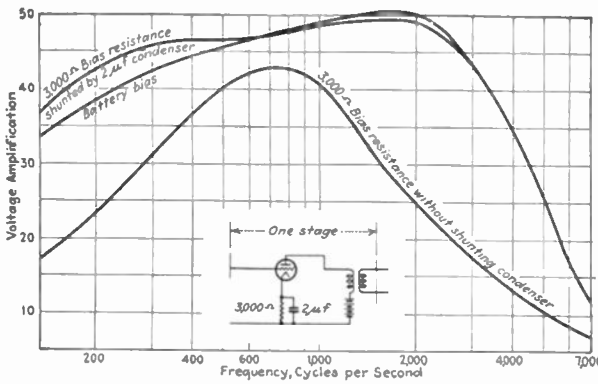


FIG. 205.—Effect of regeneration produced in an audio-frequency amplifier by self-biasing. A shunting condenser eliminates substantially all regeneration except at the lower frequencies and, in the particular case illustrated, makes the phase relations such that the regeneration improves the low-frequency response.

When a number of tubes receive bias from the same resistance, as in Figs. 203*b* and 204*b*, an exact analysis is very complicated, but for all practical purposes the only effect which need be considered is the voltage which is applied to the grid of the first tube as a result of the voltage drop developed across the bias impedance Z_c by the amplified signal currents produced in the plate circuit of the last tube. This is analogous to the corresponding assumption made in Sec. 40 in analyzing the effect of a common plate impedance, and the justification is the same. The simplified analysis of this case shows that the effect of regeneration is to change the effective amplification of the first tube according to the relation

$$\left.\begin{array}{l}\text{Amplification of first tube taking into}\\ \text{account regeneration from bias resistance}\end{array}\right\} = \frac{A_1}{1 + A_1 A_2 \mu \dfrac{Z_c}{Z_L + R_p}} \qquad (145)$$

where

A_1 = vector amplification of first stage with no regeneration
A_2 = vector amplification between grid of second tube and grid of last tube
μ = amplification factor of last tube

R_p = plate resistance of last tube
Z_L = load impedance in plate circuit of last tube
Z_c = impedance across terminals of bias resistance.

Examination of Eq. (145) shows that the regeneration increases as the total amplification A_1A_2 and the bias impedance Z_c are increased, as is to be expected.

The grid-bias voltages required by large tubes operating at direct-current plate potentials in excess of 1000 volts are frequently so large, particularly in linear amplifiers and in plate-modulated Class C amplifiers, that the usual means of obtaining the negative grid voltage are uneconomical. In such cases it is customary to use either a separate rectifier-filter system for supplying the grid bias, or a direct-current generator equipped with appropriate filter. When the rectifier-filter system is used it is necessary that a resistance be connected across the output of the filter if the tube is operated in such a way as to draw grid current, for otherwise there would be no way for this current to reach the cathode.

92. Sources of Anode Power.—The direct-current power required by the anode (*i.e.*, plate, screen grid, or space-charge grid) electrodes of vacuum tubes is usually obtained either from rectifier-filter arrangements operating on commercial alternating current, or from dry batteries. The former is used whenever possible since it is the most convenient and in the long run the more economical source of power. Dry batteries find their chief usefulness under circumstances where commercial alternating-current power is not available, or where special circumstances make it desirable to avoid the complications of the rectifier-filter arrangement.

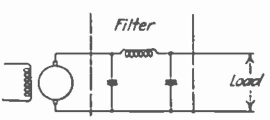


FIG. 206.—Filter circuit for eliminating ripple in output of direct-current generator.

The familiar B battery used to furnish anode power consists of either 15 or 30 small dry batteries assembled in a moisture-proof carton. The individual cells are similar in all respects except size to the usual bell-ringing and ignition cells, but special care must be taken in their construction in order to give a long shelf life (*i.e.*, the battery must not discharge itself upon standing unused for long periods).

Anode power is occasionally obtained from small storage-battery cells, but such arrangements have a high first cost and are not entirely satisfactory because the acid forms conducting films between the electrodes and allows the cells to discharge slowly.

High-voltage direct-current generators are employed to a limited extent to supply anode power for large tubes used in radio transmitters. The problems of high-voltage commutation are such, however, that machines of this type are relatively expensive and so are used only under special circumstances where the rectifier-filter arrangements are not practicable. The voltage developed by a direct-current generator always pulsates slightly as a result of the finite number of commutator bars, so that in order to obtain steady direct current from a generator it is necessary to eliminate this "ripple" by means of a filter, such as shown in Fig. 206.

93. Rectifiers for Supplying Anode Power.—As a result of a long period of evolution and development the high-vacuum thermionic rectifier and the hot-cathode mercury-vapor rectifier have become practically standard in vacuum-tube power-supply systems, while the cold-cathode gaseous rectifier, the mercury-arc rectifier, the thin-film copper oxide rectifier, and the electrolytic rectifier are used occasionally but to a diminishing extent.

High-vacuum Thermionic Rectifiers.—The high-vacuum thermionic rectifier consists of a vacuum tube containing an electron emitting cathode surrounded by an anode or plate electrode. Such a two-element tube acts as a rectifier because it will pass current only when the plate is positive with respect to the cathode, and so when placed in series with an alternating supply voltage and a load impedance will permit current to flow in only one direction.

The characteristics of the two-electrode vacuum tubes were discussed in detail in Sec. 27 and are incorporated in the curves giving the relationship between the anode current and anode voltage, such as those of Fig. 50. When the plate voltage is not too large the electrons are emitted from the cathode more rapidly than they can be drawn to the positive plate, causing a space charge to be formed in the vicinity of the cathode and making the anode current substantially independent of the cathode temperature. Under these conditions the current which the anode draws from each part of the filament is proportional to the $3/2$ power of the anode voltage with respect to that part of the filament, and when the voltage drop in the filament is negligible in comparison with the anode voltage, which is nearly always the case in high-vacuum rectifiers, the total anode current is almost exactly proportional to the $3/2$ of the anode voltage. If the anode voltage is very high however the electrons are drawn away from the cathode as fast as emitted. The plate current is then determined only by the electron emission and is independent of the plate voltage, *i.e.*, voltage saturation is present.

The important characteristics of the high-vacuum thermionic rectifier are the allowable peak plate current and the maximum allowable peak inverse voltage. The peak plate current represents the maximum electron emission which the cathode can be counted upon to supply during the useful life of the tube and is therefore determined by the cathode. Since the rectifier never allows current to flow for more than half of the time, the average plate current, *i.e.*, the direct-current output current, will never exceed one-half of the peak plate current and may be less.

The maximum allowable inverse plate voltage is the largest negative voltage that may be applied to the plate with safety, and determines the direct current voltage that can be obtained from the rectifier tube. The exact relationship between direct-current output voltage and the allowable inverse voltage depends upon the rectifier circuit employed, but in general the inverse voltage will be at least as great as the direct-current voltage and in certain rectifier connections will be π times as great.

High-vacuum thermionic rectifier tubes are constructed in much the same way as the corresponding oscillator tubes; in fact most types of rectifier tubes are merely standard filament-type three-electrode tubes with the grid omitted. The only exception to this is in the case of small rectifiers used in supplying anode power for radio receivers, where the low inverse voltages encountered permit a construction which places the plate very close to the filament. Large rectifiers are water-cooled and employ tungsten filaments just as do the water-cooled oscillator tubes, while the medium and small rectifiers are air-cooled and make use of either thoriated-tungsten or oxide-coated cathodes. The size of the filament is determined by the required maximum peak plate current, while the spacing of the electrodes and degree of vacuum determine the maximum safe inverse voltage. The losses which must be dissipated by the tube consist of the filament-heating power and the average plate loss. The average plate power will never exceed one-half the instantaneous power that is dissipated in the tube when the voltage drop between plate and filament has the value required to make the plate current equal the allowable peak value, and is usually somewhat less than the plate loss in the corresponding oscillator or power tube, because when there is no grid to shield the plate the full space current is obtained with a relatively low plate potential. The characteristics of a number of representative high-vacuum thermionic rectifiers are shown in Table X.

The high-vacuum thermionic rectifier can be built to withstand inverse voltages in excess of 100,000, and commercial tubes having peak plate currents of 7.5 amp. are available. The efficiency of the high-vacuum thermionic rectifier is high, particularly when used to develop large direct-current voltages, for the voltage drop in the tube is a relatively small fraction of the output voltage, and the cathode-heating power is only a small fraction of the output. The development of the hot-cathode mercury-vapor tube has, however, limited the field of the high-vacuum thermionic rectifier to the production of direct-current voltages greater than those which can conveniently be obtained from the mercury-vapor type of tube, and to low-power low-voltage applications (notably in broadcast receivers) where the superior ruggedness of the high-vacuum type of tube makes it more satisfactory than the hot-cathode mercury-vapor rectifier.

TABLE X.—CHARACTERISTICS OF TYPICAL HIGH-VACUUM THERMIONIC RECTIFIER TUBES

Type	Rating		Filament data			Type of filament	Type of cooling
	Maximum allowable peak plate current, milliamperes (approx.)	Maximum safe inverse voltage	Volts	Amperes	Watts		
280*	250	1,200	5.0	2.0	10.0	Oxide coated	Air
281	350	2,000	7.5	1.25	9.4	Oxide coated	Air
217-A	750	5,000	10.0	3.25	32.5	Thoriated tungsten	Air
218	750	50,000	11.0	14.75	162.0	Tungsten	Air
219	2,500	50,000	22.0	24.5	539.0	Tungsten	Air
855	500	50,000	14.5	52.0	754.0	Tungsten	Water
214	7,500	50,000	22.0	52.0	1,144.0	Tungsten	Water

* The 280 tube has two cathodes and two anodes and so is essentially two half-wave rectifiers enclosed in one envelope. The allowable plate current is given for each anode, but the filament data are for both filaments taken together.

The Hot-cathode Mercury-vapor Rectifier.[1]—The hot-cathode mercury-vapor rectifier is essentially a high-vacuum thermionic rectifier which contains mercury vapor in equilibrium with liquid mercury. The distinguishing characteristic of such a tube is that the plate is able to draw the full electron emission of the cathode when only 15 to 20 volts positive with respect to the cathode. The reasons for this behavior can be explained as follows. When the plate is more positive than the filament some of the electrons collide with mercury molecules, and if the voltage difference between plate and cathode is at least 10.4 volts, which is the ionization potential of mercury, some of these collisions will be sufficiently severe to knock electrons out of the mercury-vapor molecules (*i.e.*, will cause ionization by collision). When the plate becomes about 15 volts positive with respect to the cathode these ionizing collisions become very numerous. The electrons thus produced are attracted to the plate, but because of the low pressure of the mercury vapor (1 to 30 microns) the increase in plate current which results from the ionization is entirely negligible. The positive mercury ions which result from mercury molecules losing electrons are drawn toward the filament, but being very heavy they move much more slowly than do the electrons (only about $1/600$ as fast), so that, although the rate of production of

[1] For a more detailed discussion of tubes of this type see H. C. Steiner and H. T. Maser, Hot-cathode Mercury-vapor Rectifier Tubes, *Proc. I.R.E.*, vol. 18, p. 67, January, 1930.

positive ions is relatively slow compared with the rate of emission of electrons from the filament, the number of positive ions that are present at any one time in the interelectrode space is of the same order of magnitude as the number of electrons. The positive mercury ions therefore produce a positive space charge that is of about the same magnitude as the negative space charge produced by the electrons emitted from the filament, giving a resultant space charge that approaches zero. This neutralization of the negative space charge by the slow-moving ions permits the plate to draw the electrons from the cathode as fast as they are emitted when the plate is only 15 to 20 volts positive with respect to the cathode.

Since the positive ions eventually fall into the filament it might be thought that the filament life of hot-cathode mercury-vapor tubes would be very short. Experiments by Langmuir and Hull have shown, however, that cathode disintegration by positive-ion bombardment with mercury molecules does not take place to an appreciable extent unless the positive mercury ions have fallen through a potential that is in excess of about 22 volts. *If the plate is therefore never allowed to become more than 22 volts positive with respect to the filament the positive-ion bombardment of the filament produces no injurious effect.* It is therefore apparent that at plate voltages of 15 and 20 volts positive with respect to the cathode it is possible to produce sufficient ionization to enable the plate to draw the entire electron emission of the cathode while at the same time avoiding cathode disintegration by positive-ion bombardment.

As in the case of the high-vacuum thermionic rectifier the important characteristics of the hot-cathode mercury-vapor rectifier are the maximum allowable peak plate current and the maximum safe inverse plate voltage. The peak plate current is determined by the electron emission that can be obtained from the filament and is unaffected by the presence of the mercury vapor. The maximum safe inverse plate voltage is the sparking voltage through the low-pressure mercury vapor, and is somewhat lower than would be the case with the mercury removed.

More care must be taken in the operation of hot-cathode mercury-vapor rectifiers than is necessary with high-vacuum tubes. In the first place the bulb temperature must be maintained between definite limits because this temperature determines the pressure of the mercury vapor. If the pressure is too low (low operating temperature) the intensity of ionization is reduced to the point where the voltage drop across the tube that is needed to cause neutralization of the negative space charge exceeds the cathode disintegration value, while when the pressure is high (high operating temperature) the inverse voltage at which spark-over or flash-back takes place is reduced excessively. These effects of temperature are shown in Fig. 207. In the second place the instantaneous plate current must never be permitted even momentarily to exceed the allowable peak plate current, since in increasing the plate current there is danger that the voltage drop in the tube will reach the value at which cathode disintegration starts. Thus a momentary short circuit which would merely heat up the plate of a high-vacuum tube will cause permanent damage to a mercury-vapor tube. Finally, the filament of a hot-cathode mercury-vapor tube must be brought to full operating temperature before the plate voltage is applied, for otherwise the voltage drop in the tube during the warming-up process will exceed the cathode-disintegration value, and the filament will be permanently damaged. Where mercury-vapor types of rectifiers are employed it is customary to insert a time-delay relay in series with the rectifier plate circuit, which is thereby held open until the filament has had time to reach its operating temperature.

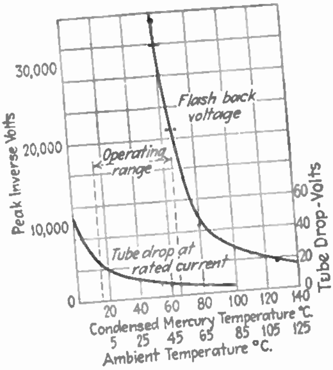


Fig. 207.—Effect of operating temperature on voltage drop and flashback voltage of a hot-cathode mercury-vapor tube. The difference between the mercury and the ambient (room) temperature represents the temperature rise of the coldest part of the tube and is determined by the design.

In connecting hot-cathode mercury-vapor rectifier tubes in parallel in order to obtain high output currents, the anodes of the two tubes should be connected to opposite ends of an inductance, to the center of which is brought the line connection. This arrangement is necessary to insure that the two tubes will carry equal anode currents. Otherwise the tube having the smallest internal voltage drop would usually carry nearly all of the current even when the difference in the voltage drops required to produce copious ionization in the tubes is only a fraction of a volt.

The distinctive constructional features of the hot-cathode mercury-vapor tube are a small plate, a large spacing between plate and filament, an oxide-coated filament requiring an extremely low voltage (never more than 5 volts) and air cooling in even the largest sizes. The small plate and air cooling are permissible because the voltage drop in the tube is so low that the anode loss is small, while the wide separation of plate and filament is made possible by the fact that the neutralization of the negative space charge by the positive ions permits a weak electrostatic field to draw the full space current. Oxide-coated filaments can be used because the low voltage drop in the tube prevents the cathode bombardment by high-velocity positive ions that is present in high-voltage high-vacuum tubes. The filaments must be designed to operate on a low voltage, for if the sum of the crest filament supply voltage and the potential required to produce copious ionization (which is 10 to 15 volts) exceeds the cathode-disintegration value, the results will be either cathode disintegration or failure of part of the filament to contribute electrons to the plate current. Typical hot-cathode mercury-vapor rectifiers are shown in Fig. 208, and the characteristics of a number of representative tubes are given in Table XI.

Compared with the high-vacuum tube, the hot-cathode mercury-vapor rectifier has the advantage of higher plate efficiency, a voltage regulation that is almost perfect, a much lower filament power, and a

FIG. 208.—Typical hot-cathode mercury-vapor rectifier tubes. Note the small plate, the ribbon-like cathode, and the comparatively large spacing between plate and cathode.

lower first cost. At the same time, the mercury-vapor type of rectifier cannot be abused without permanently damaging the cathode, and only moderately high inverse peak voltages are allowable. The hot-cathode mercury-vapor rectifier finds its chief usefulness in supplying the plate power for all but the smallest vacuum tubes, but is not capable of standing sufficient abuse to be satisfactory in broadcast receivers.

TABLE XI.—CHARACTERISTICS OF TYPICAL HOT-CATHODE MERCURY-VAPOR TUBES

Type	Rating		Filament data		
	Maximum allowable plate current, milliamperes	Maximum safe inverse voltage	Volts	Amperes	Watts
866	600	5,000	2.5	5.0	12.5
872	2,500	5,000	5.0	10.0	50.0
857	20,000	20,000	5.0	60.0	300.0

Miscellaneous Types of Rectifiers Used in Supplying Anode Power.—While the high-vacuum and mercury-vapor thermionic rectifiers are generally employed in rectifier-filter arrangements for supplying anode power, other kinds of rectifiers are occasionally used. The most important of these are the mercury-arc rectifier, the cold-cathode gaseous rectifier, the thin-film copper oxide rectifier, and the electrolytic rectifier.

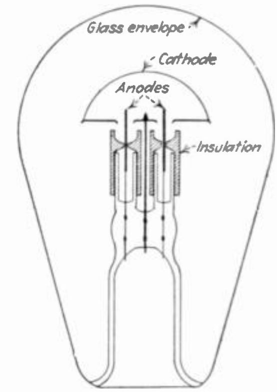


Fig. 209.—Cross section of cold-cathode gaseous rectifier. This tube has two anodes and one cathode and is intended for use in the full-wave rectifier circuit of Fig. 211*b*.

Mercury-arc rectifiers have been used to a considerable extent in Europe for supplying the plate power of large transmitting tubes. Mercury arcs are capable of withstanding very high back voltages and will readily carry any current that could possibly be required by the plate circuits of vacuum tubes, but have the disadvantage of not being self-starting and of requiring special means for keeping the arc "alive" when the power is turned on and off in code transmitters.

The cold-cathode gaseous rectifier makes use of the fact that the current which as a result of ionization flows between two electrodes immersed in a low-pressure gas is roughly proportional to the cathode area, so that if one electrode has a very small and the other a very large area, as is the case in the arrangement shown in Fig. 209, the current that flows between the two electrodes when the smaller one is positive is very great compared with the current that flows when the smaller electrode is negative. Commercial rectifiers making use of this principle were widely used at one time for supplying the plate power required by radio receivers but have been largely displaced by the high-vacuum thermionic rectifier. While the cold-cathode gaseous rectifier has the advantage of requiring no filament-heating power, its usefulness is limited by the fact that the high gas pressure limits the allowable back voltage to a comparatively low value, and by the fact that a rather high voltage drop across the tube is required to start current flowing although as soon as the path between the two electrodes becomes conducting the drop is greatly reduced. This causes the rectifier to have a poor voltage regulation and tends to develop voltage surges of considerable intensity. The successful operation of the cold-cathode gaseous rectifier depends on maintaining the gas at exactly the right pressure, and the life of such

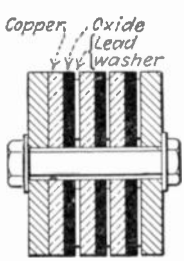


Fig. 210.—Assembly of units of copper oxide type of thin-film rectifier. The cuprous oxide film is formed directly on the surface of the copper and is an integral part of the copper washer.

tubes is terminated either by a reduction in gas pressure as a result of gas absorption by the electrodes, or by an increase of pressure resulting from gas given up by the electrodes and the glass wall.

The thin-film copper oxide rectifier makes use of the fact that when a thin film of cuprous oxide is formed upon a metallic copper surface the resistance which this film offers to electrical currents is small for currents flowing in one direction and high for currents going the opposite way. Each individual copper oxide film will stand a back voltage of only a few volts, so that it is ordinarily necessary to employ a number of films in series. This is accomplished by building up a pile of alternate rectifying and lead washers, as shown in Fig. 210, and clamping the whole together. Rectifiers of the copper oxide film type find their greatest usefulness in low, or moderately low, voltage service, and are seldom used where direct-current voltages in excess of 100 volts are desired.[1]

Electrolytic rectifiers make use of the fact that certain metals, notably aluminum, when placed in a suitable solution have films formed on their surface which permit current to pass in only one direction. In its most common form the electrolytic rectifier consists of aluminum and iron or lead electrodes placed in a bicarbonate of soda or ammonium phosphate solution. The highest peak inverse voltage that can be safely applied to an individual cell is in the order of several hundred volts, and the current that can be drawn is limited by the heating of the solution. Electrolytic rectifiers making use of a number of small cells connected in series have been employed to a considerable extent by radio amateurs in supplying power for moderate-sized transmitting equipment, but at the present time have no other important application.[2]

94. Rectifier Circuits.—The various types of rectifier connections that may be employed with a single-phase source of power are shown in Fig. 211, together with the wave form of the voltage which is developed across a resistance load.

The circuit shown at Fig. 211*a*, in which a single rectifier is placed in series with the source of alternating voltage and the load impedance, is called a half-wave rectifier circuit. It has the very great disadvantage of delivering an output voltage which is far from being a continuous direct-current potential, and of producing a direct-current magnetization in the core of the supply transformer as a result of the rectified current which flows through the secondary, and is therefore seldom used.

The rectifier circuit most commonly employed with a single-phase power source is shown in Fig. 211*b*, and consists of two rectifier units

[1] An excellent discussion of the theory of thin-film rectifiers is given by J. Slepian, Thin Film Rectifiers, *Trans. Amer. Electrochem. Soc.*, vol. 54, p. 201, 1928. Information on the performance of such rectifiers is given by L. O. Grondahl and P. H. Geiger, A New Electronic Rectifier, *Trans. A.I.E.E.*, vol. 46, p. 357, 1927.

[2] For information concerning the design of electrolytic rectifiers see "Radio Amateur's Handbook," American Radio Relay League, Hartford, Conn.

operating in conjunction with a center-tapped transformer in such a way that the two tubes alternately supply rectified current to the load, giving the so-called "full-wave" rectifier output voltage shown. Such a full-wave rectifier not only produces a direct-current voltage that is more nearly constant than does the half-wave rectifier, but also avoids direct-current saturation in the core of the supply transformer since the direct-current magnetizations in the two halves of the transformer secondary are opposed to each other and give zero resultant magnetization.

The bridge type of full-wave rectifier shown at Fig. 211*c* requires four rectifier units instead of the two called for by the center-tapped transformer arrangement, but has the advantage of requiring only one secondary winding instead of two. This circuit is seldom employed with

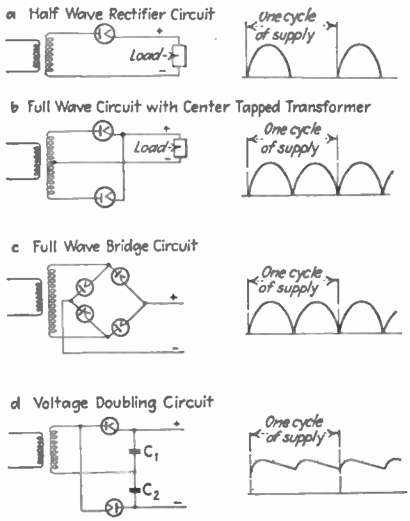


FIG. 211.—Rectifier circuits for operation with single-phase power sources, together with wave forms of voltage developed across a resistance load.

rectifiers of the thermionic type since the different cathodes are not at the same potential and so cannot be connected in parallel and supplied from a single-filament transformer secondary.

The rectifier circuit shown at Fig. 211*d* is known as the voltage-doubling circuit because it has the unique property of delivering a direct-current voltage that approaches twice the alternating voltage which is supplied by the transformer. This is accomplished because of the fact that the two condensers C_1 and C_2 are alternately charged to the full voltage of the transformer, and since the condensers are in series the output voltage can reach twice that of the alternating supply. The wave form of the output delivered by such a rectifier system depends upon the load and the size of condensers C_1 and C_2. The voltage-doubling circuit finds its chief usefulness where the required direct-current output voltage is greater than can be obtained from a single tube, as in x-ray work.

Polyphase Circuits.—When a polyphase source of alternating power is employed the number of possible rectifier connections is almost unlimited, although only a relatively few of these are of practical importance.[1] The polyphase rectifier circuits most commonly used with three-phase power sources are shown in Fig. 212 and develop voltages across a resistance load that have the wave forms indicated in the figure. The three-phase half-wave circuit is essentially three half-wave rectifiers of the type shown in Fig. 211*a*, with each leg of the secondary Y forming one phase. In such an arrangement each rectifier tube carries current one-third of the time, and the output wave pulsates at three times the frequency of the alternating-current supply. In order to avoid direct-current saturation in the transformer it is necessary to employ a three-phase rather than three single-phase transformers.

The circuit of Fig. 212*b*, which employs six tubes and two three-phase Y-connected secondaries, is essentially two three-phase half-wave rectifiers of the type shown in Fig. 212*a* connected in parallel, but with the polarities such that during the period when the output voltage of one three-phase unit is at a minimum the output of the other unit is maximum, so that the ripple in the output wave is small and has a fundamental frequency six times that of the power supply. The two three-phase units are connected in parallel through an interphase reactor (or "balance coil") which enables each three-phase unit to operate independently. If it were not for this reactor each tube would carry the load current only one-sixth of the time, whereas with the reactor each tube carries current one-third of the time, and at any instant there are always two tubes delivering current to the load. The balance coil should have sufficient inductance so that the alternating current flowing as a result of the voltage which exists across the coil has a peak value less than one-half the normal direct-current load current (*i.e.*, a peak value less than the direct current in one leg). Since the direct current flows in opposite directions in the two halves of the winding no direct-current saturation is present, and an air gap need not be provided in the core. The result is that the interphase reactor requires only a small amount of material.

The three-phase full-wave rectifier circuit shown at Fig. 212*c* gives the same output wave as does the double three-phase half-wave rectifier of Fig. 212*b* but differs in that the tubes are arranged so that full-wave rectification is obtained through each leg of the secondary winding.

[1] For a more complete list of polyphase rectifier circuits see R. W. Armstrong, Polyphase Rectification Special Connections, *Proc. I.R.E.*, vol. 19, p. 78, January, 1931. Also see D. C. Prince and F. B. Vodges, "Mercury-arc Rectifiers and Their Circuits," McGraw-Hill Book Company, Inc., New York, 1927.

This circuit requires only one three-phase secondary and no interphase reactor, but the filament transformer must have four separate secondaries.

Polyphase rectifiers are used where the direct-current power required is in the order of 1 kw. or more. Compared with the single-phase circuits, the polyphase rectifiers, particularly those of the full-wave type shown at Fig. 212*b* and *c*, develop an output voltage wave that is much closer to a steady direct-current potential than is the case with single-phase arrangements, and the more desirable polyphase circuits give a higher

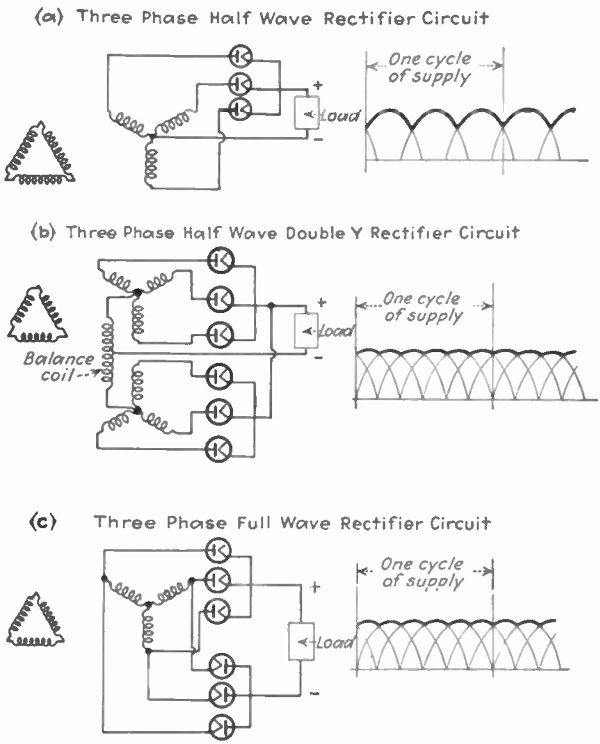


FIG. 212.—Rectifier circuits for operation with three-phase power sources, together with wave forms of voltage developed across resistance load.

output voltage in proportion to the peak inverse voltage and also utilize the possibilities of the transformer more effectively.

When thermionic rectifiers are employed the filaments are always heated by alternating current obtained from a filament transformer having a secondary, or secondaries, insulated to withstand the direct-current output voltage. In the case of small rectifiers, such as those employed in radio receivers and low-power transmitters, it is common practice to add a special secondary to the rectifier transformer for the purpose of securing filament power, but where appreciable amounts of power are involved it is preferable to employ a separate filament transformer. The number of secondaries required for filament heating ranges

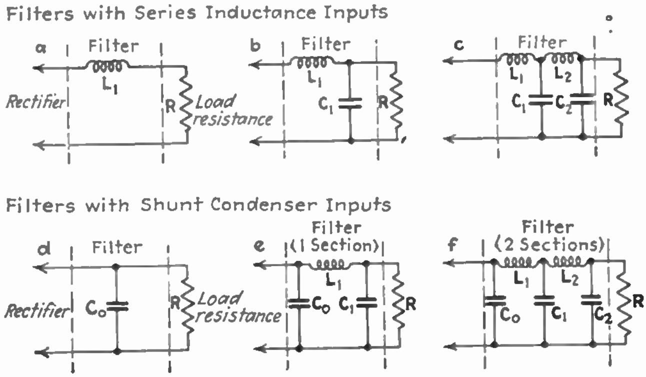


FIG. 213.—The filter circuits most commonly employed to smooth out the pulsating rectifier output into a steady direct-current voltage.

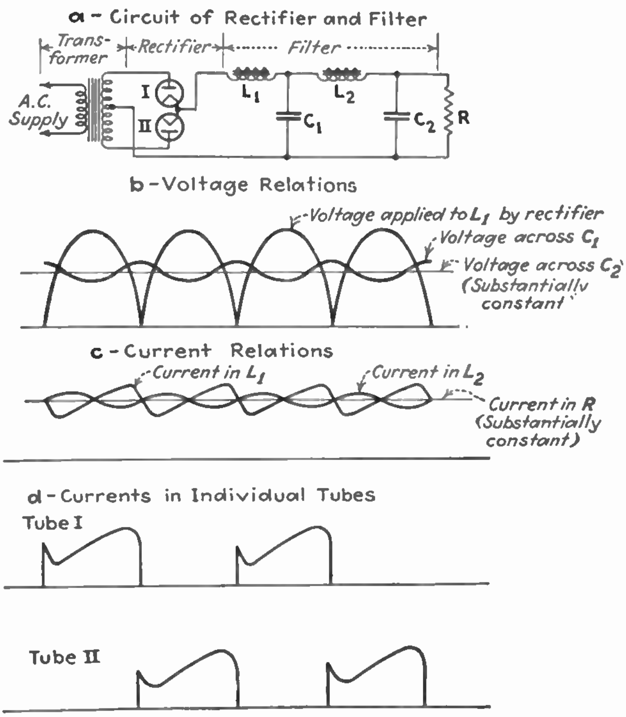


FIG. 214.—Oscillograms showing action taking place in filter having a series-inductance input when supplied power from a single-phase full-wave rectifier. These curves are idealized in that they neglect transformer leakage reactance, tube drop, and the effect of energy losses in the filter inductances and condensers.

from one in the case of the simpler circuits, to four for the three-phase full-wave circuit, where it will be noted one of the four secondaries must have sufficient capacity to operate three filaments, while the other three each supply only one filament. The filament windings are usually provided with a center tap to which the cathode connection of the rectifier is brought in order to equalize the flow of plate current in the two legs of the filament.

95. Filter Circuits Having a Series Inductance Input.—The pulsating voltage delivered by the rectifier output can be smoothed into a steady direct-current voltage suitable for applying to the anode circuit of a

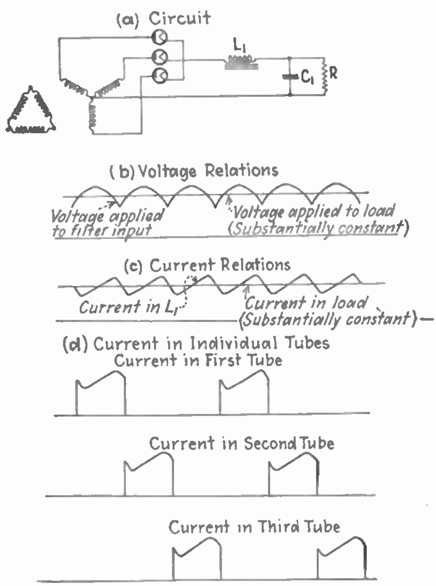


FIG. 215.—Oscillograms showing action taking place in a filter having a series-inductance input when supplied power from a three-phase half-wave rectifier. These curves neglect the effect of transformer reactance, tube voltage drop, and losses in the filter elements.

vacuum tube by being passed through an electrical network, commonly called a filter, which ordinarily consists of series inductances and shunt condensers. The most commonly used filter circuits are those shown in Fig. 213 and may be divided into two general classes according to whether the input consists of a series inductance or a shunt condenser. Each type of filter may be further subdivided according to the number of sections or filter elements involved.

The action that takes place in a properly adjusted filter having a series-inductance input can be understood from an examination of the oscillograms of Figs. 214 and 215. Consider first the case of a full-wave single-phase rectifier delivering its output to the filter of Fig. 213*c*. The current flowing into the filter tends to increase when the voltage output of the rectifier is high and tends to decrease when the rectifier voltage is low, but if the input inductance is reasonably large, as is the case in Fig.

214, these variations in current are relatively small, and to a first approximation the input current can be considered constant, with each rectifier tube carrying the full current for one-half of the time.[1] The first condenser tends to absorb what variations there are in the current entering the input inductance, with the result that the voltage that appears across the terminals of this condenser is more nearly constant than is the input current. The voltage across the first condenser is applied to the second inductance and is smoothed out still more by the action of this inductance and the second condenser, with the result that the potential appearing across the load is substantially pure direct current with a value equal to the average output voltage of the rectifier.

The action that takes place in the three-phase half-wave rectifier circuit is somewhat similar and is shown in Fig. 215. The special features to note here are that the output wave of this type of rectifier is more nearly constant than that of the full-wave single-phase type, and that each tube carries the output current only one-third of the time.

Examination of the oscillograms of Figs. 214 and 215 shows that the voltage across the first condenser C_1 is nearly constant at a value corresponding to the average voltage of the rectifier output. The potential difference across the first filter inductance hence approximates the difference between the actual instantaneous voltage of the rectifier output and the average value of this output voltage. When the voltage output of the rectifier is above the average the current tends to increase, while when it is less than the average the current through the inductance decreases. The amplitude of the current variation that results depends upon the size of the inductance in the way that is shown in Fig. 216. When the input inductance is very large the current variations are negligibly small, as shown by curve *a*, and each tube passes a wave of current that is approximately square. If the input inductance is only moderately large the input current varies as shown at *b*, and the current waves passed by the individual tubes have a ratio of average to peak value that is lower than in Case *a*. As the input inductance is reduced the current variations become larger until a point is finally reached when the rectifier ceases to draw current continuously throughout the cycle, which results in the condition shown at Fig. 216*c*. This last condition is to be avoided under normal operating conditions because it leads to a low ratio of average to peak anode current in the individual rectifier tubes and also results in poor regulation of the direct-current potential.

Analysis of Voltage Delivered by Rectifier to the Filter.—The action taking place in a filter having an input inductance of sufficient size to maintain a continuous flow of current from the rectifier can be calculated

[1] In the case of the full-wave bridge circuit there are always two rectifier tubes in series so that each tube carries the full current half of the time even though there are two tubes operating at any instant.

with an accuracy sufficient for most practical purposes by considering that the rectifier applies to the filter input a voltage having a wave shape shown by the idealized curves of Figs. 211 and 212. This neglects the leakage reactance of the supply transformer and the voltage drop in the rectifier but is justified because both of these factors are merely modifying influences in rectifier-filter systems of the type used in supplying anode power. The idealized output wave of the rectifier can be considered as consisting of a direct-current component upon which are

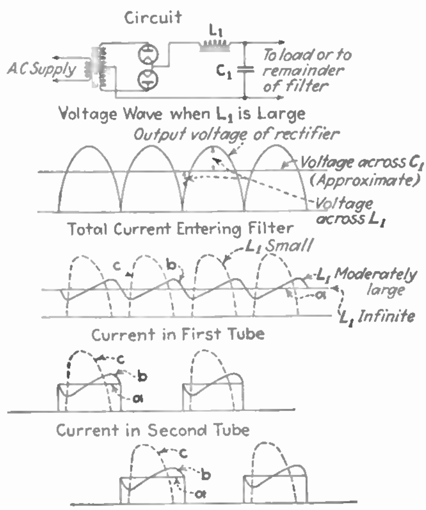


FIG. 216.—Oscillograms showing the effect of changing the size of the series input inductance when a single-phase full-wave rectifier is used.

superimposed alternating-current voltages. Thus in the case of the full-wave single-phase rectifier, the output wave has the equation

Output voltage of single-phase full-wave rectifier =

$$\frac{2E}{\pi}\left(1 - \frac{2}{3}\cos 2\omega t - \frac{2}{15}\cos 4\omega t - \frac{2}{35}\cos 6\omega t \cdots\right) \quad (146)$$

where E represents the crest value of the alternating-current voltage applied to the rectifier tube, and ω is the angular velocity $(2\pi f)$ of the supply frequency. In this case the direct-current component of the output wave is $2/\pi$ times the crest value of the alternating-current wave, the lowest frequency alternating-current component in the output is twice the supply frequency and has a magnitude that is two-thirds of the direct-current component, and the remaining alternating-current components are harmonics of this lowest frequency component. Table XII gives results of such analyses for the waves delivered by the single-phase full-wave rectifier, by the three-phase half-wave rectifier, and by the three-phase full-wave rectifier. It will be observed that in the three-phase half-wave rectifier the lowest alternating-current frequency is three times the frequency of the power supply, while in the three-phase full-wave rectifier it is six times that of the power supply. In all cases the amplitude of the alternating-current components diminishes rapidly as the order of the harmonic is increased.

TABLE XII.—CHARACTERISTICS OF RECTIFIERS OPERATED WITH A FILTER SYSTEM HAVING A SERIES INPUT INDUCTANCE

	Rectifier circuit				
	Single-phase, full-wave, center-tapped connection	Single-phase, full-wave bridge	Three-phase, half-wave	Double three-phase, half-wave	Three-phase, full-wave
Voltage Relations (Direct-current component of output voltage taken as 1.0):					
a. R.m.s. value of transformer secondary voltage (per leg)	1.11*	1.11	0.855	0.855	0.428
b. Maximum inverse voltage	3.14	1.57	2.09	2.42	1.05
c. Lowest frequency in rectifier output (F = frequency of power supply)	$2F$	$2F$	$3F$	$6F$	$6F$
d. Peak value of first three alternating-current components of rectifier output					
Ripple frequency	0.667	0.667	0.250	0.057†	0.057
Second harmonic of ripple frequency	0.133	0.133	0.057	0.014	0.014
Third harmonic of ripple frequency	0.057	0.057	0.025	0.006	0.006
Current Relations:					
e. $\dfrac{\text{Average anode current}}{\text{Peak anode current}}$	0.500	0.500	0.333	0.333	0.333
f. $\dfrac{\text{Average current per anode}}{\text{Direct-current load current}}$	0.500	0.500	0.333	0.167	0.333
Transformer Utilization Factors:					
g. Primary	0.900	0.900	0.827	0.955	0.955
h. Secondary	0.637	0.900	0.675	0.675	0.955

NOTE: This table assumes that the input inductance is sufficiently large to maintain the output current substantially constant, and neglects the effects of voltage drop in the rectifier and leakage reactance of the transformers.

* Secondary voltage on one side of center tap.

† The principal component of the voltage across the balance coil has a frequency of $3F$ and a peak amplitude of 0.250.

Calculation of Direct-current and Alternating-current Components of Filter Output.—If the voltage drop in the rectifier and the leakage reactance of the supply transformer are neglected, the direct-current voltage delivered to the load is less than the direct-current input to the filter, as calculated from Table XII, by an amount equal to the voltage drop in the resistance of the filter inductances, and the regulation of the output voltage is then the regulation that would result from the filter resistances. Actually, the leakage reactance of the supply transformer and the voltage drop in the rectifier tube cause the direct-current output voltage to be lowered somewhat and make the regulation poorer, although Table XII will give the magnitude of the direct-current output voltage delivered to the filter input with an accuracy sufficient for most purposes.

The alternating-current voltage that appears across the output of the filter is the potential which is developed across the filter output when the alternating-current voltages given in Table XII are applied to the filter input. Since the smoothing action of the filter results from the fact that the series inductances of the filter choke out these alternating-current voltages, while the shunt condensers tend to short-circuit them, the output condenser must have a reactance that is low compared with the load resistance, while each inductance must have a high reactance compared with the reactance of the condenser which immediately follows it. Furthermore the input inductance must also have sufficient reactance in relation to load resistance to satisfy Eq. (148) if current is to flow into the filter throughout the cycle.

An exact determination of the alternating-current voltage that appears across the output of the filter involves considerable labor because of the complicated electrical networks involved, but for most purposes it is permissible to simplify the analysis by assuming that the reactance of each condenser is small compared with the reactance of the inductances immediately preceding and following the condenser, and that the reactance of the output condenser is small compared with the load resistance. The fraction of the alternating-current voltage applied to the filter input that reaches the filter output is then given by the following equations.[1]

For the filter of Fig. 213*c*:

$$\frac{\text{Alternating-current voltage across load}}{\text{Alternating-current voltage applied to input}} = \frac{1}{\omega^4 L_1 L_2 C_1 C_2} \tag{147a}$$

For the filter of Fig. 213*b*:

$$\frac{\text{Alternating-current voltage across load}}{\text{Alternating-current voltage applied to input}} = \frac{1}{\omega^2 L_1 C_1} \tag{147b}$$

For the filter of Fig. 213*a*:

$$\frac{\text{Alternating-current voltage across load}}{\text{Alternating-current voltage applied to input}} = \frac{R}{\sqrt{R^2 + (\omega L_1)^2}} \tag{147c}$$

In these equations ω is the angular velocity $(2\pi f)$ corresponding to the

[1] The derivation of a relation, such as given by Eq. (147*a*), is as follows: The alternating-current current which flows in L_1 as a result of an applied voltage E is $E/\omega L_1$ if the reactance of C_1 is small compared with that of L_1. Practically all of this current flows through C_1 because of the much higher reactance of the path through L_2, so that the voltage developed across C_1 is $E/\omega^2 L_1 C_1$. This potential is then applied to L_2 and C_2, and if the reactance of C_2 is low compared with the load resistance a repetition of the method used to get the voltage across C_1 shows that the alternating-current potential across C_2 is $E/\omega^4 L_1 L_2 C_1 C_2$, and Eq. (147*a*) follows at once.

frequency of the component involved, and the alternating-current voltage applied to the input is given by Table XII for different rectifier connections. An examination of Eq. (147) shows that the filtering action increases very rapidly with the number of filter elements, *i.e.*, the number of inductances and capacities. The filter is also seen to be more effective the higher the frequency, and this, coupled with the fact that the largest component of the alternating-voltage in a rectifier output is always the one having the lowest frequency, makes it permissible to neglect all frequency components in the rectifier output except the fundamental in calculating the alternating-current voltage that will appear across the load.

Factors Involved in the Design of Rectifier-filter Systems.—The input inductance of a filter should ordinarily be of sufficient size to maintain a continuous flow of current from the rectifier under normal operating conditions, *i.e.*, the crest or peak value of the alternating current flowing in the input inductance should be less than the direct-current load current. This condition is realized by satisfying the approximate relation[1]

$$\frac{\omega L_1}{R_{\text{eff.}}} \geq \frac{\text{lowest frequency component of alternating-current voltage in rectifier output}}{\text{direct-current voltage in rectifier output}} \tag{148}$$

where ωL_1 is the reactance of the input inductance to the lowest ripple frequency, and $R_{\text{eff.}}$ is the effective load resistance, *i.e.*, the actual load resistance plus direct-current resistances of the filter inductances. The higher the load resistance, *i.e.*, the lower the direct-current load current, the more difficult it is to maintain a continuous flow of current, and with a given L_1 Eq. (148) will not be satisfied when the load resistance exceeds a critical value.

The ratio of average to peak anode current depends upon the rectifier connection and upon the size of the input inductance. Table XII gives the results for the common circuits when the input inductance is infinite. When the input inductance is finite the peak anode current is the sum of the direct-current output current and the crest value of the alternating current flowing in the input inductance. Since the largest part of this alternating current is the component of lowest ripple frequency, the following relation is approximately true, provided Eq. (148) is satisfied:

$$\frac{\text{Peak current with finite input inductance}}{\text{Peak current with infinite input inductance}} = 1 + \frac{E_1}{E_0}\frac{R_{\text{eff.}}}{\omega L_1} \tag{149}$$

[1] In order to insure a continuous flow of current in the rectifier the crest value of the alternating current flowing through the input inductance must not exceed the average or direct current. Since the direct current is equal to the direct-current voltage in the rectifier output divided by $R_{\text{eff.}}$, while the crest alternating current in the first inductance is very closely equal to the fundamental ripple-frequency voltage contained in the rectifier output, divided by the reactance wL_1 of the input inductance to this lowest frequency component of the ripple voltage, Eq. (148) follows at once.

E_1 is the amplitude of the lowest frequency component of the ripple voltage as given in Table XII, E_0 is the direct-current output voltage of the rectifier, and $R_{\text{eff.}}$ and ωL_1 have the same meaning as in Eq. (148).

The maximum inverse voltage which the rectifier tube will be called upon to withstand depends upon the rectifier connections and may vary from only slightly more than the direct-current output up to π times this potential. Values for the commonly used circuits are given in Table XII.

A transformer used to supply power to a rectifier will normally run hotter than when delivering the same amount of energy to a resistance

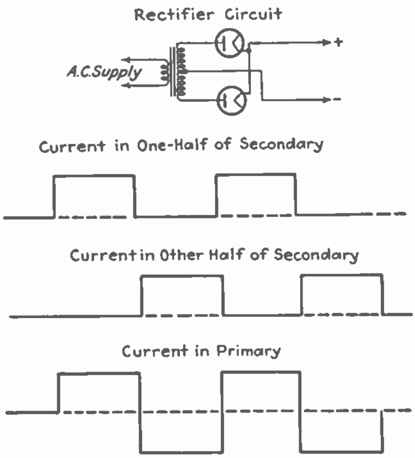


FIG. 217.—Wave shapes of current in primary and secondary windings of a center-tapped transformer supplying a single-phase full-wave rectifier operating into a filter with large series input inductance.

load because of the irregularly shaped current waves drawn by the rectifier. Thus in the case of the full-wave single-phase rectifier the current waves with a large input inductance have the shape shown in Fig. 217. The ratio of direct-current power output to the normal alternating-current rating for the same transformer losses is called the utilization factor of the transformer and depends upon the rectifier connections. Table XII gives the utilization factor of the primary and secondary windings for some of the more commonly used rectifier connections.

By making use of Eqs. (147), (149), and Table XII it is possible to design rectifier-filter combinations having an input inductance, and to calculate everything about the performance except the voltage regulation[1] with an accuracy that is sufficient for most practical purposes. The method by which this is done is illustrated by the following example:

[1] Methods have also been devised for calculating the voltage regulation, but these are so complicated as to be beyond the scope of this book. For further information see D. C. Prince and F. B. Vodges, *loc. cit.*

Example: It is desired to design a three-phase half-wave rectifier-filter system to operate from a 60-cycle power supply and to deliver a direct-current output of 2500 volts and 0.4 amp. with a ripple that must not exceed 2 per cent. If the direct-current resistance of the filter inductances is neglected the rectifier must deliver a direct-current output voltage of 2500 volts, and Table XII shows that the r.m.s. voltage which each secondary leg must develop is $2500 \times 0.855 = 2135$ volts. Since the utilization factors of the primary and secondary, as given by Table XII, are 0.827 and 0.675, respectively, each leg of the primary must have a rating of $2500 \times 0.4/(3 \times 0.827) = 403$ watts, and each leg of the secondary a rating of $2500 \times 0.4/(3 \times 0.675) = 493$ watts. Table XII shows that in order to satisfy Eq. (148), $\omega L/R_{\text{eff.}} > \frac{1}{4}$, and since $R_{\text{eff.}} = 2500/0.4 = 6250$ while $\omega = 2\pi 180$, L_1 must be not less than 1.38 henrys. Tentative calculation based on Eq. (147*b*) shows that the filter of Fig. 213*b* with $C_1 = 1.0$ μfd and $L_1 = 9.8$ henrys will keep the ripple voltage down to 2 per cent and will be generally satisfactory. Reference to Table XII and Eq. (149) shows that the peak anode current will be $0.4(1 + 0.14) = 0.456$ amp., while the maximum inverse voltage which each rectifier must stand is $2500 \times 2.09 = 5225$ volts. Type 866 mercury-vapor tubes (see Table XI) would meet these requirements. In actual practice the secondary voltage of the transformer would be made greater than 2135 volts by perhaps 10 per cent to compensate for the loss of voltage caused by the resistance of the filter inductance, the voltage drop in the rectifier, and the leakage reactance of the transformer.

96. Filter Circuits Having a Shunt-condenser Input.—When the input to the filter is a shunt condenser the action is somewhat different from that which takes place when a series inductance is used, as is apparent from the oscillograms of Fig. 218. Each time the crest alternating-current voltage of the transformer is applied to one of the rectifier anodes the input condenser charges up to this peak voltage, and then discharges into the first inductance until another rectifier anode reaches a peak potential, when the condenser is charged again. The rectifier current flows only a small part of the time, since during most of the cycle the condenser is more positive than all of the anodes. During this discharge period the condenser voltage drops off nearly linearly because of the fact that the first filter inductance draws a substantially constant current from the input condenser. The result is that the input condenser applies a saw-toothed voltage wave to the first inductance. This first inductance and the remainder of the filter then act to prevent the alternating-current voltage across the input condenser from reaching the load. The only essential difference between the filter action with series-inductance and shunt-condenser inputs is therefore that the voltage applied to the first inductance has a wave shape that in the former case is determined only by the rectifier connection, but in the latter case also depends upon the input-condenser capacity and the direct-current load current.

Approximate Method of Calculating Alternating-current and Direct-current Voltages Appearing at Filter Output.—An engineering analysis of the action taking place when the filter input is shunted by a condenser requires that a number of simplifying assumptions be made. To a first approximation the voltage that appears across the input condenser, and which is therefore applied to the remainder of the filter, can be considered as having the saw-toothed shape shown in Fig. 218c. The accuracy of such a representation increases as the amplitude of the voltage variation across the condenser decreases, but in even the most unfavorable circumstances the calculated results are of the proper order of magni-

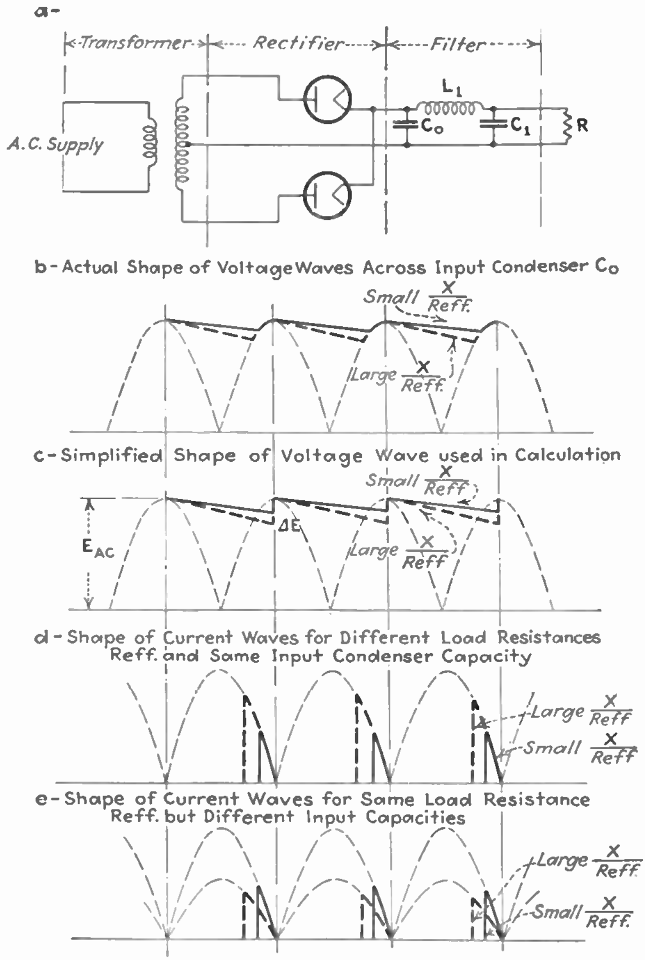


FIG. 218.—Oscillograms showing voltage across input condenser, and idealized rectifier current in typical rectifier-filter systems having shunt-condenser inputs.

tude. The maximum amplitude of the saw-toothed wave is equal to the crest value of the alternating-current voltage applied to the rectifier provided the voltage drop in the rectifier and the leakage reactance of the supply transformer are negligible, and the wave itself can be shown to have the equation[1]

[1] This can be shown as follows: The variation ΔE in voltage across the input condenser is caused by the load current I discharging for a time interval Δt that is equal

$$\text{Voltage across shunt condenser} = \frac{E_{ac}}{1 + \pi\dfrac{X}{R_{\text{eff.}}}}\left\{1 + \frac{2X}{R_{\text{eff.}}}\left(\sin \omega t + \frac{1}{2}\sin 2\omega t + \frac{1}{3}\sin 3\omega t + \cdots\right)\right\} \tag{150}$$

where

E_{ac} = crest value of alternating-current voltage applied to rectifier
ω = $2\pi f$ where f is the number of times the input condenser is charged each second, and so is the fundamental frequency of the ripple voltage
X = $1/\omega C$ = reactance of shunt input condenser at the frequency f
$R_{\text{eff.}}$ = direct-current load resistance plus direct-current resistance of filter inductances.

The important features of Eq. (150) are plotted in Fig. 219 as a function of $X/R_{\text{eff.}}$.

The voltage appearing across the output of the filter can now be calculated using the same general methods employed with filters having an inductance input, but with the difference that the voltage applied to the first inductance is given by Eq. (150) instead of being obtained from Table XII. The alternating-current voltage appearing in the filter output is caused almost solely by the lowest frequency alternating component of Eq. (150) because of the high filter attenuation offered to the higher frequencies. The ratio of the crest value of this ripple-frequency voltage to the direct-current voltage appearing across the input condenser is $2X/R_{\text{eff.}}$, and so depends upon the size of the input condenser and the direct-current load current instead of being determined solely by the rectifier connections as in the case of a series-inductance input. The actual magnitude of the alternating-current voltage in the output can be calculated with the aid of Eqs. (150) and (147). Examination of Eq.

---

to $1/f$, where f is the number of times per second that the first condenser is charged. One can therefore write

$$\Delta E = \frac{\Delta Q}{C} = \frac{I\,\Delta t}{C} = \frac{2\pi I}{2\pi f C} = 2\pi I X$$

The average, *i.e.*, direct-current, component of the saw-toothed voltage is $E_{ac} - \Delta E/2$ and so is $(E_{ac} - \pi I X)$. The load current I is equal to the direct-current voltage divided by the effective load resistance $R_{\text{eff.}}$, so that the average value of the saw-tooth wave can be expressed as

$$\text{Average value of saw-tooth wave} = \frac{E_{ac}}{1 + \pi\dfrac{X}{R_{\text{eff.}}}}$$ = D C. V

Reference to any standard work on Fourier analysis shows that the amplitude of the fundamental frequency component of the alternating-current part of a wave, such as shown in Fig. 218*c*, is $\Delta E/\pi$, and that the amplitudes of the harmonics of this fundamental frequency are inversely proportional to the order of the harmonic, so that Eq. (150) follows at once.

(150) or Fig. 220 shows that the direct-current voltage developed across the input condenser also depends on the ratio $X/R_{\text{eff.}}$ and becomes less as the direct-current load resistance and the capacity of the input condenser are decreased.

Miscellaneous Features of Rectifier-filter Systems Employing Condenser Input.—When the voltage drop in the rectifier and the leakage reactance of the supply transformer are neglected, the current which the rectifier supplies to the input condenser has a wave shape that represents a section of the sine wave of current that would flow into the condenser if the alternating-current voltage that is applied to the rectifier were applied

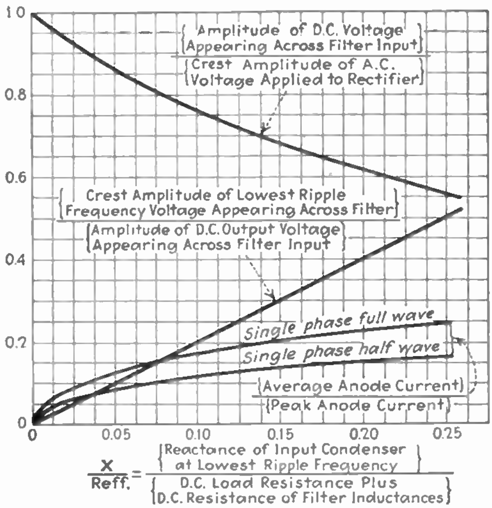


FIG. 219.—Properties of filter systems having a shunt-condenser input plotted as a function of the ratio of the reactance X of the input condenser at the lowest ripple frequency to the effective direct-current load resistance R_{eff}. These curves are approximate in that they neglect the effects of voltage drop in the rectifier and the leakage reactance of the transformer and assume that the voltage across the input condenser has the saw-tooth wave shape of Fig. 218*c*.

directly to the input condenser, as is shown at Fig. 218*d* and *e*. Strictly speaking, the rectifier current is equal to the sum of the direct-current load current and the current flowing into the condenser, but when the condenser receives charge for only a small fraction of the cycle the correction for the load current is small. The ratio of the average to peak anode current depends upon the ratio $X/R_{\text{eff.}}$ and is expressed with good accuracy by a rather complicated equation which gives the results plotted in Fig. 219. The plotted curves do not take into account the effect of leakage reactance in the supply transformer or the voltage drop in the rectifier, but as both of these factors decrease the peak currents the actual ratio will always be somewhat less than that obtained from Fig. 219.

The maximum inverse voltage that a rectifier must stand when delivering its output to a filter having a shunt-condenser input is twice the crest value of the alternating-current voltage applied to the rectifier when any of the single-phase circuits are used. With polyphase arrangements the inverse voltage is less, but as polyphase rectifiers are seldom used with filters having a shunt-condenser input the exact relation is unimportant.

Examination of Eq. (150) and Fig. 219 shows that the fundamental characteristics of a rectifier system having a condenser input are determined by the value of the ratio $X/R_{\text{eff.}}$. A small value of X/R_{eff}, *i.e.*, large input condenser in proportion on the direct-current load current, causes the alternating-current voltage applied to the first filter inductance to be small, and also gives the direct-current voltage a good regulation and a value approximating the crest alternating-current voltage applied to the rectifier. At the same time Figs. 218 and 219 show that a small value of $X/R_{\text{eff.}}$ causes the maximum instantaneous rectifier current to be large compared with the average rectifier current and so makes poor use of the peak current rating of the rectifier. The size of condenser that should be employed is therefore a compromise between conflicting factors, and the exact value for any particular case will depend upon the special circumstances involved. After the shunt-input-condenser capacity has been decided upon, the remainder of the filter is designed in the same way as when a series-input inductance is employed except that instead of satisfying Eq. (148) the first inductance should have a reactance at the ripple frequency that is at least several times the reactance X of the input condenser.

The utilization factor of transformers used in rectifier-filter systems having a shunt-condenser input is always much lower than when the filter has a series-inductance input, but the exact value follows a law too complicated to be of much practical value. The general rule is to use a transformer somewhat (perhaps 50 per cent) larger than would be required with the same direct-current output power obtained from a filter having a series-inductance input.

97. Filter Circuits—Miscellaneous Comments.—The rectifier and filter connections that should be used in a particular case depend upon the individual circumstances. Polyphase rectifiers are preferred when the direct-current power is in the order of several kilowatts or more because the ripple in the output of a polyphase rectifier is of small magnitude and of high frequency and so can be easily filtered out. Smaller outputs are most conveniently obtained from single-phase full-wave rectifiers, usually of the center-tapped transformer type shown in Fig. 211*b*. Filters with series-inductance inputs are always used in polyphase circuits and are also preferred with high-power single-phase rectifiers because of the high ratio of average to peak current that is obtained with a series-inductance input. The use of a shunt-condenser input causes the alternating-current voltage appearing across the filter input to be less than with the series-inductance input, but the ratio of average to peak rectifier current is smaller, so that the condenser input is employed only in low-power single-phase rectifiers, such as those used to supply plate power for broadcast receivers.

When a filter with a series-inductance input is operated under conditions which fail to satisfy Eq. (148) the resulting situation is intermediate between that existing when Eq. (148) is satisfied and when a shunt input condenser is used. Thus the ratio of average to peak rectifier current is higher than in the latter case but lower than in the former, while the direct-current and alternating-current voltages applied to the first inductance depend somewhat on the ratio $X/R_{\text{eff.}}$, but not as much as in Eq. (150).

The condensers used in filters must be capable of *continuously* withstanding a direct-current voltage that is equal to the crest alternating-current voltage applied to the rectifier. Ordinarily a single condenser should be used to withstand the entire voltage, rather than several condensers in series. When condensers are in series the direct-current voltage stress divides between them in proportion to their leakage resistances rather than their dielectric strength, and the leakage resistances are variable and uncertain.

The inductance coils used in the filter must have laminated iron cores, with an air gap that is sufficient to prevent the direct current from saturating the iron. The inductance that is effective to the alternating currents will vary with the superimposed direct current as discussed in Sec. 6 and will normally be greatly lowered by the presence of the direct current. The insulation between the winding and the core must be capable of withstanding the full direct-current voltage delivered by the rectifier output.

98. Filters Employing Resonant Elements.—The filters illustrated in Fig. 213 all make use of series inductances to choke out the ripple voltages, and shunt capacities to short-circuit them. The effectiveness of the filtering for any particular frequency can be increased by the use of a tuned filter element as shown in Fig. 220. In the filter shown at *a* the series inductance L_1 is shunted with a capacity C_1' of such a size as to produce parallel resonance to the frequency which it is desired to suppress. At *b* the shunt element $L_1'C_1'$ is made resonant at the desired frequency and has the effect of short-circuiting the filter as far as this particular frequency is concerned. The tapped-inductance arrangements of Fig. 220*c* and *d* are equivalent in their action to 220*b* but have the advantage of not requiring an extra inductance.

The tapped-choke arrangements are commonly employed in the filters used in broadcast receivers and can be understood by the aid of

Fig. 220*e* and *f*. Tapping the inductance causes it to act as a transformer in which the primary is the portion of the inductance L_p through which the direct current flows while the secondary L_s is the other end of the tapped choke. The primary and secondary have leakage reactances L_p' and L_s', respectively. The leakage reactance L_s' of the secondary is tuned by condenser C_o' to resonance at the frequency to be suppressed. The resulting effect is equivalent to shunting the filter with a series-resonant circuit as shown at Fig. 220*f*, which except for the small leakage inductance L_p' is equivalent to Fig. 220*b*.

Filters with resonant elements are very effective in suppressing a particular ripple frequency. The chief disadvantage of filters with

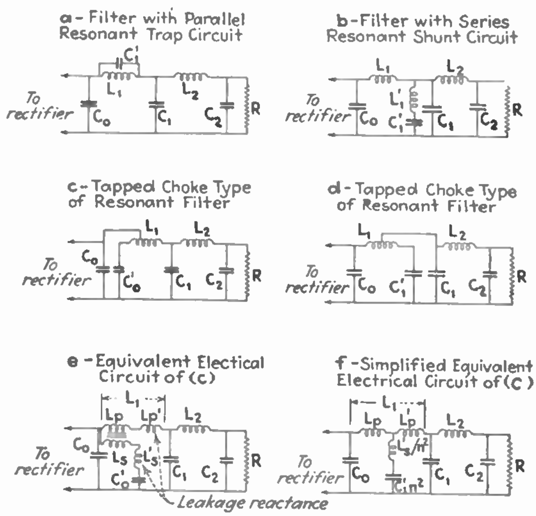


FIG. 220.—Filter circuits using tuned elements.

tuned elements is that the chokes must be constructed with considerable care if individual adjustment of the inductance is to be avoided. Furthermore, since the inductance of a coil varies with the direct-current saturation, the filter inductance, and hence the necessary tuning capacity, will vary with the load on the rectifier-filter system except for the case of Fig. 220*b*. The relative merits of resonant and non-resonant elements must therefore be decided for each particular filter, and in general will be a balance between conflicting factors.

99. Self-rectifying Circuits.—If an alternating voltage is applied directly to the plate of an oscillator in place of the usual direct-current potential the oscillator will operate on the positive half-cycles of the applied voltage and so will generate wave trains having the character shown in Fig. 221. These are known as interrupted continuous waves (abbreviated I.C.W.) and are occasionally used in radio telegraphy. Another type of self-rectifying oscillator circuit is shown in Fig. 222, in which the two plates of a two-tube oscillator are supplied with alternating-current voltages that are 180° out of phase, and a large inductance is inserted in the common filament return lead. This arrangement is electrically equivalent to a single-phase full-wave rectifier circuit with a filter consisting of a single inductance (see Fig. 213*a*). When the inductance is large the total current drawn by the two tubes is substantially

FIG. 221.—Interrupted continuous waves generated by applying an alternating-current voltage directly to the plate of an oscillator.

constant, and the generated oscillations will also be of constant amplitude. Practically, however, it is impossible to maintain the total current entirely constant, and as a result modulated oscillations are generated.

Self-rectifying oscillator circuits find their chief usefulness under conditions where the complications of a rectifier-filter system are to be avoided. These circuits are not suitable for ordinary commercial radio transmitters because the voltage that is applied to the plates of the

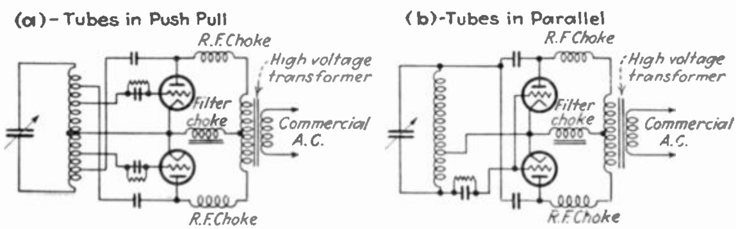


FIG. 222.—Self-rectifying oscillator circuits capable of generating oscillations of substantially constant amplitude if the filter choke is very large.

oscillators is not constant, and as a result the generated oscillations suffer both frequency and amplitude modulation.

100. Power Supply Systems for Aircraft Equipment.—The power for operating aircraft transmitting equipment can be obtained from a wind-driven generator, an engine-driven generator, or a dynamotor operated from a storage battery. The wind-driven generator is the simplest power source but is inoperative when the plane is on the ground. The engine-driven generator similarly will not function when the plane is grounded with a disabled engine. The dynamotor is not subject to these limitations but requires a storage battery capable of supplying considerable power.

Wind-driven generators are usually designed to supply both filament and plate voltages and are made in such a way as to give a substantially constant output voltage over an extremely wide range of wind velocities.

The engine-driven generator is similarly normally arranged to provide both filament and plate voltages and ordinarily involves some voltage-regulating mechanism to insure a constant output irrespective of engine speed. Dynamotors need supply only plate potential since the storage battery can be used to operate filaments.

Anode and filament power for aircraft receivers can be obtained from batteries, or from generators such as those employed in transmitters. In commercial installations the filament power for receivers is usually obtained from the plane storage battery while anode power is derived from some form of generator.

CHAPTER XII

RADIO TRANSMITTERS

101. Radio Transmitters—General Considerations.—Commercial transmitting equipment is ordinarily mounted on a framework of structural-steel members fronted by a vertical metal panel containing the controls and meters necessary for adjusting and monitoring the transmitter. All equipment appearing on the panel is at ground potential, and instruments which must be observed during adjustment or operation, and are not at ground potential, are located behind the panel and viewed through windows. The steel frame is normally enclosed with wire mesh of some sort and is provided with doors that cut off the transmitter power when opened. Typical examples of this general type of construction are to be found in Figs. 224 to 226, 231, and 234. This type of construction requires a minimum of floor space in proportion to the amount of apparatus involved, makes the transmitter accessible for inspection and repairing, and eliminates all hazard to persons.

It is highly important that the carrier frequency of a transmitted wave be maintained as nearly as possible at the assigned value, and this requirement becomes so important in short-wave and broadcast transmitters as to dominate the entire design. Consider the case of a short-wave code transmitter which has a carrier frequency of 20,000 kc and is transmitting at a speed of 1000 letters per minute. According to the data given in Sec. 83 such a transmitter requires a side band approximately 131 cycles wide. Thus with perfect stability of the carrier frequency two such transmitters could operate on frequencies differing by only 300 cycles, and a relatively limited space in the frequency spectrum could accommodate a large number of transmitters. If, however, the carrier frequency of each transmitter cannot be maintained at an assigned value to closer than 0.03 per cent, each transmitter will at times deviate from its assigned frequency by 6000 cycles, and instead of the frequency separation of 300 cycles, which would be possible with perfect stability, the frequency separation would actually have to be 12,300 cycles, or 41 times as great. This example makes clear that the accuracy with which the carrier frequency can be maintained at an assigned value is of fundamental importance in determining the number of communication channels which can be operated in a given part of the frequency spectrum. The importance of frequency stability is more or less proportional to the carrier frequency, since a given percentage variation in the frequency represents a smaller number of cycles at a low than at a high carrier frequency.

With broadcast equipment the carrier frequency must be maintained constant to a high degree of accuracy because two or more transmitters are often assigned the same carrier frequency and will interfere with each other by the production of audible beat notes unless the stability is so high that the frequency of each transmitter stays within a few cycles of the assigned value.

In addition to the requirement that a transmitter maintain its frequency accurately at the desired value it is also necessary that the transmitter be prevented from radiating energy on other frequencies. All vacuum tubes when operated under conditions which give high plate efficiency generate strong harmonics, and in order to avoid interference with other radio channels some means must be provided for preventing the radiation of these harmonics. This is accomplished by placing selective circuits between the transmitter and the antenna, and by completely enclosing the transmitter with grounded metal sheet or screen in order to prevent the radiation of harmonics directly from the transmitter. The suppression of harmonic radiation becomes of increasing importance as the power of the transmitter is raised because when the radio-frequency power is in the order of kilowatts the harmonic energy is capable of creating radio waves that can be heard many thousands of miles.[1]

102. Short-wave Code Transmitters.—In short-wave code transmitters the benefits gained by maintaining the transmitted frequency accurately at an assigned value are so great as to make it desirable to obtain the radiated frequency from a crystal oscillator whenever possible. When a crystal oscillator is used to control the frequency the entire design of the transmitter is dominated by the fact that the amount of power controlled directly by a crystal is relatively small, and also by the fact that crystals generating frequencies higher than about 4000 kc are so thin and fragile as to be unsatisfactory in commercial service. The usual crystal-controlled transmitter consists of a crystal oscillator using a tube having a rating in the order of 7.5 to 50 watts, followed by a buffer tube which acts as either an amplifier, or harmonic generator of the plate-distortion type. The crystal oscillator and the buffer tube normally have a separate power supply and operate continuously even when the key of the transmitter is up. The buffer tube is followed by harmonic generators, which are usually of the plate-distortion type arranged to develop the second harmonic. After the output of the crystal oscillator has passed through a sufficient number of harmonic generators to develop the frequency that is to be transmitted, the power

[1] For a detailed discussion of the problem of avoiding harmonic radiation, see J. W. Labus and Hans Roder, The Suppression of Radio-frequency Harmonics in Transmitters, *Proc. I.R.E.*, vol. 19, p. 949, June, 1931.

level is raised to the desired point by means of Class C (high-efficiency) amplifiers.

Typical High-power Short-wave Commercial Code Transmitter.—The characteristic features of a well-designed, crystal-controlled, short-wave code transmitter are illustrated by the high-power transmitter shown in Fig. 223. This particular transmitter is designed to operate in the frequency range 6670 to 21,500 kc (45 to 14 meters wave length) and is capable of delivering a power output of 50 kw at the lowest and 23 kw at the highest frequency within this range. The transmitter itself is divided into two parts, the first of which is an exciter unit containing the crystal oscillator, buffer amplifier, harmonic generators, and power amplifiers capable of delivering 1-kw output. The second unit is a one-stage power amplifier using water-cooled tubes, and is connected to the exciter unit by a short transmission line. This arrangement, with the final power amplifier separated from the remainder of the transmitter, is frequently used in high-power equipment because it simplifies the problem of shielding the low power level circuits from the strong fields produced by the final power amplifier. The exciter unit can be considered as a complete transmitter of moderate power, and its output can be delivered directly to an antenna if desired.

The sequence of tubes in the exciter unit is as follows:

7½-watt triode crystal oscillator.
75-watt screen-grid buffer amplifier.
75-watt screen-grid frequency doubler.
75-watt screen-grid frequency doubler (second doubler).
500-watt screen-grid tube used as frequency doubler when output frequency is in excess of 12,000 kc, and otherwise employed as power amplifier.
2, 500-watt screen-grid tubes in push-pull amplifier circuit.

The final power amplifier unit consists of four water-cooled tubes in a neutralized push-pull circuit.

The crystal oscillator uses the circuit shown in Fig. 127*a*, and the associated buffer amplifier is a standard type of screen-grid radio-frequency amplifier. The use of a buffer amplifier, suitable shielding, and a separate power source for the crystal oscillator and buffer tubes, makes the generated frequency virtually independent of the conditions existing in the remainder of the transmitter. The frequency stability is still further increased by placing the crystal in an oven which is normally maintained at a temperature of 45°C. to an accuracy of 0.25°C. The plate voltage of the crystal-oscillator tube is kept well below the rated value in order that the amplitude of the crystal vibrations will be small, while the load placed on the crystal oscillator is kept as low as possible by operating the buffer tube so that its grid never becomes positive. With these precautions the frequency that is obtained will not vary from an assigned value by more than 1 part in 4000.

The frequency doublers are of the plate-distortion type. They are operated with a grid bias greater than the cut-off value and with sufficient excitation to cause the grids to go positive for a part of each cycle. For

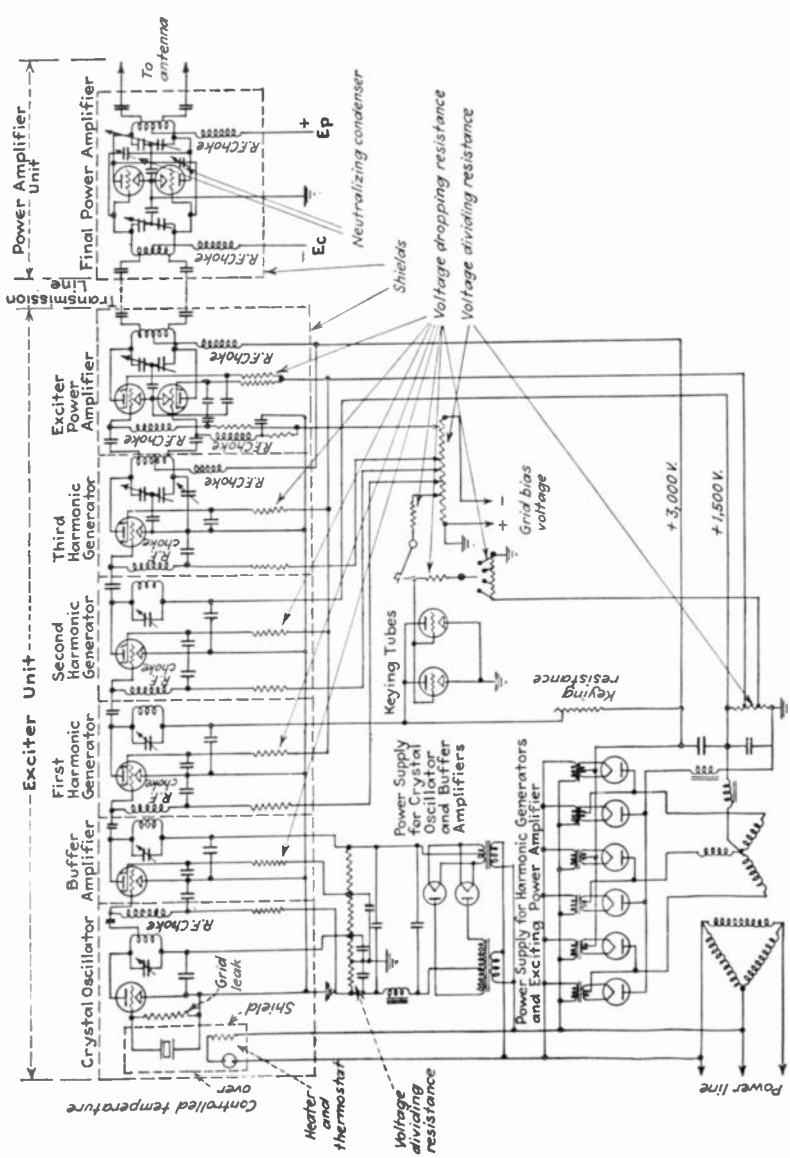


FIG. 223.—Simplified circuit diagram of a representative high-power, short-wave code transmitter (RCA 20–40 kw transmitter).

frequencies below 12,000 kc two doublers are used, and the crystal is ground to generate a frequency one-fourth of that to be transmitted, while for frequencies above 12,000 kc three doublers are employed and the crystal is constructed to give a frequency exactly one-eighth of that to be transmitted.

The power amplifiers are all of the Class C type (*i.e.*, bias greater than cut-off) and are operated with sufficient grid excitation to drive the grid positive for a portion of each cycle. Ordinary tuned radio-frequency amplifier circuits are employed, with no special features except that the two sides of the double-ended tuned circuits associated with the inputs and outputs of the push-pull amplifiers are balanced to ground in order to prevent current from circulating through the ground and introducing undesired stray couplings.

The tuned circuits employed in radio transmitters are generally referred to as "tank" circuits. The tank circuits of the low-power stages of this transmitter are tuned by means of variable air condensers and by short-circuiting turns of the inductance coils. In the high-power stages the rough tuning is accomplished by short-circuiting turns of the inductance and by an adjustable condenser, while the fine tuning makes use of a copper disk that is rotated in the field of the inductance.

Three separate anode power-supply systems are provided. The crystal oscillator and buffer amplifier operate from a single-phase center-tapped rectifier using Type 866 mercury-vapor tubes. A voltage dividing resistance across the output of the rectifier makes it possible to obtain anode, screen-grid, and control-grid bias voltages from a single rectifier. The remainder of the exciter receives its anode power from a three-phase full-wave rectifier using Type 872 mercury-vapor tubes in conjunction with a filter consisting of a series input inductance followed by a shunt condenser. This unit delivers 3000 volts direct-current, which is applied directly to the plates of the 500-watt screen-grid power-amplifier tubes. Lower voltages for the other tubes and for the screen grids are obtained either by means of voltage dropping resistances from the 3000-volt source, or from a 1500-volt tap obtained by connecting to the neutral point of the secondary star through a filter inductance. This tap causes three of the six rectifier tubes to do double duty by also functioning in a three-phase half-wave circuit to give 1500 volts output. The grid-bias voltages required by the exciter are obtained either from a 1000-volt motor generator or from a suitable rectifier-filter arrangement with the aid of a voltage-dividing resistance. The power-amplifier unit has a separate three-phase full-wave rectifier using Type 869 mercury-vapor tubes and delivering a direct-current output of 8 amp. at 12,000 volts. Twelve rectifier tubes are employed, operated in parallel pairs by means of a balance coil. The filter consists of a 0.5-henry series input inductance followed by a 3-μf shunt condenser.

It is to be understood that this description of the transmitter has dealt only with the outstanding features and has omitted innumerable details which must be taken into account in the actual construction in order to obtain satisfactory results.[1] For example it is necessary to shield the parts of the transmitter from each other by means of copper shields, the location of which is indicated by the dotted lines of Fig. 223. Radio-frequency choke coils and by-pass condensers must also be provided in many places not shown in this schematic diagram if parasitic oscillations and regeneration are to be avoided, and it is also necessary

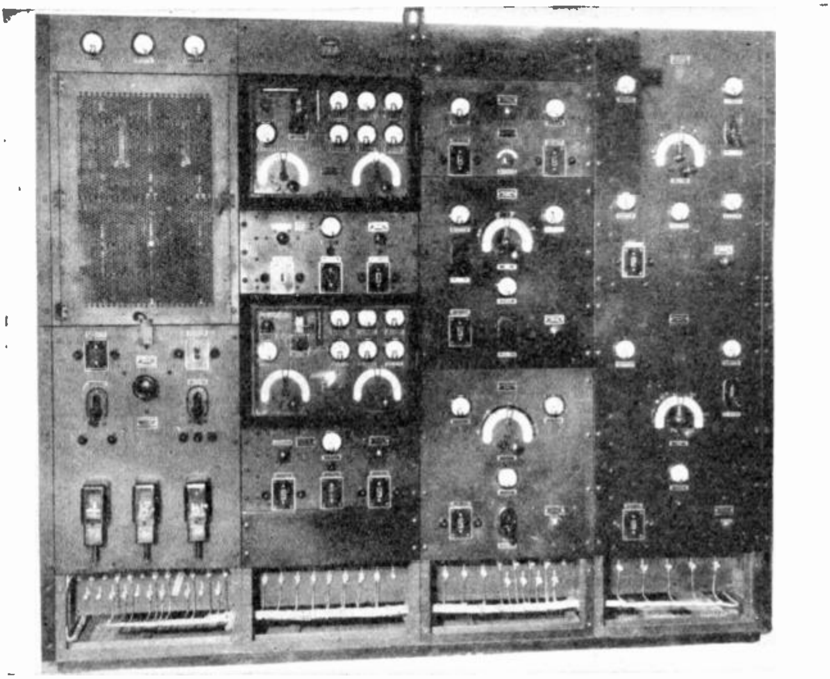


FIG. 224a.—Exciter unit of transmitter of Fig. 223—front view of panel.

to arrange the various parts of the transmitter in relation to each other in such a way as to minimize undesirable stray couplings while making the connections as short and as direct as possible.

Photographs showing different views of the complete transmitter are given in Figs. 224, 225, and 226 and serve to bring out the more obvious features of construction.

While the high-power transmitter that has been described incorporates all of the typical features of a short-wave crystal-controlled code transmitter, it is of course possible to modify the layout in many details. Thus the buffer tube is often made to serve as a harmonic generator by tuning its tank circuit to twice the frequency generated by

[1] For further details concerning this transmitter, see I. F. Byrnes and J. B. Coleman, 20–40 Kilowatt High-frequency Transmitter, *Proc. I.R.E.*, vol. 18, p. 422, March, 1930.

the crystal. Triodes are also commonly used as harmonic generators instead of screen-grid tubes, and can be used as radio-frequency amplifiers in place of screen-grid tubes provided neutralized circuits are employed. In some cases the harmonic-generator tubes are called upon to generate the third, or even the fourth harmonic instead of the second, although the power output on these higher harmonics is small. Sometimes tuned circuits use only the capacities supplied by tubes, leads, etc., and are adjusted to resonance by varying the circuit inductance.

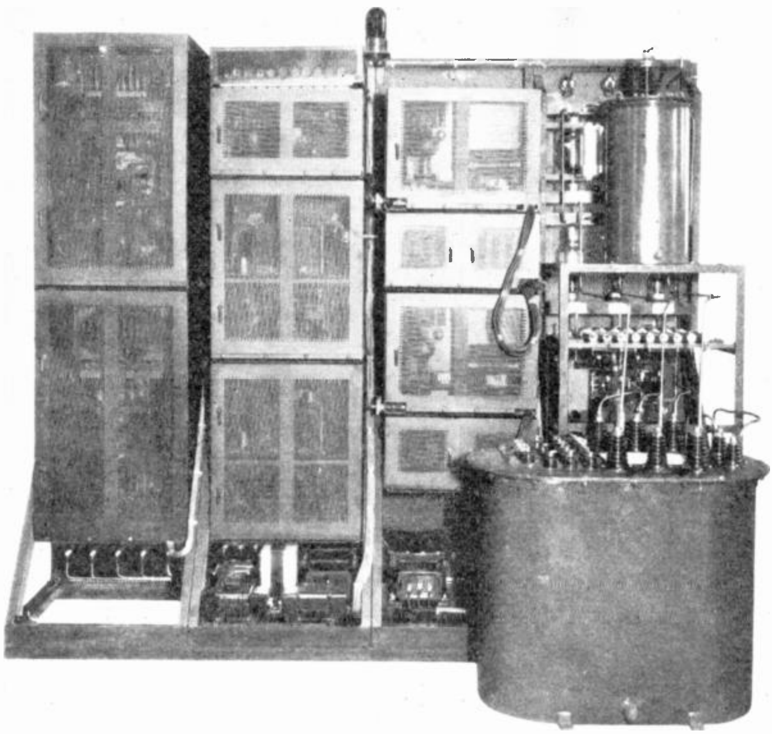


FIG. 224b.—Exciter unit of transmitter of Fig. 223—rear view with shields in place.

Transmitters Not Employing Crystal Control.—Although crystal-controlled transmitters represent the best practice in short-wave transmission there are circumstances which sometimes make it either impossible or uneconomical to go to the complicated layout required when a crystal oscillator is employed. In such circumstances a master-oscillator power-amplifier type of transmitter is commonly used. This arrangement consists of an ordinary triode oscillator exciting a screen grid or a neutralized triode radio-frequency power amplifier, as shown in Fig. 227. The master-oscillator power-amplifier type of transmitter is relatively simple compared with the crystal-controlled transmitter because the oscillator can be of sufficient capacity and of the correct frequency to excite the power amplifier directly. Master-oscillator power-amplifier arrangements have fair frequency stability because of the fact that the load on the oscillator is small, and because the power amplifier serves to separate the oscillator from external influences which would otherwise affect the frequency.

Short-wave radio transmitters sometimes employ an oscillator delivering power directly to the radiating system. This eliminates harmonic

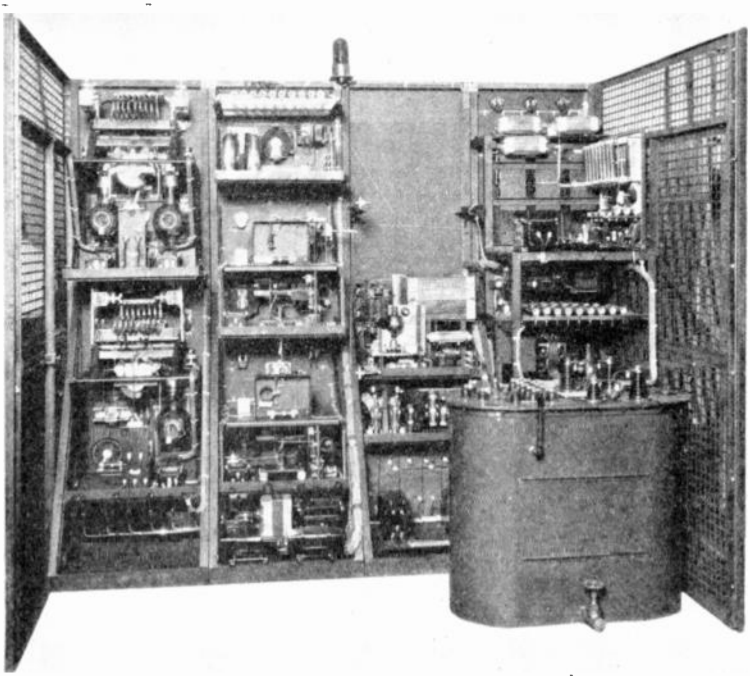


FIG. 224c.—Exciter unit of transmitter of Fig. 223—rear view with covers of shields removed.

generators and power amplifiers but has the disadvantage of very poor frequency stability, caused in part by the load from the output, as explained in Sec. 53, and in part by the fact that the coupled output circuit helps determine the frequency. Hence anything that affects the antenna, such as wind, temperature, etc., will change the frequency. In order that the frequency stability of this type of transmitter may be as high as possible it is desirable to use a tuned oscillator circuit in which the capacity is large in proportion to the inductance, together with an adjustment that causes the reactive volt-amperes circulating in this "tank" circuit to be large. The radiating system is then loosely coupled to the oscillator in order to minimize the effects which variations in the constants of the radiating system will have on the frequency. Such an arrangement does not deliver the largest possible power to the antenna

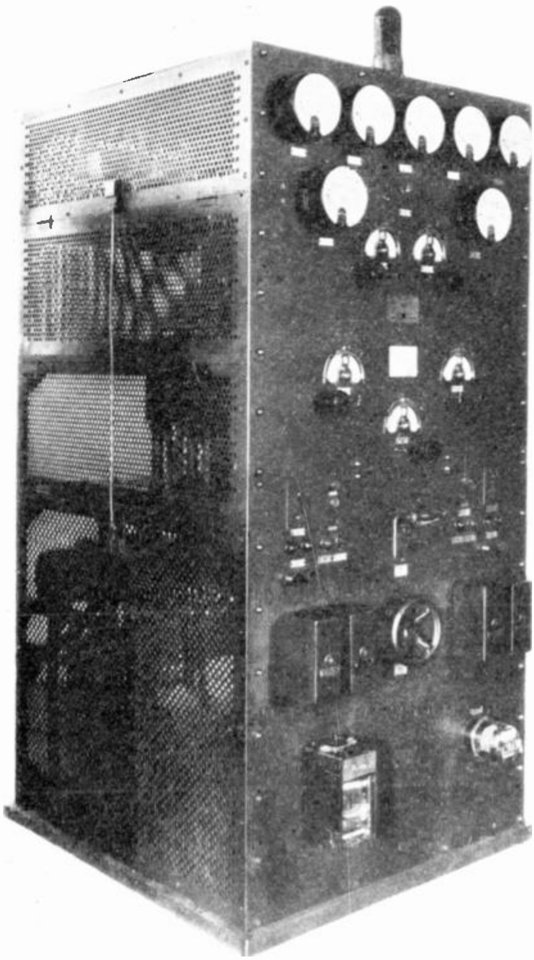


FIG. 225a.—Front view of 20–40 kw output power amplifier.

but is more satisfactory than a much larger power at a lower frequency stability. Transmitters making use of an oscillator feeding directly into the radiating system are used primarily by amateurs, where it is merely necessary that the radiated frequency be kept within a rather wide band. A schematic diagram of such an amateur transmitter is shown in Fig. 228.[1]

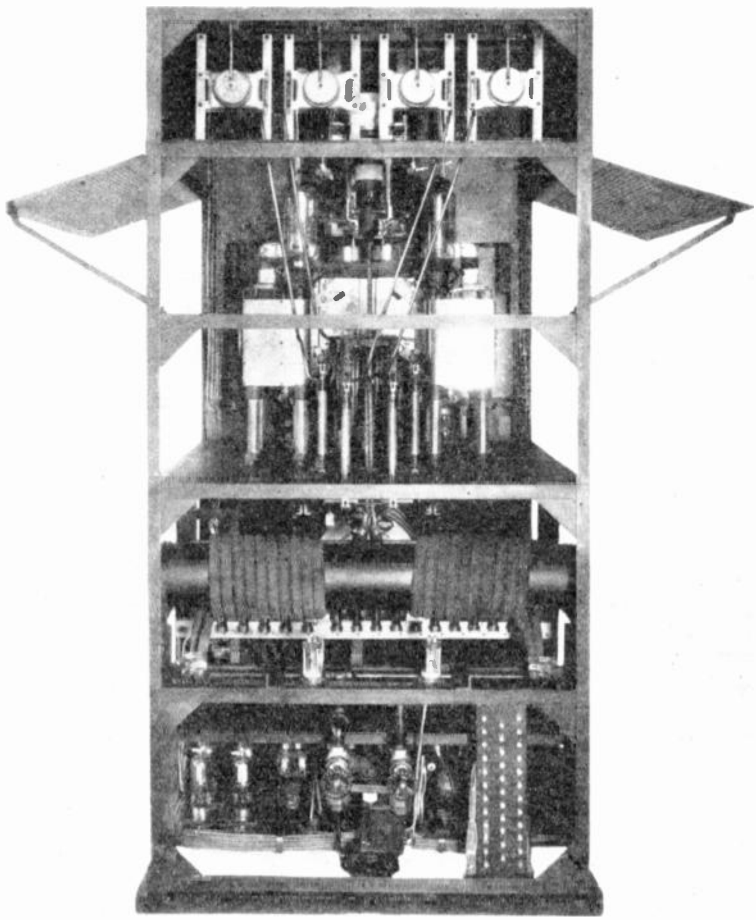


FIG. 225b.—Rear view of 20–40 kw output power amplifier.

103. Moderate- and Long-wave Code Transmitters.—Code transmitters intended for operation at moderate and long waves, *i.e.*, at frequencies below 550 kc, differ from the corresponding short-wave code transmitters primarily in that the required degree of frequency stability is considerably lower. This is because a given percentage variation of a

[1] Numerous examples of self-oscillator, and master-oscillator power-amplifier circuits are to be found in the pages of *QST*, the organ of the American Radio Relay League.

low carrier frequency represents a small number of cycles compared with the same percentage variation of a high-frequency carrier. The master-oscillator power-amplifier type of transmitter is hence standard for code work on frequencies below 550 kc, and some use is made of transmitters in which an oscillator delivers energy directly to the radiating system.

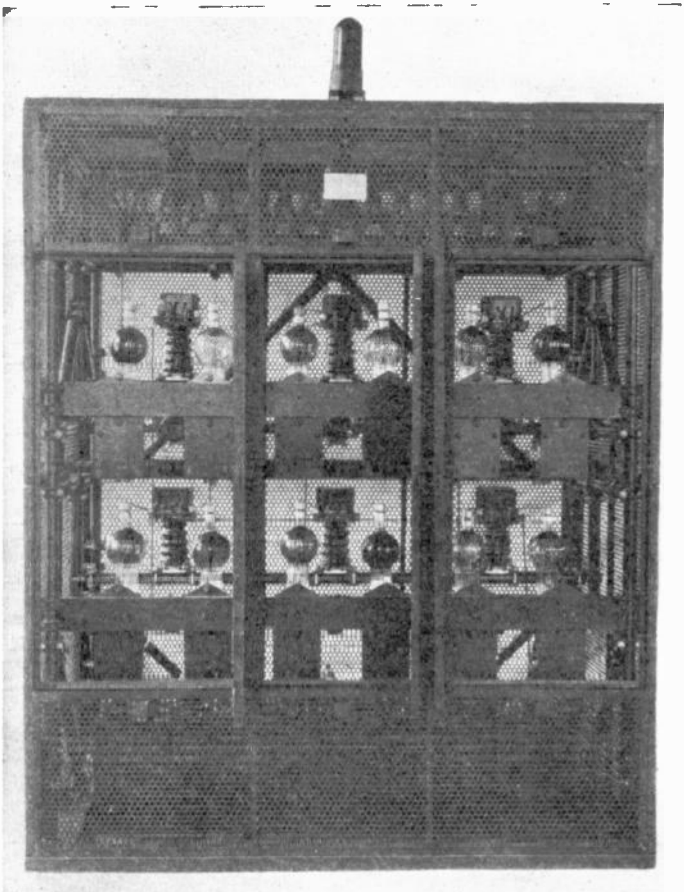


FIG. 226a.—Front view of rectifier and filter for 20–40 kw power amplifier.

Master-oscillator power-amplifier transmitters of low power generally consist of a master oscillator and a single screen-grid or triode power amplifier operating directly into the radiating system, while when the power is large an intermediate power amplifier is often placed between the master oscillator and the final amplifier. A schematic diagram of a commercial 2-kw ship transmitter intended for service in the frequency range 125 to 500 kc is shown in Fig. 229. This consists of a 50-watt master oscillator, an intermediate power amplifier composed of four 50-watt tubes in parallel, and a final amplifier which is impedance coupled to the intermediate amplifier and consists of two 1-kw tubes in parallel. The different parts of this transmitter are provided with complete shielding in order to improve the frequency stability, reduce the tendency toward self-oscillation in the amplifier, and permit operation near metal

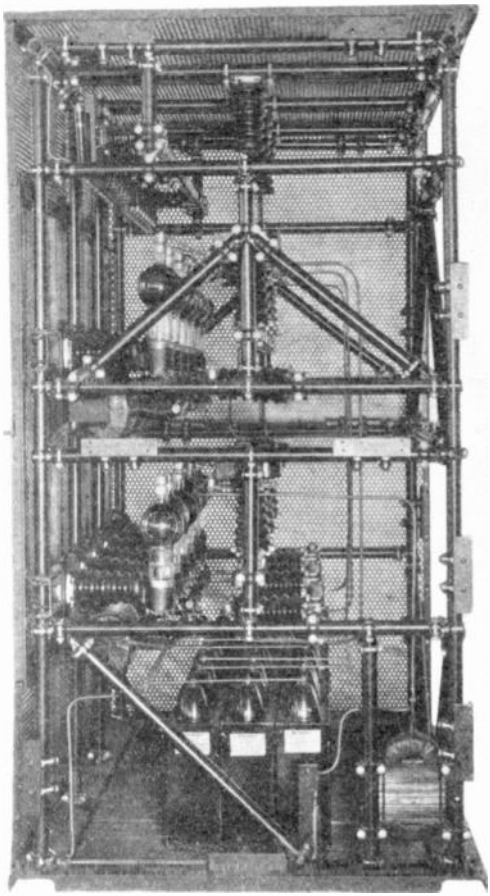


FIG. 226b.—Side view of rectifier and filter for 20–40 kw power amplifier.

bulkheads. A later model of this transmitter has been built using screen-grid tubes in the intermediate and output amplifiers, and with a tuned circuit to couple these stages.

Many code transmitters intended for use in the frequency range 125 to 500 kc are provided with some means for interrupting the radiation at a rate of 500 to 1000 times per second. The resulting signals are known as interrupted continuous waves (often abbreviated I.C.W.), and in a receiver of the broadcast type produce code signals having a pitch corresponding to the rate of interruption at the transmitter. The interruptions are ordinarily produced by applying alternating current directly to the plate, or by superimposing an alternating-current voltage on the

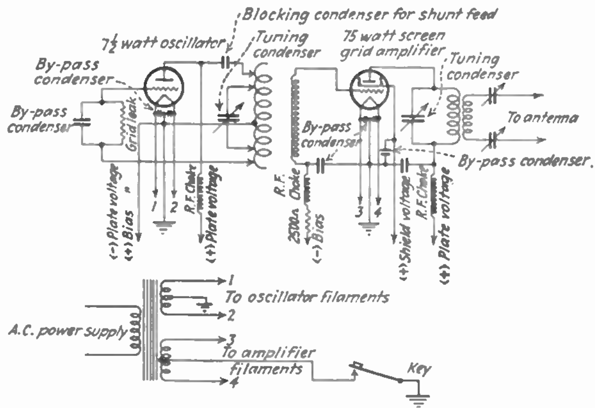


FIG. 227.—Circuit diagram of typical low-power master-oscillator power-amplifier short-wave transmitter.

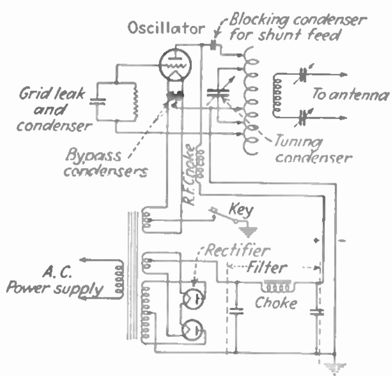


FIG. 228.—Circuit diagram of short-wave transmitter in which the oscillator tube delivers its output directly to the radiating system.

direct-current plate potential, or by the use of some sort of a motor-driven interrupter, called a "chopper," located somewhere in the transmitter.[1]

104. Keying of Code Transmitters.—The output of a crystal-controlled transmitter is ordinarily turned on and off in accordance with the characters of the telegraph code by means of a keying system which

[1] A number of typical moderate-wave code transmitters are described by I. F. Byrnes, Recent Developments in Low-power and Broadcasting Transmitters, *Proc. I.R.E.*, vol. 16. p. 614, May, 1928.

operates on one of the low level amplifier or harmonic generator tubes in such a way as to remove the alternating-current excitation from the grids of the larger tubes. The full power output is thus controlled by keying only a small amount of energy and a low-power relay can be used. The crystal oscillator and its buffer tube normally have an independent source of anode power and are allowed to operate continuously in order that the generated frequency may reach an equilibrium value that is unaffected by the rapidity and character of the keying.

While there are many ways by which a power-amplifier or harmonic-generator tube may be prevented from delivering output to a succeeding

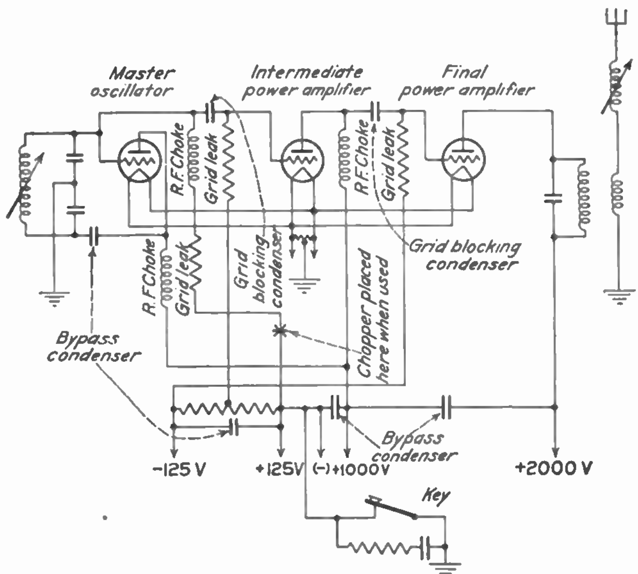


FIG. 229.—Simplified circuit diagram of 2-kw ship transmitter designed for use in the frequency range 125 to 500 kc. This transmitter consists of a master oscillator, an intermediate and an output amplifier, and is keyed by opening the ground connection of the plate and grid return leads.

tube, two of these find the widest application. The first is illustrated in the schematic diagram of the high-power transmitter shown in Fig. 223. Keying is accomplished here in the first doubler stage by reducing the plate voltage supplied to the doubler tube to the point where the output is insufficient to bring the instantaneous grid potential of the succeeding tube above cut-off. The keying unit consists of two 50-watt tubes in parallel, the plates of which are fed in parallel with the plate of the first doubler tube through a common resistance from the 3000-volt power supply. When the key is down, a negative bias exceeding the cut-off value is placed on the grids of the keying tubes so that the keying unit draws no current and allows normal voltage to be applied to the plate of the doubler tube. When the key is up, a slightly positive voltage is applied to the grids of the keying tubes, causing these tubes to draw a large plate current through the series resistance and hence to reduce the potential that is applied to the plate of the doubler tube to a very low value. The second widely used keying method consists of an arrangement which places a large additional bias voltage on the grid of one of the amplifier or harmonic-generator tubes when the key is up and removes the extra bias when the key is down. The added bias is made sufficient to block the tube and so prevents operation of the transmitter.

In the master-oscillator power-amplifier type of transmitter the ideal keying arrangement would permit the oscillator to operate continuously from a separate power source and would key in the power-amplifier circuit. From a practical point of view, however, a separate power source for the master oscillator is seldom justified. The keying method commonly employed with master oscillators is illustrated in the transmitters shown in Figs. 227 and 229. It consists in grounding the filaments of all tubes and bringing the grid- and plate-return leads to ground through the key. When the key is down, the grid-return and negative plate leads are grounded and the tubes operate in a normal manner; while when the key is up, the positive plate lead and the filament assume the same potential and the negative plate lead becomes negative with respect to the filament. Since the negative plate supply lead is connected to the grids, a high negative voltage with respect to the filament is placed on the grids, and this blocks the tubes. Transmitters consisting of an oscillator which delivers its power directly to the radiating system are usually keyed in the same manner, as shown in Fig. 228.

Keying Troubles.—In keying self-oscillator and master-oscillator power-amplifier types of short-wave transmitters trouble is always experienced from frequency instability. This is because the generated frequency depends to a certain extent upon the temperature of the tubes and circuits, and during a keying impulse, such as a dot or dash, heating takes place with the result that the frequency of the first and last parts of a dash will be different, resulting in what is often called "chirping." Furthermore the average temperature depends upon the character of the keying, being greater with a series of dashes than with a series of spaced dots. Again the temperature and hence the frequency will be different after a short rest period than it will be after continuous keying. Commercial oscillator and master-oscillator short-wave code transmitters minimize such troubles by operating the tubes at a conservative rating and by sending out dots during otherwise idle intervals in order to keep the temperature somewhere near normal.

Some keying methods fail to give clean-cut dots and dashes when the speed of transmission is high. This is the case, for example, when the key is placed in the primary of the rectifier transformer. After the key is closed a brief interval must elapse before the normal direct-current voltage is obtained from the filter output, while after the key breaks the circuit the transmitter will continue to operate for a brief time by making use of the energy stored in the filter inductances and condensers. The result is that tails are formed after each impulse, and with rapid keying the successive code characters tend to overlap and so become indistinct. The remedy for such a condition is a less effective filter, or slower speed of transmission, or a different keying method.

When the key is located where it must break the direct current that flows through an inductance there is a tendency for sparking to take place at the key contacts. This trouble can be eliminated by shunting the key with a condenser and resistance in series, as shown in Fig. 229. The condenser should be relatively large, *i.e.*, in the order of several microfarads for a low- or moderate-power transmitter, while the resistance should be of the same order of magnitude as the direct-current resistance of the circuit which is being keyed.

Code transmitters will frequently cause interference in near-by radio receivers which are tuned to receive a frequency entirely different from that being transmitted. This interference is in the form of clicks or thumps which occur as the key makes and breaks contact. Such key clicks result from suddenly starting and stopping the oscillations of the code transmitter and in this way producing side-band frequencies that are widely different from the carrier frequency. In the theoretical case, where the oscillations are started and stopped instantly by the making and breaking of the key contacts, the side-band frequencies contained in the code signal extend in a continuous spectrum from zero to infinite frequency. The amplitude of the energy on any given frequency is of course very small but is sufficient to cause interference with radio receivers which are in the immediate vicinity. Key clicks can be eliminated by any arrangement that will prevent the oscillations from building up and dying down too rapidly. In transmitters such as that shown in Fig. 223, where there are a number of tuned circuits between the key and the final output to the radiating system, the selectivity obtained by sending the oscillations through this chain of circuits is usually sufficient to reduce key clicks to a minimum. In other cases it may be necessary to provide special time-delay circuits to prevent the oscillations from building up and dying out as rapidly as would otherwise be the case. Such circuits usually consist of an inductance shunted by a high resistance, and the parallel combination then placed in series with a plate lead of one of the tubes. The time required for the current to build up and die down in the inductance gives the necessary time lag, while the high shunting resistance is for the purpose of preventing excessive surge voltages from developing across the inductance.

105. Radio-telephone Transmitters.—A transmitter for producing radio-telephone signals is essentially a code transmitter to which has been added some means of modulating the output power. While there are as many varieties of radio-telephone transmitters as there are combinations that can be formed from all the possible modulating systems combined with all the possible types of code transmitters, nearly all radio-telephone transmitters are of the crystal-controlled or master-oscillator type, with the modulation obtained by means of a plate-modulated Class C type of amplifier. This arrangement is used because it insures a high degree of frequency stability, prevents the modulation from varying the carrier frequency and hence from introducing frequency modulation, and because the modulation can be accomplished at a relatively high-power level.

Broadcast Transmitters.—The transmitters used in radio broadcasting represent the highest development in radio-telephone transmitters, and practically all other types of telephone transmitters are modifications of broadcast equipment. Crystal control is used in all first-class broadcast transmitters in order that the maximum possible frequency stability will be obtained. Since the broadcast frequencies lie in the range 550 to 1500 kc it is possible to grind crystals to generate the desired carrier frequency, and harmonic generators are unnecessary. The output of the crystal oscillator is amplified by means of a series of neutralized or screen-grid Class C amplifiers, and is then fed to a modulated amplifier which is normally of the plate-modulated Class C type. The modulated output is then either radiated directly or is amplified further by means of linear (Class B) amplifiers. Linear amplification instead of Class C amplification must be used after modulation in order to avoid distortion. While it is possible to use distortionless (Class A) amplification after modulation this is never done in practice because the plate efficiency of such amplifiers is very much less than the plate efficiency of the linear amplifier.

Schematic circuit diagrams of typical broadcast transmitters are shown in Figs. 230, 232, 233, and 235. The transmitter of Figs. 230 and 231 is a standard Western Electric 1-kw transmitter giving 100 per cent modulation and is used in many broadcasting stations.[1] The tube sequence in the radio-frequency chain is as follows:

50-watt crystal oscillator.
50-watt Class C amplifier.
50-watt modulated Class C amplifier.
250-watt linear (Class B) amplifier.
1-kw water-cooled linear (Class B) amplifier.

[1] For a more detailed description of this and other Western Electric broadcast transmitters see Edward L. Nelson, Radio Broadcasting Transmitters and Related Transmission Phenomena, *Proc. I.R.E.*, vol. 17, p. 1949, November, 1929.

The audio-frequency system consists of a 50-watt tube operated as an audio-frequency amplifier and resistance-coupled to a 250-watt output tube which modulates the Class C amplifier. The crystal oscillator is impedance coupled to the first amplifier and oscillates because the coupling inductance supplies a plate load impedance having an inductive reactance. The modulated amplifier and the Class B amplifiers are neutralized by the Rice method (see Sec. 44). It will be observed that the power rating of the modulator tube is considerably higher than that of the modulated amplifier, and that the modulated amplifier operates at a lower plate voltage. These features are necessary if complete modulation of the radiated wave is to be obtained.

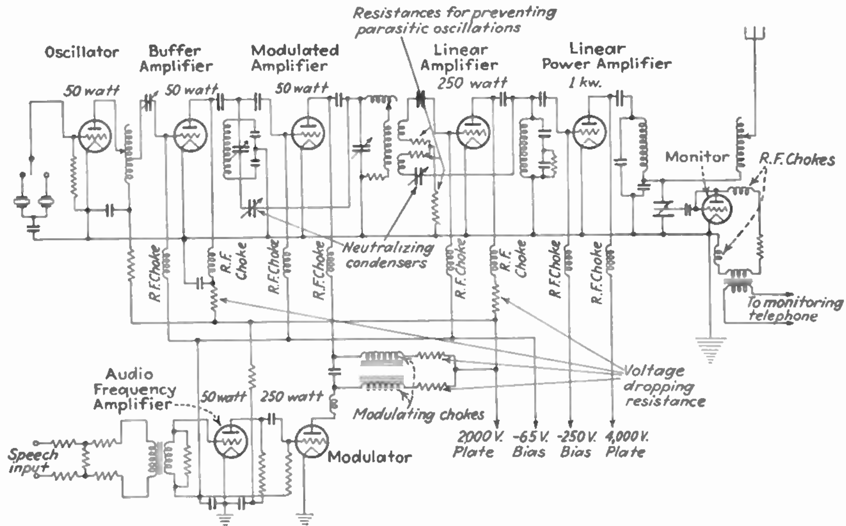


FIG. 230.—Simplified circuit diagram of standard Western Electric 1-kw broadcast transmitter capable of 100 per cent modulation.

The schematic circuit of a Western Electric 50-kw broadcast transmitter capable of delivering a completely modulated carrier wave of 50 kw is shown in Fig. 232. The tube sequence is as follows:

50-watt crystal oscillator.
50-watt Class C amplifier.
50-watt modulated amplifier.
2 250-watt tubes as linear push-pull amplifiers.
2 water-cooled tubes as linear push-pull amplifiers.
6 water-cooled tubes as linear push-pull amplifiers.

The audio-frequency system is the same as in the 1-kw outfit, and in fact the only essential difference between the 50-kw and 1-kw transmitters is in the amount of linear amplification that is used following the modulated amplifier.

The schematic diagrams for two RCA broadcast transmitters are shown in Figs. 233 and 234.[1] These differ in construction from the Western Electric transmitters in that only one stage of linear amplifica-

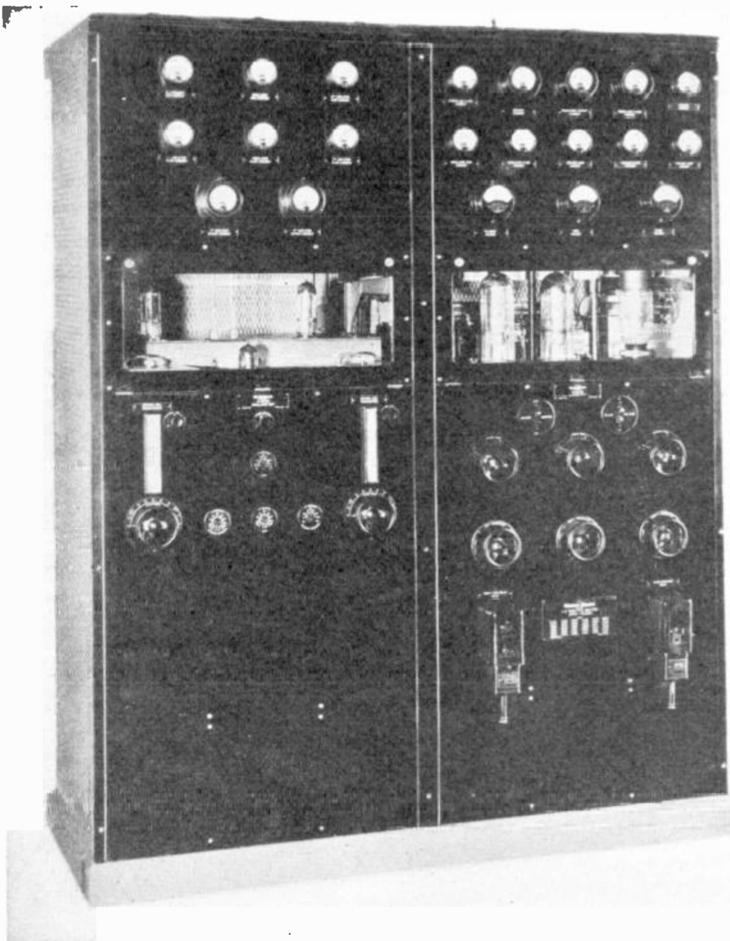


FIG. 231a.—Western Electric 1-kw broadcast transmitter—front view.

tion is used following the modulated amplifier. The tube sequence in the 100-watt transmitter of Fig. 233 is as follows:

7.5-watt triode crystal oscillator.
7.5-watt screen-grid buffer amplifier.
7.5-watt screen-grid Class C amplifier.

[1] For a more detailed description of these and other RCA-Victor broadcast transmitters see I. J. Kaar and C. J. Burnside, Some Developments in Broadcast Transmitters, *Proc. I.R.E.*, vol. 18, p. 1623, October, 1930.

2 7.5-watt triodes in parallel as modulated Class C amplifiers.
250-watt linear amplifier.

The audio-frequency system consists of a three-stage amplifier, the first two stages employing 7.5-watt triodes, and the final stage four

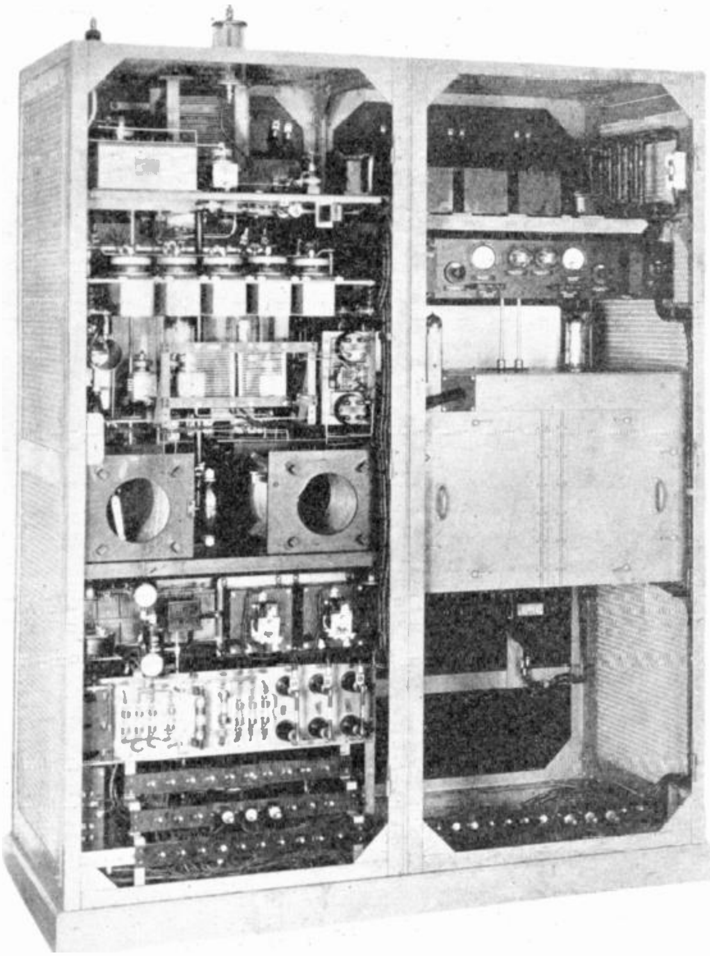


FIG. 231b.—Western Electric 1-kw broadcast transmitter—rear view.

Type 842 power tubes in parallel. The modulation is accomplished with the aid of a plate reactor common to both modulated and modulator tubes, while a voltage-dropping resistance in series with the modulated-amplifier tubes makes possible the complete modulation of the output.

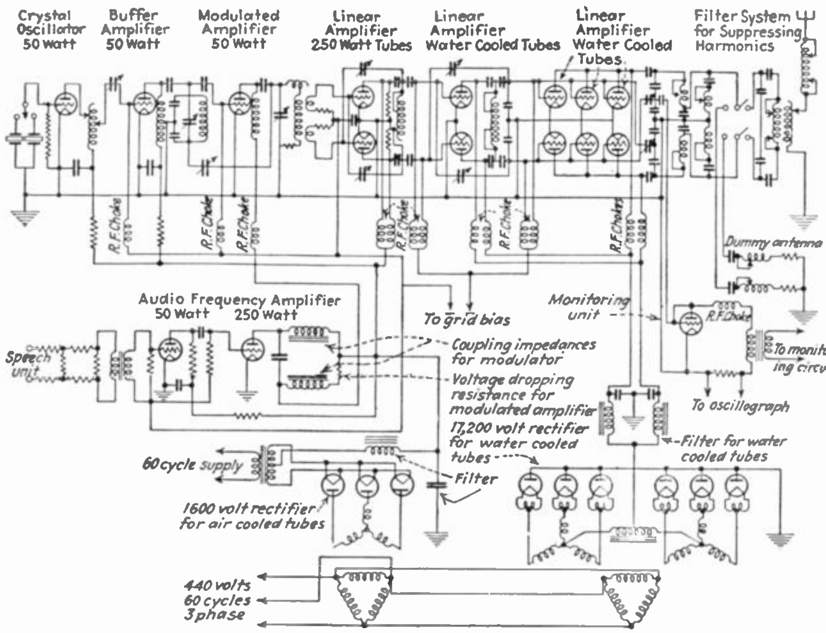


Fig. 232.—Simplified circuit diagram of Western Electric 50-kw broadcast transmitter.

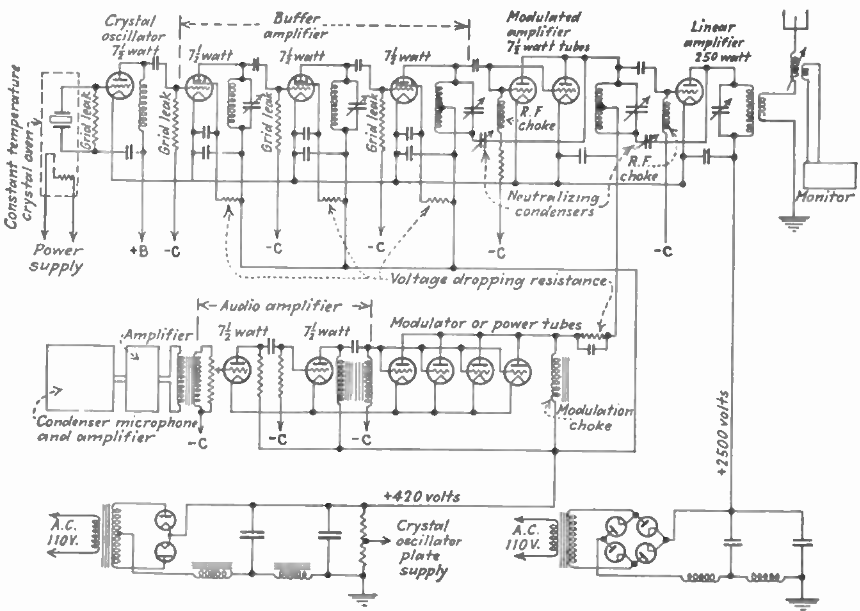


Fig. 233.—Circuit diagram of RCA 100-watt broadcast transmitter.

The schematic diagram of a 5-kw broadcast transmitter is shown in Fig. 235. The tube sequence in the radio-frequency chain is as follows:

7.5-watt crystal oscillator.
7.5-watt screen-grid buffer amplifier.
7.5-watt screen-grid Class C amplifier.
75-watt screen-grid Class C amplifier.
250-watt triode as a modulated Class C amplifier.
2 water-cooled tubes in push-pull as linear amplifiers.

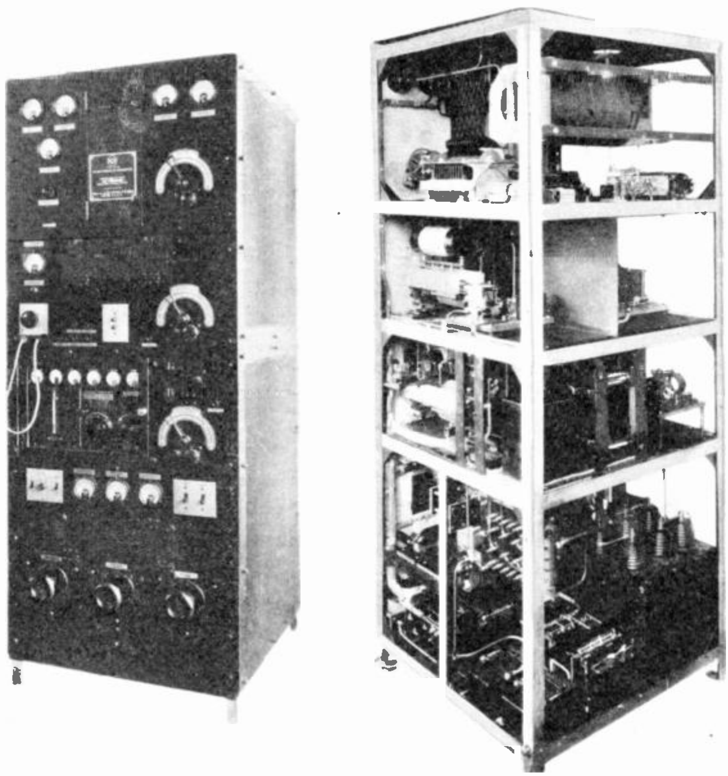


FIG. 234.—Photographs of RCA 100-watt broadcast transmitter.

The audio-frequency system consists of two stages of 50-watt tubes, followed by a stage of two 250-watt power tubes in parallel which completely modulate the one 250-watt Class C amplifier. The two triode stages in the radio-frequency chain are neutralized by the Rice method, and the output stage is provided with damping resistances to suppress the parasitic oscillations which tend to occur with this method of neutralization.

The principal difference between the two makes of transmitters that have been described is that in one the modulation is carried out at a relatively low power level, and several stages of linear amplification are employed, while in the other the modulation takes place at a higher power level, and there is only one stage of linear amplification. Proper attention to circuit conditions will give satisfactory results with either arrangement. Modulation at a low power level has the advantage of giving considerable economy in tubes, since much less audio-frequency

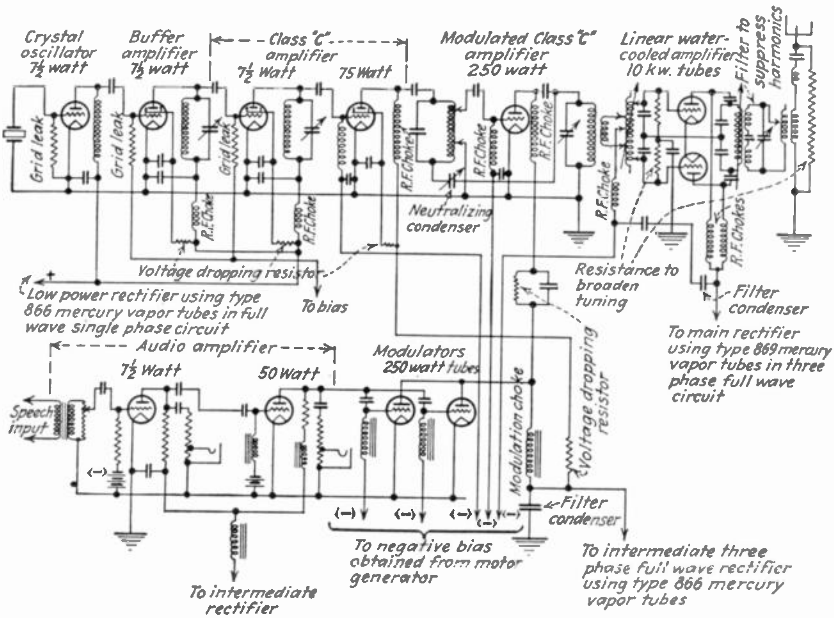


FIG. 235.—Simplified circuit diagram of RCA 5-kw broadcast transmitter.

power is required than when a large tube is to be modulated. This saving in audio-frequency power is obtained by increasing the number of stages of linear amplification, and as linear amplifiers tend to distort more than do audio amplifiers more care must be taken to give the same over-all performance when modulation takes place at a low power level.

The performance of a high-class commercial broadcast transmitter leaves little to be desired in range of frequencies which can be modulated upon the carrier, or in the avoidance of amplitude distortion. The frequency-response characteristic of the 50-kw transmitter of Fig. 232 is shown in Fig. 236 and is flat to within about 5 per cent over the frequency range 30 to 10,000 cycles. When completely modulated the harmonics generated in this transmitter do not exceed 5 per cent and are less with lower degrees of modulation. Performance such as indicated by these figures is of an altogether different order of magnitude of perfection than is obtained from even the very best broadcast receivers.

In the higher power broadcast transmitters extreme care is taken to avoid radiating harmonics of the carrier frequency. This is made necessary by the fact that when the carrier output is 50 kw a very small

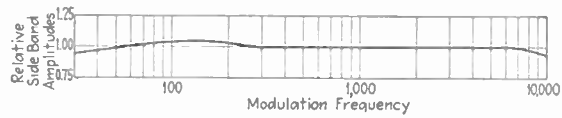


FIG. 236.—Relative side-band amplitude as a function of modulation frequency for the Western Electric 50-kw broadcast transmitter of Fig. 232.

percentage of harmonic radiation represents a considerable number of watts and so will cause interference at appreciable distances. Harmonics are prevented from reaching the antenna by the use of special coupling methods which operate to discriminate against the harmonics while permitting free passage of the desired carrier. In addition the transmitting equipment is shielded to prevent direct radiation.

Short-wave Radio-telephone Transmitters.—Short-wave radio-telephone transmitters are similar to broadcast transmitters except for modifications

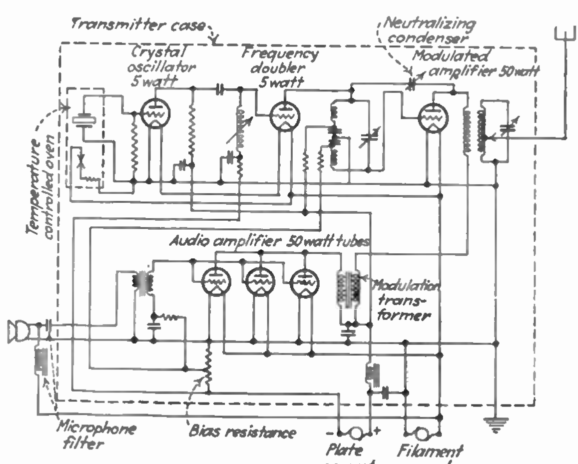


FIG. 237.—Circuit diagram of 50-watt short-wave Western Electric telephone transmitter designed for use on aircraft.

made necessary by the fact that crystals cannot be ground to generate extremely high frequencies. Schematic diagrams of typical short-wave radio-telephone transmitters are shown in Figs. 237 and 239. The first of these is an airplane transmitter capable of delivering 50 watts of completely modulated carrier and of operating in the frequency range 1500 to 6000 kc (200 to 50 meters). The transmitter consists of a 5-watt oscillator followed by a 5-watt frequency doubler which excites a 50-watt neutralized triode operating as a modulated Class C amplifier. The

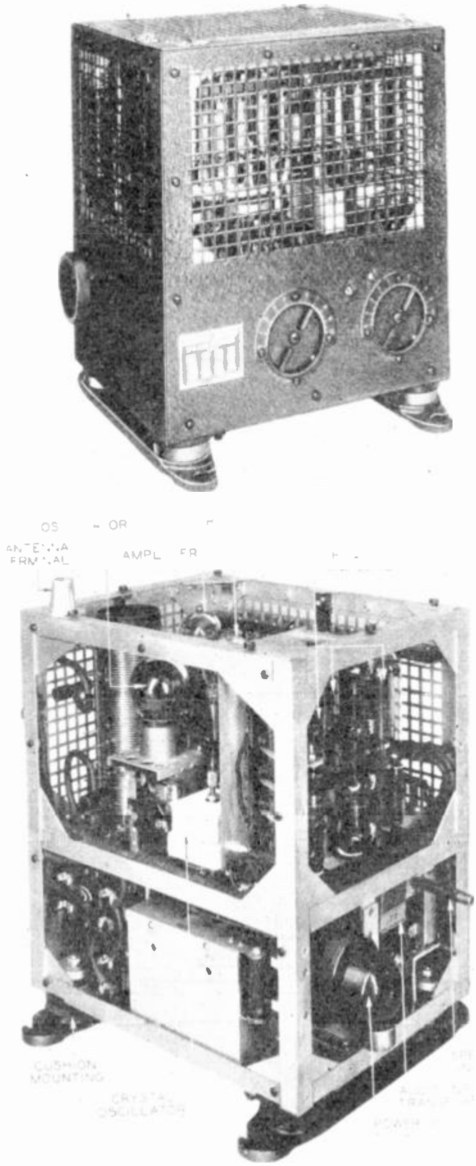


FIG. 238.—Photographs of the airplane short-wave telephone transmitter of Fig. 237.

audio-frequency system consists of three 50-watt modulator tubes in parallel with their grids excited directly from a sensitive microphone operating into a microphone transformer. The modulator and modulated tubes operate at the same plate voltage and employ the transformer circuit of Fig. 186*c*, in which the ratio of transformation is arranged so that the core of the modulation transformer suffers little direct-current magnetization, thus permitting the use of a small core. Photographs of the complete transmitter are shown in Fig. 238.

Figure 239 shows the circuit diagram of a typical low-power short-wave radio-telephone transmitter, such as used by radio amateurs. The required carrier frequency is obtained from a crystal oscillator followed by harmonic generators. A completely modulated Class C amplifier is employed, followed by a linear push-pull amplifier employing tubes rated at 75 watts. Since this amplifier is capable of developing about

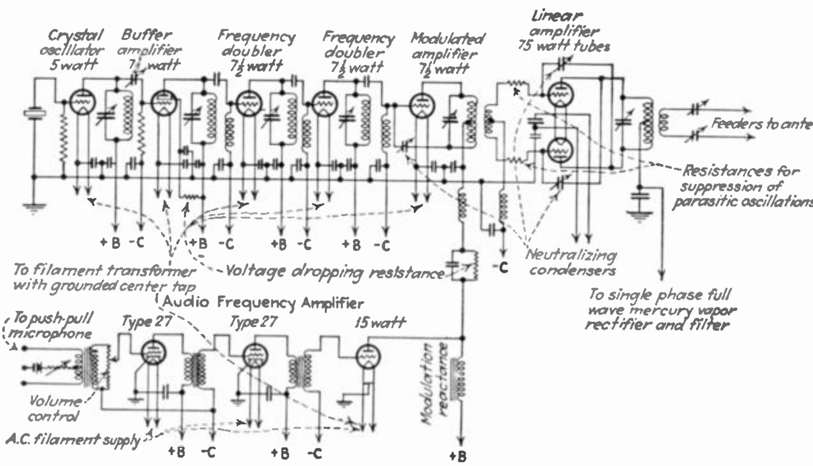


FIG. 239.—Circuit diagram of low-power short-wave telephone transmitter such as used by amateurs on wave lengths of 20 meters.

150 to 200 watts of output at the modulation crest, the carrier power is in the order of 37.5 to 50 watts when complete modulation is allowed for.

Long and Moderately Long-wave Radio-telephone Transmitters.—There are only a few radio-telephone transmitters in commercial service operating in the frequency range 500 to 125 kc. These transmitters are all of the master-oscillator power-amplifier type with the modulation introduced into one of the Class C amplifiers. An example of such a transmitter is shown in the schematic circuit diagram of Fig. 240, which is a United States Coast Guard 500-watt transmitter intended for use in the frequency range 125 to 500 kc, and can be employed for either telegraph or telephone communication. The transmitter consists of a 50-watt oscillator, an intermediate frequency amplifier made of three 50-watt tubes in parallel, and a 1-kw output amplifier. Telephone communication is obtained by modulating the plate circuit of the output amplifier by a 1-kw modulator tube, the grid of which is excited by a 50-watt audio-frequency amplifier tube.

The only radio-telephone transmitter operating on frequencies below 100 kc is the transatlantic long-wave radio-telephone system. This radio channel is unique in that it employs single side-band transmission.

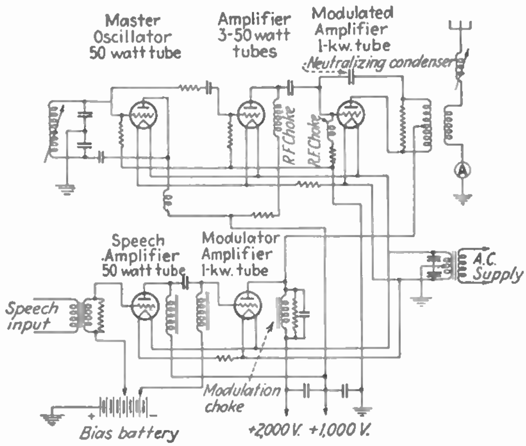


FIG. 240.—Circuit diagram of radio-telephone transmitter in which the frequency is controlled by a master oscillator. This transmitter is intended for use in the frequency range 125 to 500 kc, and is not arranged to give 100 per cent modulation.

The layout of the transmitter is shown schematically in Fig. 241. The speech input is applied to the balanced modulator shown in Fig. 196 and is modulated upon a carrier frequency of 33,700 cycles. The output of the balanced modulator contains no carrier wave and is passed through a filter that eliminates one of the side bands. The remaining side band

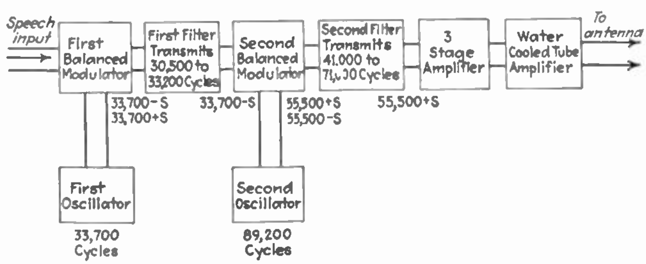


FIG. 241.—Schematic layout of long-wave transatlantic telephone, showing how the single side band is generated and amplified to a high-power level.

is then applied to a second balanced modulator and is modulated upon a carrier wave of 89,200 cycles. The resulting output is passed through a second filter which transmits only the lower side band. All of these operations take place at a power level in the order of 1 watt. The single side-band output of the second filter is amplified by a three-stage distortionless Class A amplifier consisting of a 5-watt, a 50-watt, and three

750-watt tubes, respectively, and then is further amplified by a linear amplifier consisting of 20 water-cooled tubes which are capable of delivering 150 kw of single side-band energy to the radiating system.[1]

One of the principal problems encountered in the design of moderate- and long-wave radio-telephone transmitters results from the fact that the ratio of side-band width to carrier frequency becomes greater as the carrier frequency is reduced. In order to maintain a uniform response for all side-band frequencies it is therefore necessary that the Q of the tuned circuits in which the modulated wave flows be proportional to the carrier frequency. At low carrier frequencies the Q must hence be relatively low (*i.e.*, the circuits must have high losses) if the higher modulation frequencies are not to be discriminated against. This problem is so acute in the case of the long-wave transatlantic telephone that the output of the final power amplifier is coupled to the antenna circuit by means of a band-pass filter arrangement, the secondary capacity of which is supplied by the antenna capacity. This gives a flat-topped and relatively broad response curve that makes it possible to transmit a single side-band 2500 cycles wide without undue distortion.

Miscellaneous Types of Radio-telephone Transmitters.—The equipment that has been described represents what is considered as the best practice for ordinary requirements. Other types of radio-telephone transmitters are, however, used to a limited extent under special circumstances. The first broadcast transmitters were all of the plate-modulated oscillator type (Heising constant-current system of modulation). This arrangement introduces excessive frequency modulation, however, because the oscillator frequency depends upon the plate voltage, and as a result has been superseded by arrangements employing a modulated amplifier excited by a crystal-controlled or master oscillator. The modulated oscillator is however still used extensively for experimental equipment where frequency modulation and frequency stability are not of fundamental importance. A small amount of commercial equipment makes use of the grid-modulated Class C amplifier instead of employing plate modulation, and if tubes with suitable characteristics can be constructed it is probable that this method of modulation will become of increasing importance. In experimental and laboratory equipment other methods of modulation, such as the grid-modulated oscillator, the absorption system of modulation, screen-grid modulated amplifiers, etc., are often employed.

[1] For further information on the long-wave transatlantic telephone one should consult the following articles: R. A. Heising, Production of Single Side-band for Trans-atlantic Radio Telephony, *Proc. I.R.E.*, vol. 13, p. 291, June, 1925; A. A. Oswald and J. C. Schelleng, Power Amplifiers and Trans-atlantic Radio Telephony, *Proc. I.R.E.*, vol. 13, p. 313, June, 1925; Ralph Bown, Trans-atlantic Radio Telephony, *Bell System Tech. Jour.*, vol. 6, p. 248, April, 1927.

106. Miscellaneous Features of Radio Transmitters.—Large radio transmitters are usually made so that the transmitter may be placed in operation by pressing a single push button. The circuit thus closed causes the various contacts to be made in the proper sequence and with the requisite time delay, and interlocks are provided to prevent damage to the transmitter if some part fails to function. Transmitters are also provided with a certain amount of relay protection which will take care of overvoltages, undervoltages, failure of the circulating water in water-cooled tubes, short circuits, etc. Most broadcast transmitters are provided with dummy antennas for use in testing and for warming up the transmitter. The dummy antenna consists of a resistance, inductance, and capacity, so proportioned as to represent the exact electrical equivalent of the radiating system. The transmitter itself acts just the same when delivering its output to the dummy antenna as when operating with the actual antenna, but no radiation takes place, and hence interference with other communications is avoided.

Simultaneous Two-way Telephone Conversation on the Same Frequency. Radio-telephone channels that are links in wire telephone systems sometimes make use of the same frequency for transmission in both directions. This presents special problems because the radio receiver that picks up the signals from the distant transmitting station is necessarily in close proximity to the local transmitting station which is operated on the same frequency. Trouble is avoided by making use of relays which switch the land lines to the transmitter or to the receiver, depending on whether the speech is coming into, or going out of, the radio system. The normal relay position connects the receiving equipment to the land lines and disconnects the transmitters. If, however, voice currents are received at the radio station from the land line a relay disconnects the receiving equipment from the land lines and places the transmitter on the circuit instead. In order that no speech will be lost during the brief interval when the relays are changing the connections, a time delay provided by an artificial line or a long speaking tube is placed between the point in the circuit where the voice energy operates the relay and the point where the relay contacts change the connections. A schematic diagram of the circuits involved is shown in Fig. 242.[1]

Parasitic Oscillations.—In transmitters which involve a number of tubes there is a great tendency for parasitic oscillations to be generated. Such oscillations are of high frequency and are the result of stray couplings, and resonant circuits formed by tube capacities, connecting wires, etc. Such oscillations absorb energy that would otherwise be available

[1] Arrangements of this type are used in most radio links which attempt to provide two-way talking service, even when different transmission frequencies are used in the two directions. For example, see G. C. Crawford, Echo Elimination in Trans-atlantic Service, *Bell Lab. Record*, November, 1927.

at the frequency of transmission, and may introduce distortion. Parasitic oscillations are particularly troublesome in transmitters employing large tubes because of the high mutual conductance, the high electrode capacities, and the long leads that go with such tubes. The tendency for undesired oscillations to be produced increases with the number of tubes, particularly when tubes are operated in parallel. Parasitic oscillations are suppressed by placing resistances at various places in the circuits, by using care in the arrangement of parts, and by providing short connecting leads. Copper shields can also be used to advantage. The problem of parasitic oscillations limits the number of tubes that can be operated satisfactorily in parallel and makes it desirable to employ one large tube instead of several smaller ones in parallel in obtaining a given output.

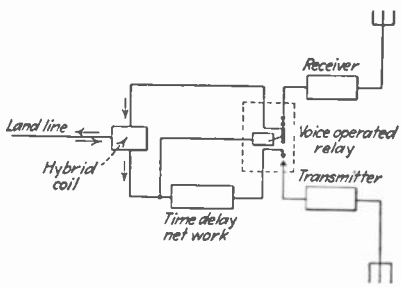


FIG. 242.—Circuits involving voice-operated relay and time-delay network for making possible two-way radio-telephone communication using the same frequency for transmission in both directions.

Monitoring.—Satisfactory performance of radio transmitters is checked by monitoring, to make certain that the signals are of satisfactory quality and that the assigned transmission frequency is being maintained. The frequency can be checked by means of a wave meter, or by comparison with a secondary standard of frequency embodying a crystal oscillator. In the case of crystal-controlled transmitters the monitoring need be done only occasionally, but when master-oscillator, or self-oscillator, short-wave transmitters are employed it is necessary to check the frequency every few minutes.

In code transmitters the character of the transmitted signals is usually observed at the transmitting station by listening on an insensitive well-shielded radio receiver. Where a number of transmitters can be monitored from one point it is also common practice to provide a tape recorder or an oscillograph to be used in conjunction with the monitoring receiver to make more accurate records of the code characters as transmitted. In high-speed work such automatic methods of recording must be used to detect faulty keying.

In radio-telephone transmitters the character of the signals being sent out is checked by rectifying a small part of the energy delivered to the radiating system, and viewing the rectified current in an oscillograph, or delivering it to a loud-speaker so that the signals can be observed by ear. All modern broadcast transmitters include a monitoring rectifier as part of the transmitter equipment (see Figs. 230 and 232, for example). In operating radio-telephone transmitters it is necessary to maintain continuous observation of the power level of the speech input to the audio system. The volume should be adjusted so that the audio power is sufficient to modulate the transmitter completely during the loud passages.

Tubes Sizes.—The relative size of the tubes in the different stages of a radio transmitter is determined by the voltage amplification that is obtainable from a single tube and depends upon the constants of the tube, the operating voltages, the losses in the plate tank circuit, and the grid losses of the tube receiving excitation from the tank circuit. Because of the many factors involved it is impossible to lay down any exact rules for the proper ratings, but in general it is considered best practice to avoid using successive tubes having widely different power ratings. This is particularly true in master-oscillator transmitters, where a master oscillator that is small in comparison with the power amplifier will be found to generate a frequency that is affected to a marked extent by the amplifier adjustments. The usual practice is to use successive tubes having ratings that are in the ratio of about 5 or 10 to 1 for the low power stages and to make use of a somewhat higher ratio where water-cooled tubes are involved.

In tubes used as frequency doublers it is difficult to obtain a second-harmonic output voltage that is appreciably greater than the fundamental-frequency exciting voltage, and when higher order harmonics are being generated the output voltage will normally be much less than the exciting voltage. Harmonic-generator tubes cannot, therefore, be counted upon to amplify the voltage appreciably even under favorable circumstances.

107. Ultra Short-wave Transmitters.—Wave lengths shorter than 10 meters (30,000 kc) have few commercial uses because the ionized layer in the upper atmosphere will not reflect these high-frequency signals earthward, and as a result communication cannot be carried on satisfactorily around the curvature of the earth. A great deal of experimental work is being done at these very high frequencies, but up to the present time there has been no tendency towards a standardized type of transmitting equipment. For this reason, and also because the technique is continually changing, no detailed description will be given of equipment suitable for transmitting on very short waves.[1]

Wave lengths in the range from 1 to 10 meters (30,000 to 300,000 kc) can be generated by oscillators of the tuned-circuit type modified in minor respects as made necessary by the very high frequencies involved. Most of the experimental transmitters operating within this frequency range use ordinary triode oscillators delivering their output directly to the radiating system. Crystal-controlled generators, master oscilla-

[1] The reader who desires to become acquainted with this field should consult the latest technical magazines for information.

tors, and power amplifiers are seldom used in experimental transmitters below 10 meters because of the extreme difficulties involved.

The only satisfactory method that has been devised for generating radio waves of a length less than about 1 meter is by the use of electron oscillations, using either triodes or magnetrons, as explained in Sec. 58 and 81, respectively, and all of the transmitters for frequencies above 300,000 kc (waves shorter than 1 meter) make use of such oscillators directly connected to the radiating system. Means have been devised for modulating the electron oscillations so that radio-telephone signals as well as telegraph code characters can be transmitted.

CHAPTER XIII

RADIO RECEIVERS

108. Broadcast Receivers—General Considerations.—The usual broadcast receiver consists of a number of stages of radio-frequency amplification, a power detector, and an audio-frequency amplifier capable of delivering one or more watts of undistorted power to the loud-speaker. The radio-frequency amplifier may be either of the tuned-circuit type, in which the resonant frequency is adjusted to the frequency to be received, or of the superheterodyne type, in which the frequency of the incoming signal is changed to a predetermined value that is amplified by a fixed amplifier. The amount of radio-frequency amplification is normally such that the voltage which is applied to the grid of the detector tube is in the order of 50,000 to 500,000 times the voltage that is induced in the antenna. Such amplification develops a detector input sufficiently large to give power detection with its accompanying advantages of low distortion and high efficiency.

Sensitivity, Selectivity, Fidelity.—The important characteristics of a broadcast receiver are the sensitivity, the selectivity, the fidelity, and the power capacity. The sensitivity represents the ability of the receiver to respond to small radio-signal voltages, and is measured quantitatively in terms of the voltage that must be induced in the antenna by the radio signal to develop a standard output from the power amplifier. This standard output has been arbitrarily chosen as 0.05 watt in a non-inductive load resistance having a value corresponding to the load resistance into which the power amplifier is designed to operate, and the sensitivity is arbitrarily defined as the effective value of the carrier voltage that must be induced in the antenna to develop this standard output when the carrier is modulated 30 per cent at a frequency of 400 cycles. The sensitivity is measured with the radio receiver tuned to give maximum response at the carrier frequency involved and with the volume controls adjusted for maximum volume. Inasmuch as broadcast receivers may be used with all types of antennas, and the performance is usually somewhat affected by the antenna characteristics, it has been found desirable to compare different radio receivers on the basis of a standard antenna possessing an inductance of 20μh, a capacity of 200$\mu\mu$f, and a resistance of 25 ohms. Typical curves showing the sensitivity of broadcast receivers as a function of carrier frequency are shown in Figs. 244

and 253. The sensitivity of broadcast receivers is seldom higher than several microvolts, and in the less sensitive sets is in the order of $100\mu v$. The highest sensitivity that can be usefully employed under even the most favorable circumstances is in the order of $1\mu v$, since the voltages developed in the input circuit to the first tube by thermal agitation are sufficiently loud to drown out signals of lower intensity. In cities and in the neighborhood of electrical appliances, man-made interference raises the noise level to much higher values, and the maximum usable sensitivity may be as low as $100\mu v$.

Selectivity is the property that enables a radio receiver to discriminate between radio signals of different carrier frequencies. Selectivity cannot be defined in a single term as can sensitivity but must be expressed in the form of curves, such as those of Figs. 244 and 253, which show the amount by which the signal input must be increased in order to maintain the standard output as the carrier frequency is varied from the frequency to which the receiver is tuned. These curves therefore indicate the extent to which interfering signals are discriminated against, and in general will depend somewhat on the carrier frequency.

Fidelity is the accuracy with which the radio receiver reproduces the intelligence contained in the modulated wave of voltage that is induced in the antenna system. Strictly speaking, fidelity should include frequency, amplitude, and phase distortion, but when applied to radio receivers is taken to refer only to frequency distortion. The fidelity of a radio receiver is expressed in curves such as those of Figs. 244 and 253, which give the variation in audio-frequency output voltage as the modulation frequency of the wave is varied, and will in general vary somewhat with the carrier frequency. In order to facilitate comparison the output is always expressed in terms of the ratio of actual output to the output obtained when the modulation frequency is 400 cycles.

The power capacity of a radio receiver, while not subject to exact definition because "allowable distortion" is not an objective term, is usually controlled by the undistorted output obtainable from the power amplifier and so can be estimated from the tube characteristics, as discussed in Sec. 39. It is possible, however, for the power limit to be fixed by the detector, or even the radio-frequency amplifier, in which cases the undistorted output available will be less than the rating of the power tube.

The sensitivity, selectivity and fidelity curves shown in Figs. 244 and 253 are determined by connecting the standard artificial antenna in place of the usual antenna, and by introducing in series with the artificial antenna sufficient carrier voltage (modulated 30 per cent at 400 cycles) to develop the standard output. Apparatus for generating small carrier voltages of known amplitudes and known degrees of modulation is commercially available and reduces the determination of receiver performance to a routine testing operation.[1]

Constructional Features.—Broadcast receivers are ordinarily mounted upon a chassis of sheet iron bent in the form of an inverted tray and suitably punched to receive the sockets, coils, transformers, etc., which are eyeleted, riveted, or bolted in place. The chassis is slipped into a cabinet, which is essentially a piece of furniture rather than an electrical device. The loud-speaker is usually mounted directly on the cabinet and is connected to the chassis by a flexible cord terminated in a plug.

The coils used in the tuned circuits are usually of small size and are shielded by means of aluminum or copper cans. The variable tuning condensers are nearly always mounted on a common shaft and are usually shielded from each other by means of metal shields placed along the shaft between the individual units. The individual radio-frequency amplifier tubes are also shielded from each other by some means. Generous use is made of by-pass condensers in the radio-frequency circuits in order to short-circuit common couplings that would otherwise exist between the different stages of the radio-frequency amplifier. Care is also taken in the arrangement of parts to insure that a minimum of stray coupling will exist between various amplifier stages.

The power transformer, filter inductances and condensers, rectifier tubes, and power amplifier are usually collected together to form what is called a "power pack," which is either segregated in one part of the main chassis or mounted in a separate unit.

The wiring of a radio receiver is arranged as far as possible in such a way as to separate those circuits which carry radio-frequency currents from those which are at ground potential. This is accomplished by making the grid and similar high-potential leads as short and direct as possible, and by-passing all other connections directly to the chassis by suitable condensers. In general the high-potential leads are kept above the chassis, while the connections that are at ground potential as far as radio-frequency currents are concerned are kept underneath the chassis and are made of flexible insulated wire run without regard to the position of other wires. Any radio-frequency lead which must go underneath the chassis is made as short as possible and is either shielded or kept some distance from the other wires.

Photographs showing these general methods of construction in typical cases are shown in Figs. 246 to 250 and 252.

109. Typical Broadcast Receivers.—Modern broadcast receivers are virtually all of the tuned radio-frequency or the superheterodyne type.

[1] For detailed information dealing with the proper procedure for conducting receiver tests one should consult the reports of the Committee on Standardization which are published in the *Year Book of the Institute of Radio Engineers.* The 1931 *Year Book* gives a complete discussion on page 131.

The typical tuned radio-frequency receiver usually consists of three stages of radio-frequency amplification making use of screen-grid tubes, a power detector of the plate-rectification type, and a push-pull power amplifier. Heater-type tubes are employed throughout except for the power amplifier, which uses alternating current directly on the filaments. The tuned radio-frequency amplifier employs one of the screen-grid circuits shown in Fig. 169 or some modification thereof, and the coils in the different stages are carefully matched to have the same inductance, so that the radio-frequency amplifier circuits can be tuned from a single control. The final lining up of the tuned circuits is done after the set

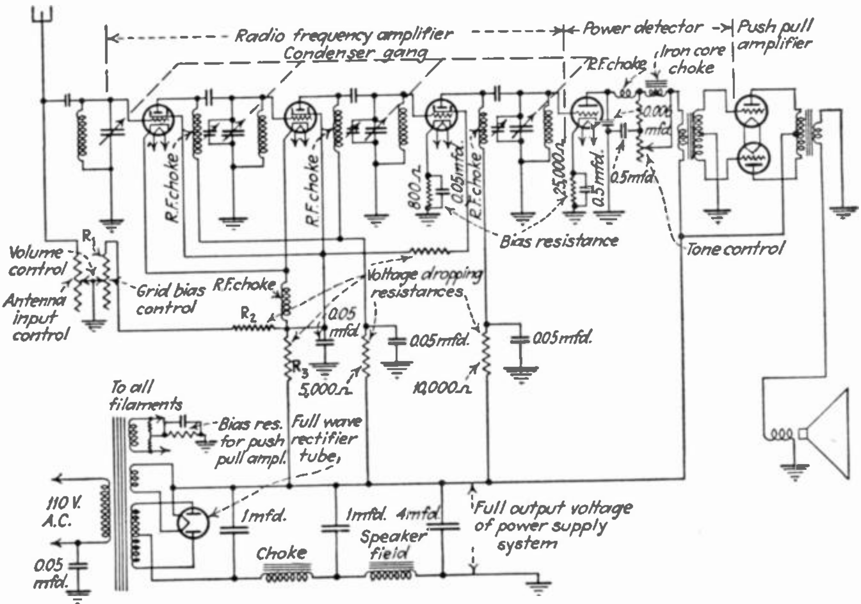


FIG. 243.—Circuit diagram of typical tuned radio-frequency type of broadcast receiver.

is constructed and is accomplished by the use of small adjustable "trimming" condensers or by bending the plates of the variable condensers in such a way as to make all circuits tune together. The input to the first radio-frequency amplifier tube usually consists of a tuned circuit operated from the same single control as are the other circuits and loosely coupled to the antenna circuit through a small coupling inductance or a small condenser. The detector is almost universally a plate power rectifier, making use of a Type '27 general-purpose or Type '24 screen-grid heater tubes. The output of the detector must be provided with a filter to prevent radio-frequency voltages from reaching the grid of the succeeding audio-frequency amplifier tube and causing overloading. Such filters are shown in Figs. 243, 245, and 251 and usually consist of a radio-frequency choke coil in series with the plate of the detector tube and shunted to ground by condensers which offer low impedance to broadcast frequencies while having a high impedance at audio frequencies.

Typical Tuned Radio-frequency Type of Receivers.—The circuit diagram of a typical tuned radio-frequency receiver is shown in Fig. 243. The receiver consists of a three-stage tuned radio-frequency amplifier using the shunt-feed circuit of Fig. 169*c* and having a tuned input circuit capacitively coupled to the antenna. A grid-bias type of power detector is employed, which operates directly into a push-pull power amplifier. The anode power is obtained from a Type '80 full-wave rectifier tube, the output of which is passed through a two-section filter having a condenser input. The second inductance of this filter is supplied by the field winding of the speaker. The full voltage of the rectifier-filter system is applied to the plates of the detector and power tubes, while the plates and screen grids of the screen-grid tubes are supplied through suitable voltage-dropping resistances. The grid bias for the power tubes is obtained by means of a resistance between the filament center tap and ground, which makes the filaments more positive than ground by the voltage drop which the plate current of the power tubes produces in this resistance. The detector and third screen-grid tube are self-biased by separate resistances between each cathode and ground, while the bias on the control grids of the first two screen-grid tubes is controlled by the voltage drop produced across R_1 and R_2 by the current which the rectifier output produces in the circuit $R_1R_2R_3$. This bias can be varied by the resistance R_1, which serves as the volume control. The heaters are connected in parallel with the filaments of the power tubes and so are positive with respect to ground, which tends to reduce the alternating-current hum.

Broadcast receivers of the tuned radio-frequency type made by different manufacturers will usually differ in some details from the above example. Most of these variations are in the radio-frequency amplifier, which may contain four stages instead of three, may use one of the other types of screen-grid radio-frequency amplifier circuits shown in Fig. 169, and so on. Band-pass selector circuits of the type discussed in Sec. 20 are sometimes used between the first radio-frequency tube and the antenna in order to reduce cross-talk. The power detector frequently makes use of a heater-type screen-grid tube instead of the heater-type general-purpose tube of Fig. 243, and in some cases a stage of audio-frequency voltage amplification is placed between the detector and the power tube. The details of the power-supply system and of the arrangements for producing grid-bias voltages will differ to a considerable extent in different makes of receivers.

The performance of a typical tuned radio-frequency type of broadcast receiver is shown at Fig. 244. The selectivity curves show a higher selectivity at the low-frequency end of the broadcast band than at the high-frequency end, which is typical of nearly all tuned radio-frequency

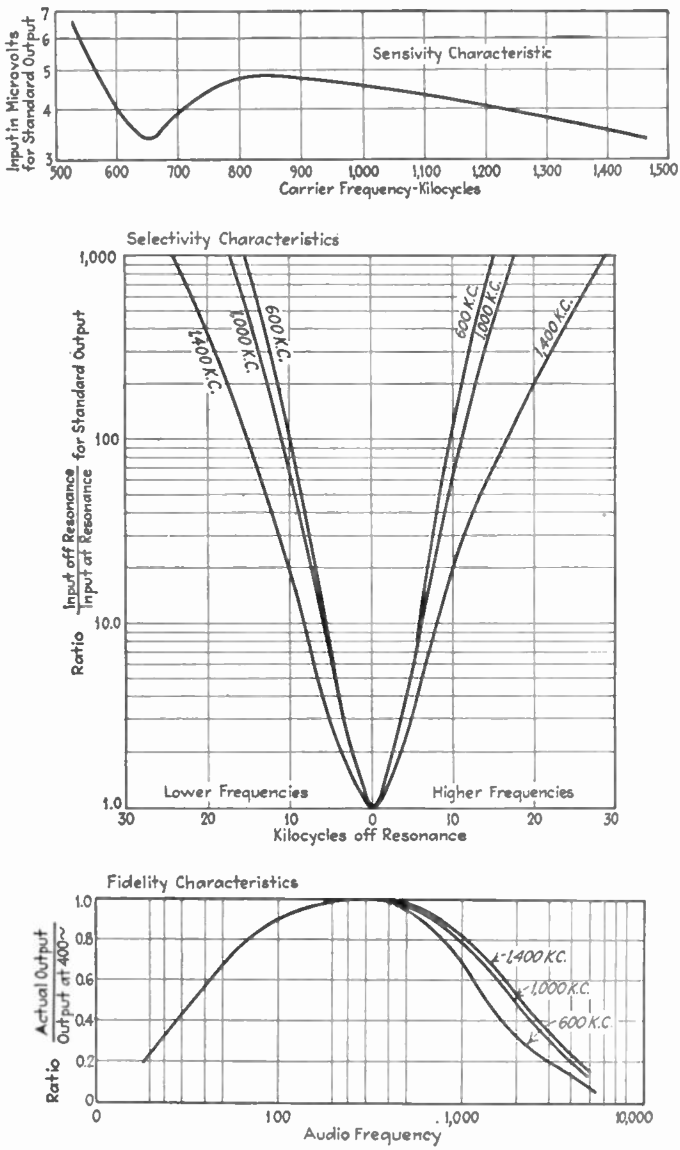


FIG. 244.—Sensitivity, selectivity, and fidelity curves of a tuned radio-frequency type of broadcast receiver.

receivers. This characteristic results from the fact that the effective Q of the coils in the amplifier tends to be independent of frequency, causing the band width to which the receiver responds to increase at the higher frequencies. The fidelity curves of Fig. 244 show a distinctly poorer response at the low-frequency end of the broadcast band than at the high-frequency end. This is typical of receiving sets employing tuned radio-frequency amplification, and results from the fact that the selectivity at the low-frequency end of the broadcast band is high enough to discriminate against the high side-band frequencies of the radio signal, while at the high broadcast frequencies the circuits tune so broadly that there is relatively little reduction of the higher audio frequencies, and the fidelity curve approximates the amplification characteristic of the audio-frequency system.

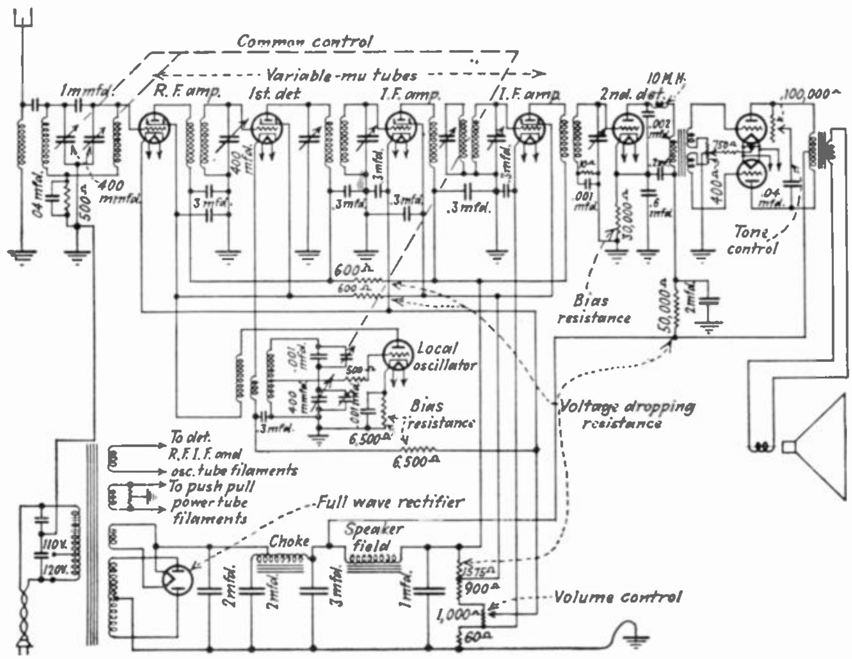


Fig. 245.—Circuit diagram of typical superheterodyne type of broadcast radio receiver.

Typical Superheterodyne Type of Receivers.—In superheterodyne receivers the input from the antenna is usually passed through a one-stage tuned radio-frequency amplifier which is sufficiently selective to eliminate signals that differ greatly from the desired frequency. The output of this amplifier is heterodyned with a local oscillator adjusted to give a difference (or "intermediate") frequency in the range 150 to 200 kc. The heterodyne signal that results from this combination is then rectified by an anode rectifier (usually called the "first detector" or "mixer,"), the output of which represents the original signal with its carrier frequency changed to a value corresponding to the difference frequency between the local oscillation and the original carrier, as explained in Sec. 65. This beat-frequency output of the first detector is then amplified by a so-called intermediate-frequency amplifier, which is a non-adjustable radio-frequency amplifier, usually of the band-pass type. The output of the intermediate-frequency amplifier is rectified by a power detector of the anode type and delivered directly to the power amplifier.

The circuit diagram of a typical superheterodyne receiver is shown in Fig. 245. This receiver has one stage of tuned radio-frequency amplification, two stages of intermediate-frequency amplification, a power detector of the bias type, and two Type '45 power tubes in push pull. A band-pass filter circuit is provided between the antenna and the first tube to minimize cross-talk trouble. The intermediate-frequency ampli-

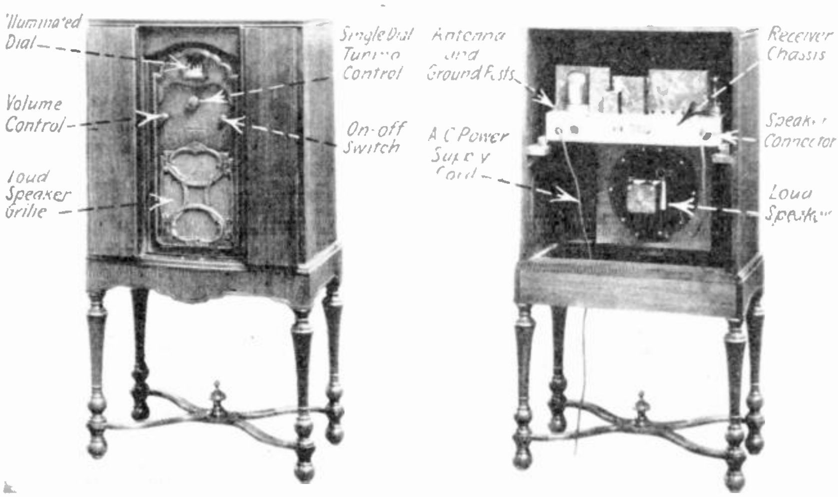


FIG. 246.—Front and rear views of receiver cabinet, showing the method of mounting chassis and loud-speaker.

fier contains two band-pass circuits and one simple tuned secondary transformer-coupled stage as seen in the figure. The intermediate-frequency transformers are compact air-core coils shown in Fig. 249, with small adjustable condensers for tuning.

The anode power is obtained from a rectifier-filter arrangement similar in the main to the power supply of the receiver of Fig. 243, except that a tapped first inductance is employed. The grid bias for the radio-frequency amplifier, first detector, and first intermediate-frequency amplifier are obtained from a potentiometer across the plate supply, the adjustment of which provides the volume control. These tubes, together with the second intermediate-frequency amplifier, which is operated with a fixed grid-bias potential, are of the variable-mu type. The local oscillator and second detector are biased by a resistance in series with the cathode, while the power tubes are provided with a resistance between ground and the center point of the filaments. The adjustment marked "hum balance" is for the purpose of balancing out the hum which would otherwise be present if the push-pull tubes had unlike characteristics.

Photographs of this receiver are given in Figs. 246 to 250, and serve to bring out the important constructional features common to all receivers.

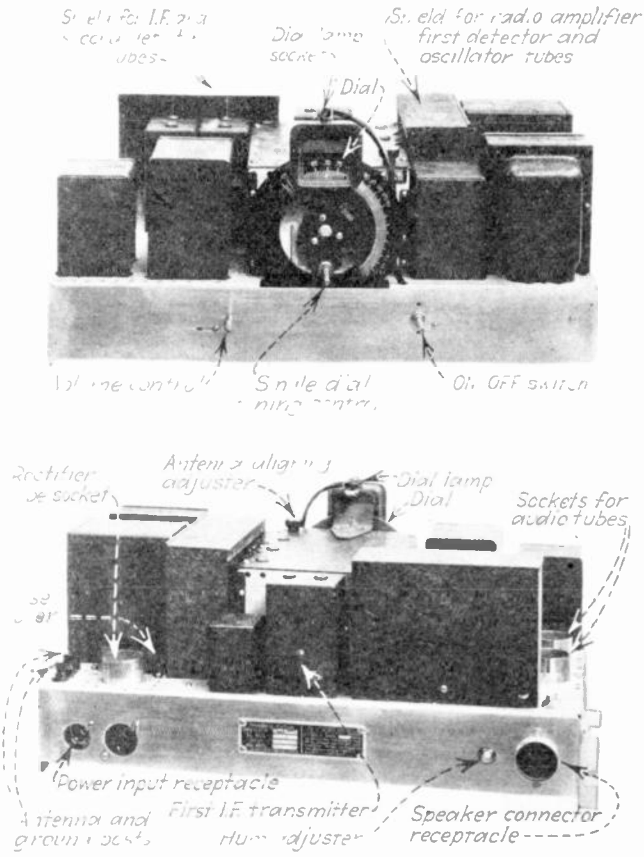


FIG. 247.—Front and rear views of chassis of the superheterodyne receiver of Fig. 245.

The circuit diagram of another superheterodyne receiver is shown in Fig. 251. This receiver is of the midget variety, designed to sell for a low price, and so is built in the simplest possible manner that is consistent with good performance. The receiver consists of one stage of transformer-coupled tuned radio-frequency amplification, a screen-grid first detector, a single stage of band-pass intermediate-frequency amplification, a power detector, and a push-pull power amplifier. Anode power is obtained from a full-wave rectifier tube associated with a one-

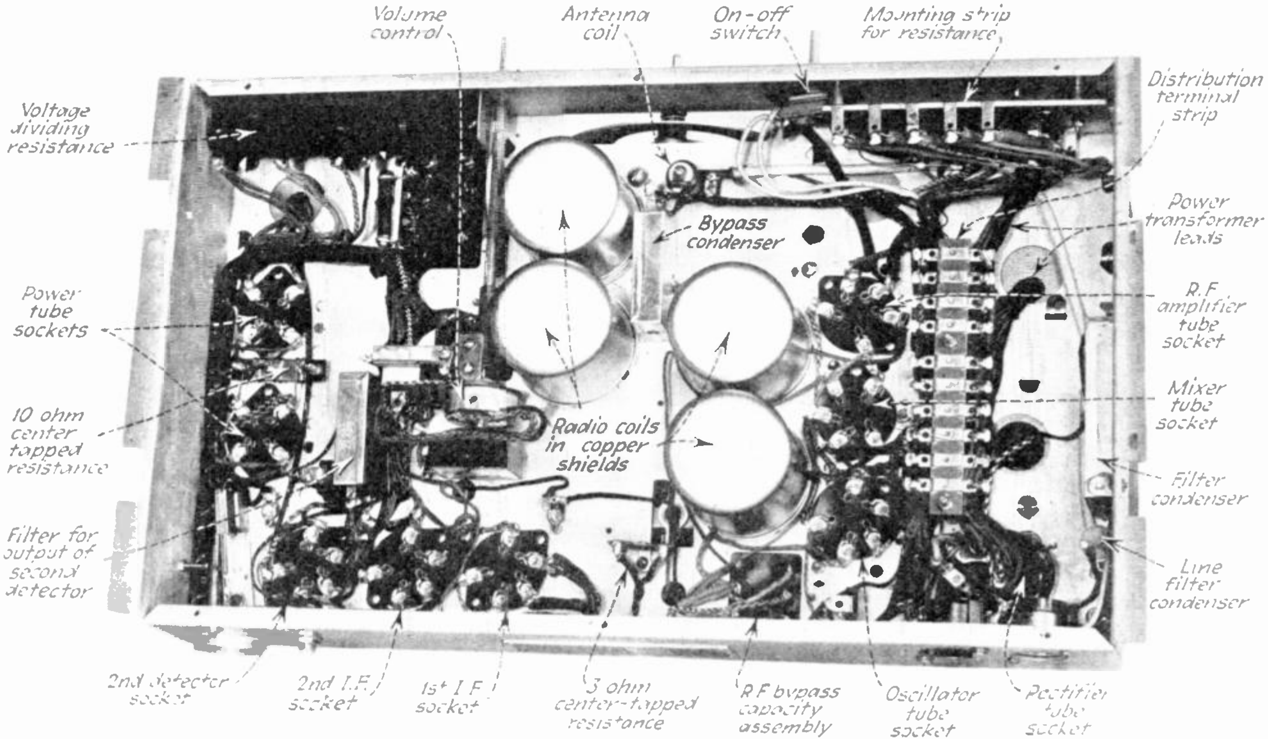
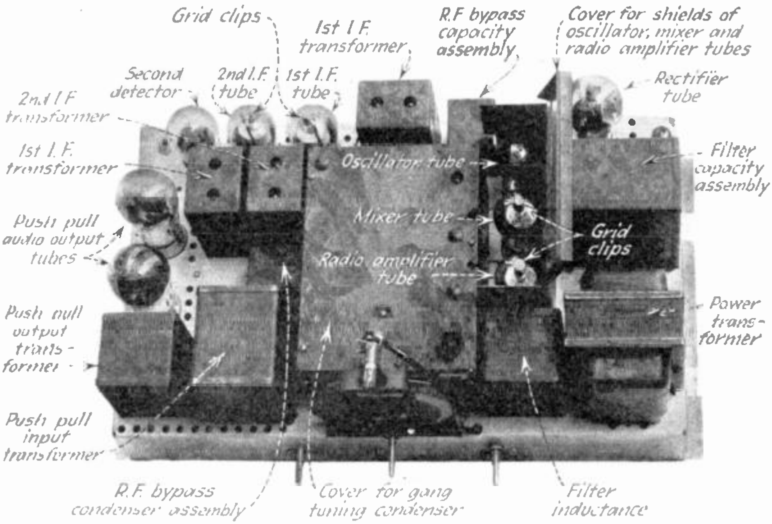


FIG. 248a.—Bottom view of chassis of receiver of Fig. 245, with cover removed.

section filter using the speaker field as the inductance. The radio-frequency and intermediate-frequency amplifier tubes are of the variable-mu type and are used to control volume by means of a variable resistance



Top view.
FIG. 248b.—Top view of chassis of receiver of Fig. 245.

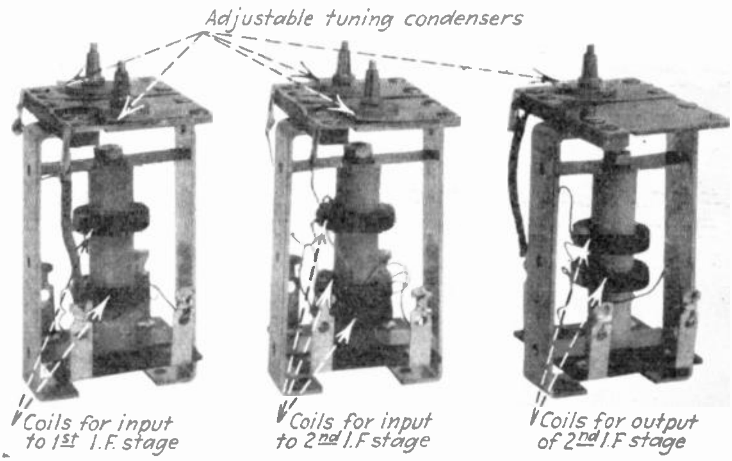


FIG. 249.—Intermediate-frequency transformers used in superheterodyne receiver of Fig. 245.

which operates from the power-supply system and makes the two cathodes positive with respect to the grids. The remaining heater-type tubes

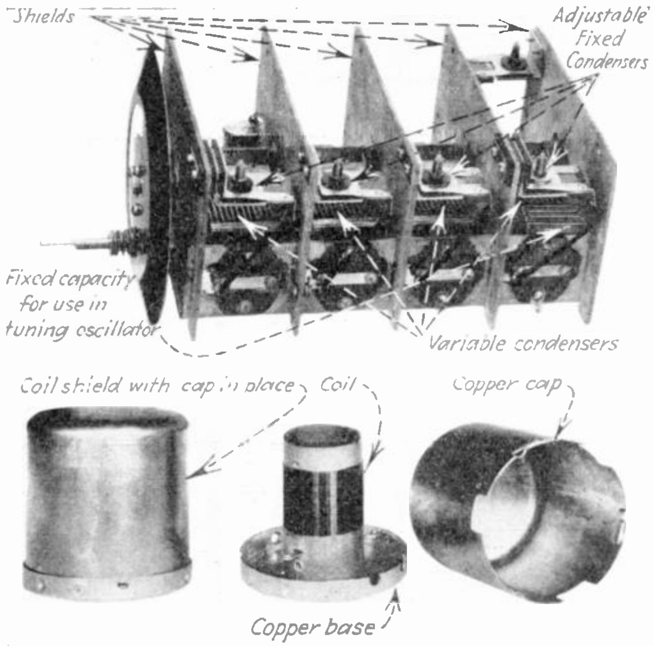


FIG. 250.—Condenser gang of four condensers on a common shaft, and shielded tuning coils used in superheterodyne of Fig. 245.

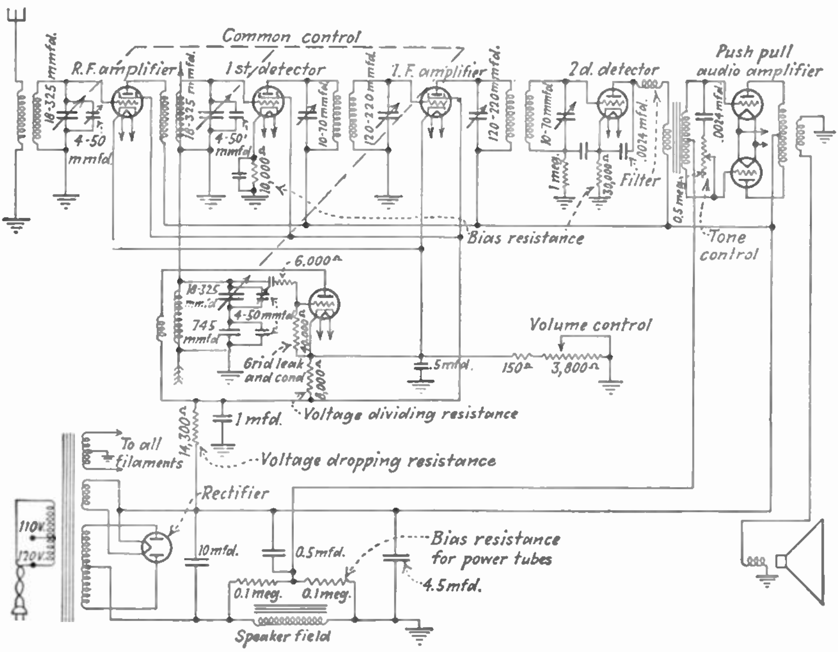


FIG. 251.—Circuit of a superheterodyne receiver of the midget type.

are biased by resistances between cathode and ground, while the negative grid voltage for the power amplifier is obtained from the direct-current voltage drop in the filter inductance by a voltage-dividing resistance placed across the terminals of the speaker field. Photographs of the chassis are shown in Fig. 252 and clearly bring out the very great compactness of this receiver.

A number of special problems are involved in the design of superheterodyne receivers as a result of the frequency-changing operation

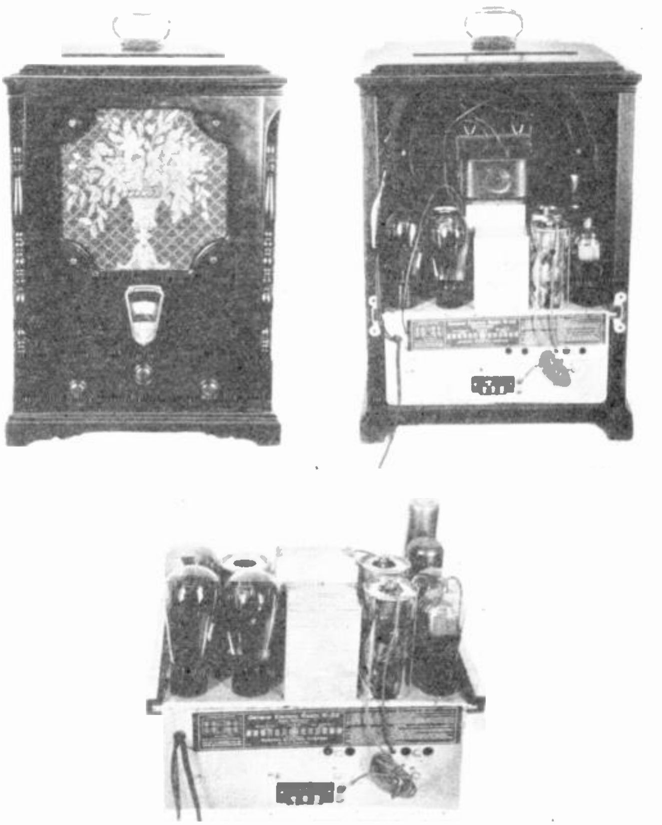


FIG. 252.—Views of midget superheterodyne receiver of Fig. 251. Note the compactness.

taking place in the receiver. For example it will be observed that for each local oscillator frequency there are two possible carrier frequencies which will give the same difference frequency. Thus if the intermediate-frequency amplifier operates at 180 kc, a desired signal having a frequency of 1000 kc will be received when the local oscillator is adjusted to 1180 kc, but such a local oscillator frequency will also heterodyne with a 1360-kc carrier to produce a 180-kc difference frequency. One

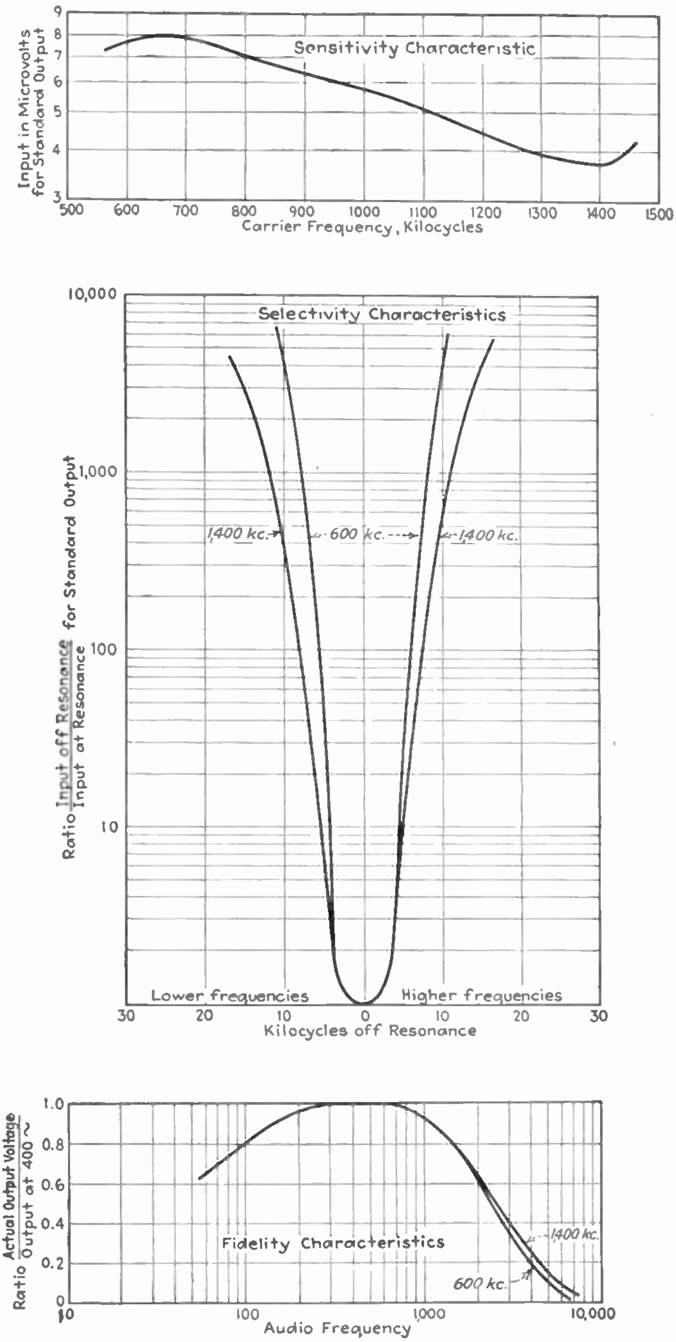


FIG. 253.—Typical sensitivity, selectivity, and fidelity curves of superheterodyne receiver. Note the difference between these selectivity and fidelity curves and those shown in Fig. 244 for a tuned radio-frequency receiver.

of the chief functions of the tuned radio-frequency input amplifier is to prevent simultaneous reception of two stations in this way. By tuning this amplifier to the desired signal, the undesired or "image" frequency is discriminated against, and if the intermediate frequency is not too low, *i.e.*, is in the order of 175 kc, the unwanted interfering signal is prevented from reaching the first detector. The input amplifier also serves to prevent energy of the local oscillator from reaching the antenna, where it would be radiated and interfere with other receivers. In carrying out the frequency conversion it is necessary to arrange the local oscillator so that the voltage delivered to the first detector is approximately constant at all frequencies and has a value that will give satisfactory heterodyne detection. The fact that the intermediate-frequency oscillator must operate at a frequency that differs from, but is exactly related to, the frequency being received introduces complications where it is desired to do all the tuning from a single control. This problem is ordinarily handled by placing a fixed condenser in series with the variable tuning condenser, and also shunting this condenser by a small capacity. When these shunting and series condensers are suitably chosen, and the oscillator inductance coil is made the proper size, the tuning condenser used to vary the oscillator frequency can be identical with the condensers in the tuned radio-frequency amplifier stages and the proper difference frequency will be obtained at all times.

Curves showing the sensitivity, fidelity, and selectivity of a typical superheterodyne receiver are shown in Fig. 253. When compared with the corresponding curves of the tuned radio-frequency receiver several outstanding differences will be observed. While the sensitivity curve is relatively flat in both cases it will be observed that the selectivity and fidelity of the superheterodyne receiver are both substantially independent of carrier frequency. The reason for this is that the selectivity of the superheterodyne receiver is determined primarily by the intermediate-frequency amplifier, which by the use of band-pass circuits can be made to have a substantially uniform response over the desired frequency band, while sharply discriminating against all other frequencies. The slight variation in selectivity and fidelity with frequency, which is observed in Fig. 253, is caused by the tuned circuit between the antenna and the first detector.

Miscellaneous Comments.—A comparison of the tuned radio-frequency and superheterodyne types of broadcast receivers shows that the latter is capable of giving a much higher quality performance but does so at the expense of complications and extra tubes introduced by the frequency-changing operation. It is easier to obtain high sensitivity in the superheterodyne type of receiver because large amplifications are easier to manage at the intermediate frequency than at broadcast frequencies. The fact that the superheterodyne receiver obtains most of its radio-frequency amplification and most of its selectivity from a fixed amplifier operating at a low radio frequency also makes it relatively easy to obtain a sensitivity, selectivity, and fidelity that are constant throughout the broadcast band. Furthermore the fact that the intermediate amplifier is fixed means that it can be constructed to have a pronounced band-pass characteristic.

While most broadcast receivers operate from commercial alternating-current power sources a few receivers are designed for battery operation. Such sets are used in city areas where the power supply is direct current, in installations on automobiles, on farms, and where portability is important. They make use of circuits similar to those used in alternating-current sets and give approximately the same performance. The filament power is furnished by either dry cells or a storage battery while the plate, screen-grid, and grid-bias voltages are obtained from dry cells assembled in the familiar B and C battery units.

110. Miscellaneous Types of Broadcast Receivers.—The broadcast receivers described in the preceding section are the result of a long process

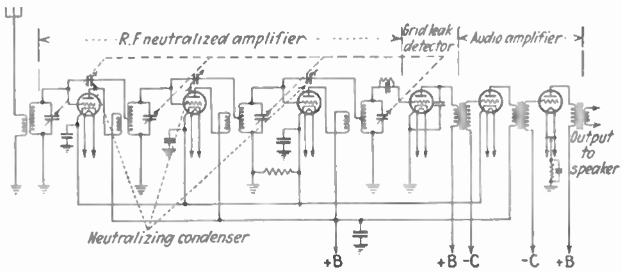


FIG. 254.—Circuit diagram of broadcast receiver using neutralized radio-frequency amplifiers.

of evolution and differ in many respects from radio receivers popular in the past. Before the development of the screen-grid tube most broadcast receivers consisted of a two- or three-stage tuned radio-frequency amplifier using general-purpose triodes, followed by a weak-signal detector of the grid-leak type, a stage of audio-frequency voltage amplification, and a power tube. The effects of the grid-plate capacity in the radio-frequency tubes were neutralized by one of the methods described in Sec. 44, of which the most widely used were the Neutrodyne (Hazeltine) system of Fig. 100*a* and *c*, and the Rice circuit of Fig. 100*b*. An example of such a receiver employing the Neutrodyne circuit is shown in Fig. 254.

Radio receivers employing neutralized radio-frequency amplification are in general inferior to those using screen-grid tubes. This is primarily because it is impossible to neutralize exactly the effects of the grid-plate tube capacity, with the result that extremely high over-all radio-frequency amplifications cannot be obtained without excessive trouble from regeneration. Even with only one or two stages there is some regeneration taking place, and since this is greatest at the higher broadcast frequencies the sensitivity of neutralized receivers tends to be greatest at the high-frequency end of the broadcast band. The use of screen-grid tubes eliminates substantially all energy transfer through the tube capacity,

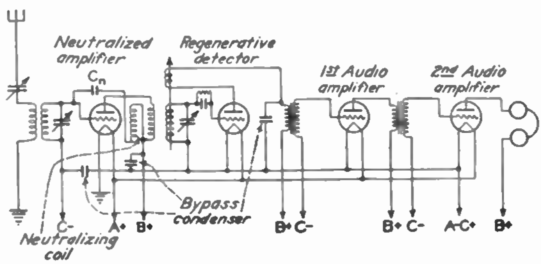


FIG. 255.—Broadcast receiver using one stage of neutralized radio-frequency amplification, followed by a regenerative detector. Such receivers were once very popular with home-set builders.

simplifies the entire construction, and makes it possible to use a greater amplification in one receiver than could possibly be handled with a neutralized amplifier.

A type of receiver that was once particularly popular with people who enjoyed building their own sets is shown in Fig. 254. This consists of one stage of neutralized radio-frequency amplification, a regenerative weak-signal detector, an audio-frequency voltage amplifier, and a power tube.

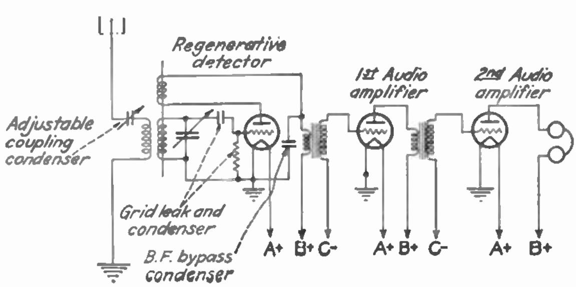


FIG. 256.—Regenerative receiver of the type used in the early days of broadcasting.

The very earliest sets used to receive broadcast signals employed no radio-frequency amplification, and consisted of a regenerative weak-signal detector followed by one or two stages of voltage amplification which operated a telephone head set directly without the use of a power amplifier. An example of such a receiver is shown in Fig. 256.

Regeneration such as employed in the sets of Figs. 255 and 256 represents a simple and inexpensive means of obtaining radio-frequency amplification but has a number of very serious drawbacks. In the first place considerable skill must be used in the operation of the regeneration control in order to obtain satisfactory results. In the second place regeneration reduces the effective resistance of the tuned circuit, as explained in Sec. 66, and this increases the sharpness of resonance and hence tends to discriminate against the higher side-band components of the signal, particularly at the lower broadcast frequencies. Finally it is impossible to operate a regenerative receiver without sometimes inadvertently increasing the regeneration until oscillations are produced. These oscillations cause the annoying squeals so familiar in the radio receivers of some years ago, and unless there is a buffer tube between the regenerative detector and the antenna, the oscillating radio receiver will act as a miniature transmitter that will interfere with other receivers in the neighborhood. With the development of satisfactory radio-frequency amplifiers, regenerative amplification has become unnecessary and is no longer relied upon in good receivers to increase the sensitivity.

In addition to the receivers of Figs. 254, 255, and 256 many other arrangements have enjoyed a certain amount of popularity at one time or another. These include receivers which use untuned radio-frequency amplifiers and in which all the tuned circuits are located between the antenna and the first radio-frequency amplifier tube, receivers in which the same tube is used simultaneously for radio and audio amplification (reflex amplifiers), etc., as well as innumerable variations of more common practice.

111. Cross-talk.—Cross-talk is the name given to interference produced between radio signals, and can be divided into two principal types. The first kind of cross-talk is produced by heterodyne detection of two signals having a frequency difference lying within the tuning range of the receiver. For example when one broadcast station is operating on 1400 kc and another on 600 kc, heterodyne detection of the two carrier waves will result in the production of an 800-kc difference frequency, which will be heard when the receiver is tuned at or near this difference frequency. The production of such cross-talk requires that the two interfering signals reach the grid of the first radio-frequency tube and that the characteristic curve of this tube be curved at the operating point. The magnitude of the cross-talk output can be determined by considering the first radio-frequency tube to be a weak-signal detector of the anode type. The results of such an analysis show that the magnitude of the cross-talk is proportional to the product of the interfering signal voltages that reach the grid, and increases with the curvature of the tube characteristic. Heterodyne cross-talk is therefore most prominent when the interfering signals are produced by powerful local stations, and when the volume control places the operating point of the radio-frequency tubes very close to cut-off. This type of cross-talk can be suppressed almost completely by using a tuned input circuit between the grid of the first tube and the antenna, since such a circuit will prevent at least one of the interfering signals from reaching the grid of the first tube.

The second kind of cross-talk is heard under the following circumstances: The receiver is tuned to a powerful local station—the "desired" signal—which is so strong as to require a low setting on the volume control. At the same time there is another powerful local station—the "unwanted" signal—operating on a frequency not greatly different from that of the station being received. During the interval in which the desired station is sending out an unmodulated carrier wave the modulation of the unwanted signal will be heard, but if the desired station ceases to radiate its carrier wave the interfering signals from the unwanted station disappear. Such cross-talk is caused by the unwanted signal modulating the carrier wave of the desired signal, and is much more troublesome than the first kind of cross-talk, because it can occur when the frequencies of the unwanted and desired signals are only slightly different.

Cross-talk Factor.—The magnitude of the cross-talk of the second type which is produced in a given tube can be conveniently measured in terms of the percentage of modulation which the unwanted signal produces on the carrier wave of the station being received, divided by the percentage of modulation of the unwanted signal. This is known as the cross-talk factor, which to a first-order approximation is given by he following equation:[1]

$$\text{Cross-talk factor} = \frac{\text{modulation produced on desired signal}}{\text{modulation of unwanted signal}}$$
$$= \frac{E_2^2}{2G_m}\frac{\partial^3 i_p}{\partial e_g^3} \qquad (151)$$

where

E_2 = amplitude of unwanted carrier

$G_m = \partial i_p/\partial e_g$ = mutual conductance (transconductance) of tube.

It is apparent from this equation that the cross-talk factor is independent of the amplitude of the carrier wave of the desired signal, but is proportional to the third derivative (*i.e.*, the rate of change of curvature) of the characteristic curve of the tube, and to the square of the voltage which the unwanted carrier develops at the grid of the tube. Cross-talk is therefore most prominent when the volume control is set to give low volume (*i.e.*, when a powerful local station is being received) and when the unwanted signal is produced by a powerful local transmitter operating on a frequency not too different from that to which the receiver is tuned. This is because a low setting on the volume control causes the tubes to

[1] See Stuart Ballantine and H. A. Snow, Reduction of Distortion and Cross-talk in Radio Receivers by Means of Variable-mu Tetrodes, *Proc. I.R.E.*, vol. 18, p. 2102, December, 1930.

be operated near the cut-off point where the characteristic curve has a high third derivative, and because an interfering signal such as described is capable of developing an appreciable voltage at the grid of the first radio-frequency tube.

Methods of Minimizing Cross-talk.—One way of preventing this type of cross-talk interference is to place selective circuits between the antenna and the input of the first tube and to limit the range of the volume control. The selective circuits reduce the amplitude of the unwanted carrier that reaches the grid of the first tube (*i.e.*, reduce the factor E_2 in Eq. (151)), while limiting the minimum volume avoids operating the tubes near the cut-off point where the third derivative of the characteristic curve is large. In order to avoid severe cross-talk when ordinary screen-grid or triode tubes are used in the radio-frequency amplifier it is necessary to use rather elaborate selective input circuits, such as several tuned circuits, or means must be provided for reducing the signal voltage which the antenna delivers to the radio receiver when local stations are being received. This latter arrangement is sometimes realized in a "local-distance" switch which can be used to disconnect the antenna when local signals are being received, and in other cases is provided by a compound volume control which operates on the antenna input to the receiver as well as on the radio-frequency amplification.

The most satisfactory method of minimizing cross-talk interference is by using tubes of the variable-mu type in the radio-frequency amplifier. As explained in Sec. 82 such tubes have a very gradual cut-off, which gives their characteristic a low curvature (*i.e.*, a low third-order derivative) when the mutual conductance of the tube is extremely low, which minimizes cross-talk at low volumes. Variable-mu tubes make a local-distance switch unnecessary and do not require more than a single tuned circuit between the antenna and the first tube.

Modulation Rise.—When the volume control places the operating point of ordinary triodes and screen-grid tubes close to cut-off, the curvature of the tube characteristic is such that the amplification depends to a certain extent upon the amplitude of the signal. When a modulated wave is received under these conditions the result is a distortion because the amplitude of the signal varies during the modulation cycle. The principal effect of this distortion is to cause the amplified signal delivered to the detector to be modulated a greater degree than is the original signal induced in the antenna. This is a result of the fact that the peaks of the modulation cycle are amplified more than the troughs. The magnitude of this effect in a tube is given to a first approximation by the following formula:[1]

$$\frac{\text{Modulation of amplified output}}{\text{Modulation of original signal}} = 1 + \frac{E_o^2}{4G_m}\frac{\partial^2 G_m}{\partial E_c^2} \tag{152}$$

[1] See Ballantine and Snow, *loc. cit.*

where E_o is the amplitude of the carrier wave applied to the tube, E_c the grid-bias voltage, and G_m is the mutual conductance (trans-conductance) of the tube. It is apparent from an examination of Eq. (152) that the rise in modulation depends upon the third-order derivative of the tube characteristic and upon the square of the applied carrier voltage. In order to avoid the distortion represented by modulation rise, the tubes must be operated so that the third-order curvature of the characteristic is not too great. This means either that variable-mu tubes must be employed, or if ordinary tubes are used, the volume control must not bring the operating point close to cut-off.

112. Miscellaneous Features of Broadcast Receivers. *Volume Control.*—All modern broadcast receivers are provided with some means for controlling the volume of the loud-speaker output produced by a given signal. A volume control is necessary to prevent overloading the tubes on powerful near-by stations and also to permit the volume of the reproduced signal to be fixed independently of the field strength of the transmitting station. Satisfactory results require that the volume be controllable over such a wide range that powerful local stations may be reduced to virtual inaudibility, and that the selectivity, fidelity, distortion, and frequency to which the set is tuned be independent of the volume control adjustment.

In modern receivers the volume is nearly always controlled by varying either the grid-bias or the screen-grid voltage of the radio-frequency amplifier stages. These arrangements take advantage of the fact that increasing the negative grid bias or decreasing the screen-grid potential reduces the mutual conductance of the screen-grid tube and hence lowers the amplification. The volume control in some cases operates on all the screen-grid tubes between the antenna and final detector, while in other cases it is not applied to the final radio-frequency tube. In superheterodyne receivers the control can also be applied to the first detector. In order to minimize troubles from cross-talk there is an advantage in arranging the circuit so that the several tubes involved contribute unequally to the control of the volume.[1]

In receivers employing variable-mu tubes in the radio-frequency amplifier the volume is controlled by varying the grid bias. This is accomplished in practically all receivers by biasing the cathode positive with respect to ground by means of a variable voltage derived from the power-supply system. This method of biasing is illustrated in serveral forms by Figs. 243, 245, and 251. When the receiver output is controlled by varying the screen-grid voltage it is customary to use a potentiometer.

[1] For a discussion of the advantages which result from arranging the volume control so that the different tubes do not approach cut-off simultaneously, see Stuart Ballantine and H. A. Snow, *loc. cit.*

In radio receivers which do not employ variable-mu tubes it is common practice to include in the volume control some means for varying the radio-frequency voltage which is applied to the grid of the first tube. This may consist of a device for disconnecting the antenna, in which case it is termed a local-distance switch; or it may include an arrangement which operates in conjunction with a control on the grids of the tubes and reduces the antenna input at low volumes. Such a compound-volume control is used in the receiver of Fig. 243. The object of these arrangements is to make it unnecessary to operate the radio-frequency amplifier tubes very close to the cut-off point when receiving powerful local signals, and thus to avoid cross-talk interference.

In some receiving sets the volume is automatically kept at a level that is substantially independent of the strength of the signal being received. This is accomplished by rectifying a portion of the carrier wave applied to the grid of the final detector and passing this rectified current through a resistance which is arranged so that an increase in the rectified current increases the negative grid bias of the volume-control tubes. Thus an increase in the signal strength gives more rectified current and reduces the amplification, while a reduction in the carrier voltage delivered to the final detector lowers the rectified current and decreases the negative bias on the volume-control tubes, thereby raising the receiver amplification. The exact details of the circuit arrangements for carrying out these operations can be varied considerably. A typical example is shown in the receiver of Fig. 259.[1]

Receivers equipped with an automatic volume control will maintain the carrier voltage applied to the grid of the final detector virtually constant over extremely wide ranges of signal strength. This has the very great advantage of maintaining the receiver output at the desired level during ordinary fading cycles and furthermore prevents the disagreeable blasts which would otherwise occur when a receiver adjusted to full sensitivity is accidentally tuned through a powerful local signal while searching for a much weaker signal. At the same time the automatic volume control has the disadvantage of causing the background noises to vary when the signal being received is fading badly. During the fading out interval the amplification increases to the point where the background noises become very pronounced, while during the "in" part of the fading cycle the background noises will disappear. The presence of the automatic volume control complicates the tuning of the receiver since the receiver tends to deliver the normal output even when slightly

[1] A general discussion of automatic volume-control systems together with several different circuit arrangements is to be found in the paper H. A. Wheeler, Automatic Volume Control for Radio Receiving Sets, *Proc. I.R.E.*, vol. 16, p. 30, January, 1928. Also see Dorman D. Israel, Sensitivity Controls—Manual and Automatic, *Proc. I.R.E.*, vol. 20, p. 461, March, 1932.

detuned. In order to overcome this disadvantage it is common practice to provide a meter in the plate circuits of the volume-control tubes to indicate the adjustment at which these tubes draw minimum plate current.

Receiving sets provided with automatic volume control always include some means of manually adjusting the level at which the output is maintained constant. This can be done by adjusting the resistance through which the rectified current flows in controlling the volume, by the use of an auxiliary bias adjustment operating on the same tubes as the automatic control (see Fig. 259) or by varying the volume of the audio-frequency amplifier system by one of the means discussed in Sec. 41.

Tone Control.—Many broadcast receivers are provided with a so-called "tone control," for the purpose of discriminating against the higher pitched sounds contained in the signal being received. The tone control ordinarily consists of an adjustable resistance-condenser combination (sometimes an inductance is also included) which is associated with the audio-frequency amplifier. Typical tone controls are shown in the receivers of Figs. 243 and 245. It is an open question whether there is any justification for the use of a tone control, which is fundamentally a device for making the reproduced speech or music unlike that originally created at the broadcasting station, but aggressive merchandising has led the public to expect this feature, which is therefore usually provided.

Alternating-current Hum.—A certain amount of low-pitched alternating-current hum is frequently observed in the output of alternating-current operated radio receivers. This hum may be caused by inadequate filtering of the plate power supply, by alternating current flowing in the field windings of the loud-speaker, by the alternating current flowing in the filaments or heaters of tubes, by alternating-current voltages introduced into the circuits by induction, etc. The presence of appreciable alternating-current hum indicates that there is either something wrong with the receiver or that the design is faulty. In order to keep the hum low it is necessary to provide an adequate filter that will deliver a substantially direct-current voltage to the grid, plate, screen-grid, and speaker field. Hum resulting from alternating-current heating of the tube cathodes is minimized by using heater-type tubes throughout except in the power stage and by biasing the heaters positively with respect to the cathode, as explained in Sec. 90. The power tubes should be specially designed to give low alternating-current hum, as explained in Sec. 90, and the grid and plate returns must be brought to the mid-potential of the filament. Alternating-current hum from induction is avoided by enclosing the power-supply transformer in an iron case and by locating it properly with respect to the radio set. The filament leads should furthermore be cabled in order that the magnetic field may be a minimum.

The tendency for alternating-current hum to be introduced into the radio-frequency amplifier depends upon the same factors that give rise to cross-talk. This is because hum is produced in radio-frequency amplifiers by modulation of the carrier wave, and the same conditions that produce cross-talk are also favorable for the production of modulation. It is therefore to be expected that means for avoiding cross-talk, such as variable-mu tubes, will decrease that part of the alternating-current hum which appears when a carrier wave is tuned in, but which disappears when the carrier is removed.

Noise Level.—When a radio receiver is adjusted for high sensitivity there is usually a background of noise in the form of crackles, hissing, etc., which is particularly noticeable during the silent intervals when the carrier wave being received is not modulated. This noise can be divided into two types, namely, that which is the result of interfering radio waves that are received by the antenna, and that which arises in the radio receiver and is present even when the antenna is disconnected.

That portion of the noise which is picked up by the antenna consists in part of static, that is, of radio waves produced by natural causes such as storms, atmospheric disturbances, etc. The remainder of the noise received by way of the antenna arises from man-made disturbances which produce radio waves or which induce radio-frequency voltages directly into the antenna by induction. This type of noise arises in many ways, such as by electrical discharges over faulty insulators in a power line, sparking of commutators and contacts, the ignition systems of automobile and airplane motors, x-ray equipment, opening and closing of switches in electrical circuits, etc., and is particularly strong in thickly populated areas where it is the factor that limits the useful sensitivity of the receiving set.

Inasmuch as the noise produced by natural and man-made causes is in the form of radio-frequency voltages in which the energy is distributed more or less uniformly throughout the frequency spectrum, there is no way by which the receiver can eliminate this noise except by the use of directional antennas and by maintaining the response band as narrow as is consistent with the side-band width to be received. In urban areas it is possible to keep the noise level down to a certain extent by giving attention to suppressing radiation from the worst sources of interference, but even when all reasonable precautions have been taken it is still found that the noise level in built-up districts is usually much higher than in less thickly populated regions where the amount of electrical equipment in operation is small.

The operation of receivers in close proximity to internal-combustion engines, as is necessary in aircraft and automobile radios, presents special problems. Under such conditions it is absolutely necessary that the ignition system be carefully shielded to prevent disturbances reaching the antenna and receiver.[1]

The noise that appears in the receiver output when the antenna is disconnected is the result of thermal agitation of the electrons in the input circuit, "shot effect" in the plate circuit of the input tube, etc. as discussed in Sec. 45, and which produce an effect that is lumped together as "fluctuation noise." The fluctuation noise is uniformly distributed throughout the frequency spectrum and sets an absolute limit to the sensitivity that can be usefully employed in a radio receiver. The discussion of Sec. 45 shows that this noise will be a minimum when the tubes are operated with sufficient electron emission to give a full space charge, when the gas pressure in the tubes is kept low (preferably less than 10^{-4} mm), and when the width of the band to which the receiver responds is as narrow as the required fidelity will allow. An investigation of the fluctuation noise developed in actual receivers when operated with sufficient amplification to develop a standard output from a given signal voltage has confirmed these conclusions and indicates that the selectivity of the radio receiver is the dominating factor in determining the fluctuation noise under normal conditions.[2]

The noise level in a radio receiver is usually increased by the presence of an unmodulated carrier, and in fact such a carrier wave often appears to be accompanied by a hiss which rises in intensity as the receiver is brought into resonance with the unmodulated carrier. This behavior results from heterodyne action between the relatively weak noise voltages normally present in the set and the relatively strong unmodulated carrier wave. The loudness of the noise depends to a considerable extent on the amplitude of the carrier and is relatively small when the carrier is absent because of the decrease in detector efficiency which takes place when the removal of the relatively strong carrier causes the detector to change over from a power to a weak-signal rectifier.

Microphonic Effects in Radio Receivers.—In some radio receivers any slight shock, such as a gentle tap, produces a ringing sound in the loudspeaker as a result of vibrations set up in the tubes (see Sec. 45). Such microphonic effects were very common in early receivers because of the large amount of audio-frequency amplification employed, but with the use of power detectors and with improvements in tubes have ceased to be of much importance. Microphonic action can also take place in a radio receiver as a result of the vibrations set up in the tubes, tuning

[1] A discussion of the problems of ignition shielding is given in the papers: H. Diamond and F. G. Gardner, Engine-ignition Shielding for Radio Reception in Aircraft, *Proc. I.R.E.*, vol. 18, p. 840, May, 1930; Paul O. Farnham, A Broadcast Receiver for Use in Automobiles, *Proc. I.R.E.*, vol. 18, p. 321, February, 1930.

[2] See Nelson P. Case, Receiver Design for Minimum Fluctuation Noise, *Proc. I.R.E.*, vol. 19, p. 963, June, 1931.

condensers, chassis, etc., by the sound waves originating in the loud-speaker, and cause some audio frequencies to be reinforced while others are discriminated against. Distortion of this sort is eliminated by using rigid construction, by taking care in locating the loud-speaker with respect to the chassis, and by mounting the chassis so that vibrations of the receiver cabinet will not be transmitted to it.

113. Design of Radio Receivers.—It is impossible to make an exact and complete mathematical design of a broadcast receiver because such factors as coil losses, regeneration, and stray couplings prevent the radio-frequency amplification from being accurately predictable. As a result, the design of the circuits for a broadcast receiver is to a considerable extent a matter of judgment guided by such calculations as can be made. The first step in the design of a broadcast receiver is to provide a power amplifier capable of delivering the maximum undistorted power output which it is desired to obtain from the receiver. This is done by reference to a table of tube characteristics, which will give the required tube type, plate voltage, and grid-bias potential. The next step is to estimate the radio-frequency carrier voltage that when applied to the detector and completely modulated will deliver sufficient audio-frequency output to excite the power tubes to their full power capacity. This is done by the principles outlined in Chap. VIII, where it was shown that the audio-frequency output voltage of a power detector is given by the formula:

$$\text{Audio-frequency output voltage} = DmE_oA \tag{153}$$

where A is the amplification of the detector considered as an audio-frequency voltage amplifier, D is the detector efficiency, m is the degree of modulation, and E_o the carrier amplitude. By making an estimate of the probable detector efficiency it is possible to calculate with a fair degree of accuracy the carrier amplitude which when fully modulated will excite the power amplifier either directly or through a stage of voltage amplification. Detector grid-bias and plate voltages are then selected to accommodate a signal of twice this amplitude in order to allow for the detection of completely modulated waves.

Calculation of Required Radio-frequency Amplification.—With the approximate radio-frequency carrier voltage which must be delivered to the detector determined in this way, it is now possible to estimate the amount of radio-frequency amplification that will be necessary to give a desired sensitivity. In receivers of the tuned radio-frequency type the amount of amplification that can be satisfactorily handled is 15 to 30 per stage. It would be easy to obtain greater amplification, but the gain per stage is purposely kept low to avoid troubles from regeneration, and to permit the use of small, easily shielded coils. Such coils have high resistance (low Q) which makes their resonance curve broad enough to avoid side-band trimming at the lower broadcast frequencies and also makes it easier to align the circuits for single-dial control. The actual amplification obtained depends upon the regeneration in the amplifier and so cannot be predicted accurately, although estimates based on previous experience will give the right order of magnitude. Since the amplification tends to vary with the frequency to which the receiver is adjusted, the usual design procedure is to provide the required gain under the most unfavorable circumstances and then to compensate for the surplus amplification at other frequencies by some method of equalization. In determining the radio-frequency amplification that is required, it is necessary to take into account the voltage step-up in the selective circuits that are located between the antenna and the grid of the first tube.

In the superheterodyne type of receiver the total radio-frequency amplification is the product of the step-up of the antenna tuning circuits, the radio-frequency amplification before the first detector, the intermediate-frequency amplification, and the amplification of the first detector. The gain of the first detector is equal to the amplification which this tube would give when acting as an intermediate-frequency amplifier, multiplied by the efficiency of detection. The efficiency of detection depends on the amplitude of the local oscillator, as explained in Sec. 65, and will ordinarily be 50 per cent or higher.

An example showing how these design principles are applied as a guide in the design of an ordinary receiver is furnished by the following case. Suppose it is desired to obtain a maximum of 3 watts output in a receiver which is to have a sensitivity of 5μv. Reference to Table VI shows that this output can be obtained by using two Type '45 tubes in push pull at a plate voltage of 250 and a grid-bias potential of 50 (alternating-current operation). The signal voltage required to develop this output is in the order of 48.5 volts per tube crest value (the grid bias required for direct-current operation) and so is 97 volts for the two tubes. Assuming the power detector has a voltage amplification of 20 at audio frequencies and is to excite the power tubes directly, then reference to Eq. (153) shows that if the detector efficiency is estimated at 50 per cent, the maximum carrier voltage that the detector will require is $97/(20 \times 0.5) = 9.7$ volts crest value. When completely modulated, the crest radio-frequency voltage applied to the grid will be twice this, or 19.4 volts, which is also the required detector grid bias for anode rectification. The detector plate voltage must be approximately this grid bias multiplied by the amplification factor of the tube. The required radio-frequency amplification is now determined by calculating the carrier voltage which when modulated 30 per cent and applied to the detector grid will develop 0.05 watt in the output of the power amplifier. Since with linear detection the power output is proportional to the square of the degree of modulation and the square of the carrier voltage, the signal required at the detector to develop the standard output is $(0.05/3.00)^{1/2}$ $9.7/0.30 = 4.16$ volts. In order that a carrier voltage of 5μv effective (7.07μv crest) induced in the antenna may develop 4.16 volts at the detector grid, the over-all radio-frequency amplification from the induced voltage in the antenna to the detector must be $(4.16/7.07) \times 10^6 =$ 609,000. In a tuned radio-frequency type of receiver this could be obtained by using three stages of radio-frequency amplification with a tuned circuit between the antenna and the first tube. Assuming that the amplification averages 30 per stage and that the step-up in the antenna tuning coil is 25, the over-all amplification is $25 \times 30 \times 30 \times 30 =$ 675,000, which would be ample. In a superheterodyne type of receiver the required amplification could be obtained by two stages of intermediate-frequency amplification, one stage of radio-frequency amplification, and an antenna tuning circuit. Thus assuming that the efficiency of the first detector is 50 per cent, that the gain per stage of intermediate-frequency amplification is 15, and that the antenna step-up and amplification of the input amplifier are 20 and 25, respectively, this gives an over-all amplification of $20 \times 25 \times (15 \times 0.5) \times 15 \times 15 = 760{,}000$.

With the general layout determined in the manner outlined it becomes possible to plan the more important details of the power-supply system. This will be called upon to deliver a direct-current voltage corresponding to that required by the power amplifier and must have a current capacity equal to the sum of the plate currents of all the tubes plus the sum of the screen-grid currents. In some cases a small additional current is also required to supply grid-bias voltages. The standard rectifier-filter arrangement ordinarily employed in receivers is the full-wave single-phase center-tapped transformer circuit operating into a two-section filter having a condenser input. The grid-bias voltages can be obtained by resistances placed between cathodes and ground, by resistances in the plate-return leads, or by biasing the cathodes positively with respect to ground by means of current derived from the power-supply system. The amount of bias required for the power tubes and detector is fixed by factors already discussed, while the bias for the radio-frequency amplifier tubes need not be large because these tubes are never called upon to handle inputs in excess of several volts. With screen-grid heater tubes it is customary to employ about 3 volts bias, which allows 1 volt for the electron emission velocity and provides for a 2-volt signal. The exact details of the arrangements for supplying grid-bias and anode voltages can be varied in many respects, as is made apparent by a comparison of the arrangements used in Figs. 243, 245, and 251.

The broad outline of the receiver as obtained above can now be used as a guide in carrying out the actual design details, and it is here that the skill and experience of the designer are of greatest value. He is faced with many alternatives in power-supply systems, grid-bias methods, volume-control arrangements, and is furthermore always faced with the problem of striking the proper balance between the conflicting requirements of low cost and compactness on the one hand, and high quality of performance on the other. In order to obtain the most satisfactory design it is necessary to build experimental receivers incorporating the details which have been laid out on paper and then to make such further modifications in the design as tests upon these receivers indicate are desirable.

114. Receivers for the Reception of Telephone Signals of Other than Broadcast Frequencies.—Receivers of this class are essentially broadcast receivers modified as required because of the difference in the frequency to be received. Receivers of the tuned radio-frequency

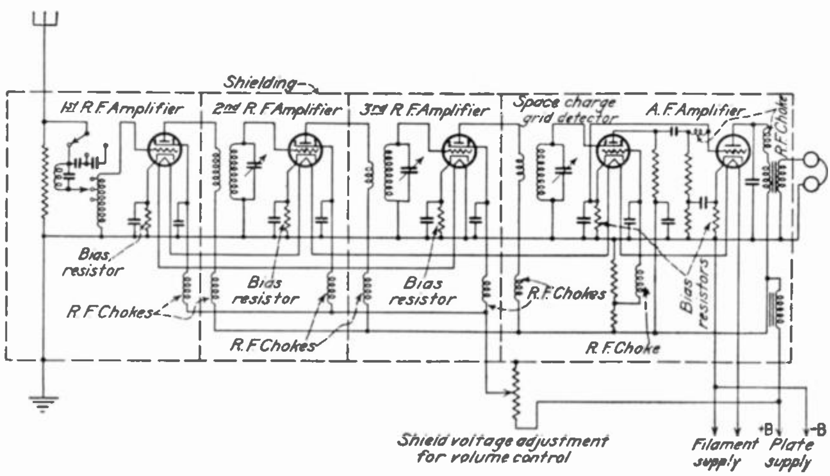


FIG. 257.—Circuit diagram of Western Electric airplane receiver for the frequency range 1500 to 6000 kc.

type for other than broadcast frequencies are thus essentially the corresponding broadcast receivers, the coils and condensers of which have been modified to give the desired tuning range. An example of such a tuned radio-frequency receiver for the reception of short-wave telephone signals is the Western Electric airplane receiver of Fig. 257. A conventional three-stage tuned radio-frequency amplifier using screen-grid tubes is employed, followed by a space-charge grid power detector of the bias type, and a power amplifier. The coils and condensers are carefully matched, and the tuning condensers are mounted on a common shaft to give single-dial control. A photograph of a complete receiver is shown in Fig. 258. The Western Electric Company also manufactures a long-wave airplane receiver, which, as far as circuit arrangements are concerned, is absolutely identical with the short-wave set just described.

Receivers of the tuned radio-frequency type have a number of disadvantages when used on frequencies above 2000 kc. The amount of amplification which can be obtained at these frequencies without instability from regeneration through the residual electrode capacity is low even when screen-grid tubes are employed. The selectivity is also rather poor because the Q of the tuned circuits tends to decrease at the higher frequencies, whereas the Q must be proportional to the frequency if constant selectivity is to be maintained. It is furthermore difficult to line up a number of tuned circuits operating from a common control

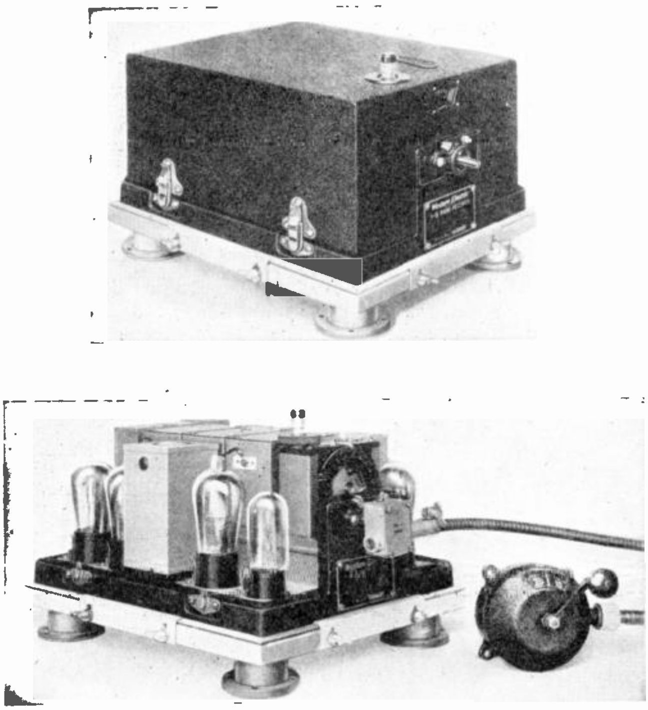


Fig. 258.—Photographs of Western Electric airplane receiver, showing constructional details.

when these circuits use the very small capacities that are employed in tuning high-frequency circuits. For these reasons the tuned radio-frequency type of amplifier has only limited usefulness in the frequency range from 1500 to 6000 kc, and is practically never employed at higher frequencies. Where the highest type of performance is desired it is always preferable to use a superheterodyne receiver.

The circuit diagram of a typical short-wave superheterodyne receiver used in commercial ship-to-shore radio-telephone extensions of the land telephone system is shown in Fig. 259. In this receiver a band-pass filter using four tuned circuits is placed between the antenna and the grid of the first detector tube to prevent interference from image signals and to minimize cross-talk. The first detector is of the grid-bias type and employs a screen-grid tube with the local oscillations introduced in the plate circuit. A three-stage 300-kc intermediate-frequency amplifier of the band-pass type is used, followed by a power detector. In order to prevent variations in the receiver output during fading periods an automatic volume control is provided. This consists of two additional stages of intermediate-frequency amplification, followed by a rectifier.

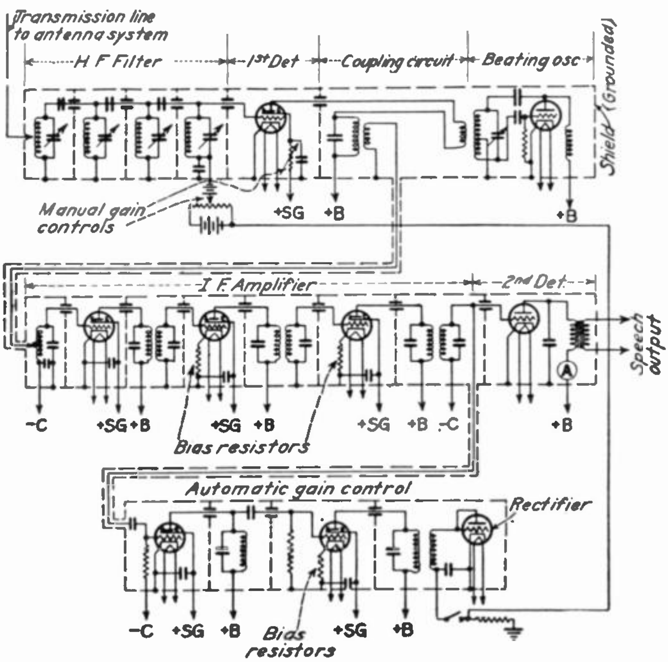


FIG. 259.—Circuit diagram of short-wave superheterodyne receiver used in commercial radio telephony.

The direct current in the rectifier output is passed through a resistance which is arranged in such a manner as to increase the negative bias on the grid of the first detector when the amplitude of the received signal increases, and thus automatically to reduce the amplification. Manual control of the volume to take into account wide variations in signal strength is provided for by a potentiometer on the screen grid of the first detector and by additional grid-bias control of this same tube.

In short-wave superheterodyne receivers the intermediate frequency is a relatively small percentage of the carrier frequency being received, so that the frequency difference between the desired signal and its image signal is relatively small. This makes it necessary to use very efficient pre-selector circuits (illustrated in Fig. 259) in front of the first detector, if interference from image signals is to be cut down to a minimum. If effective pre-selector circuits are not employed it is necessary to take considerable care in coupling the local oscillator to the first detector, since with close coupling the local oscillator will synchronize with any strong signal that forces its way to the first detector and happens to have a frequency approximating that of the local oscillator.

Short-wave radio-telephone signals can also be received on regenerative short-wave code receivers of the type shown in Fig. 261, by operating the regeneration control just below the oscillating point. It is possible in this way to obtain fairly satisfactory results, although the sensitivity and selectivity will be inferior to those obtained with a good superheterodyne receiver. Reception of long-wave telephone signals on regenerative circuits of this sort does not give satisfactory results, however, because of the side-band trimming that takes place when the received frequency is low and the circuit resistances are made very small by regeneration.

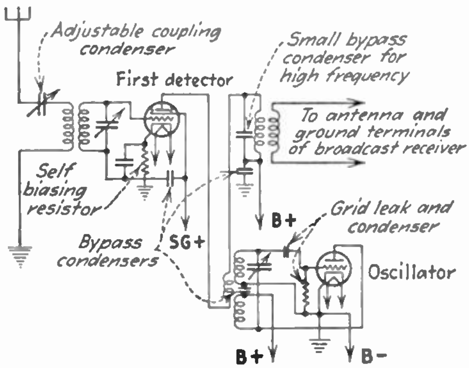


FIG. 260.—Circuit diagram of short-wave adapter for a broadcast receiver.

Any broadcast receiver can be readily transformed into a short-wave superheterodyne receiver by using heterodyne action to convert the short-wave signal to a broadcast frequency. Arrangements of this sort are known as short-wave adapters or converters and ordinarily consist of a heterodyne oscillator and a first detector, the output of which is delivered to the broadcast receiver which acts as an intermediate-frequency amplifier. An example of such a short-wave adapter is shown in Fig. 260.

In radio-telephone systems employing the single side-band method of transmitting it is necessary to supply the missing carrier frequency at the receiver by means of a local oscillator. Thus in the long-wave transatlantic telephone system discussed in Sec. 105 the receiver is of the superheterodyne type operating at an intermediate frequency in the order of 30 kc. The carrier is supplied by heterodyning the output of the intermediate-frequency amplifier with a local oscillator of suitable frequency and then rectifying the resultant heterodyne signal by a second detector to reproduce the original speech.

115. Receivers for Telegraph Signals.—Receivers for handling telegraph signals are essentially broadcast receivers to which there has been added some means for interrupting the code characters at an audible rate to permit reception with a telephone receiver. While there are a number of ways by which this breaking up of the dots and dashes can be accomplished, the method universally employed in practice consists in combining with the signal a local oscillation differing from the signal frequency by an audible amount, such as 1000 cycles, and rectifying the resultant heterodyne signal as explained in Sec. 65.

One of the chief merits of the heterodyne method of reception is that the beat frequency which is produced depends upon the frequency of the signal being received. This makes it possible to distinguish between signals which differ so little in frequency that ordinary tuned circuits cannot separate them. For example, if it is desired to receive from a station transmitting on 20,000,000 cycles, while an interfering station is operating on a frequency of 20,000,200 cycles, it would be obviously impossible to eliminate the undesired signal by means of tuned circuits since the two carrier frequencies differ by only 0.001 per cent. However, heterodyning with a local oscillation having a frequency of 20,001,000 cycles produces beat notes with the desired and undesired signals of 1000 and 800 cycles, respectively, and these frequencies can be readily separated either by ear or by a selective circuit because they differ by 25 per cent.

The most widely used type of code receiver consists of a single stage of screen-grid radio-frequency amplification, an oscillating detector, and two stages of audio-frequency amplification. The circuit diagram of a commercial short-wave receiver of this character is shown in Fig. 261. Such receivers normally possess three controls: one for tuning the input circuit of the radio-frequency amplifier; one for tuning the output circuit of the radio-frequency amplifier, which is also the tuned circuit controlling the frequency of the oscillating detector; and one for controlling the regeneration and hence the amplitude of oscillations. Interchangeable plug-in coils can be used to enable the same receiver to cover a wide frequency range. Code receivers of this type, while relatively simple and involving only a small number of tubes, possess remarkable sensitivity because of the large regenerative amplification which is obtainable from an oscillating detector.

In order to obtain the best results with a heterodyne receiver using an oscillating detector it is necessary to pay some attention to the circuit conditions. Maximum regenerative amplification is always obtained when the amplitude of oscillations is as low as possible, *i.e.*, with the regeneration control set at the minimum value that will sustain oscillations. With this adjustment threshold or fringe howl frequently occurs, as explained in Sec. 67, where means of eliminating this trouble are also discussed. The particular type of regenerative circuit and the method used to control the amount of feed back that takes place are unimportant. The different circuit arrangements will nearly always give equivalent results and differ primarily in the extent to which the regeneration control must be readjusted to maintain maximum sensitivity as the receiver is tuned to different frequencies. The different methods of controlling regeneration also differ in the effect which the feed back has on the frequency of oscillations. In the receiver of Fig. 261 the feed back is accomplished by means of a plate inductance connected in a shunt-feed arrangement, and the amount of regeneration is controlled by a vari-

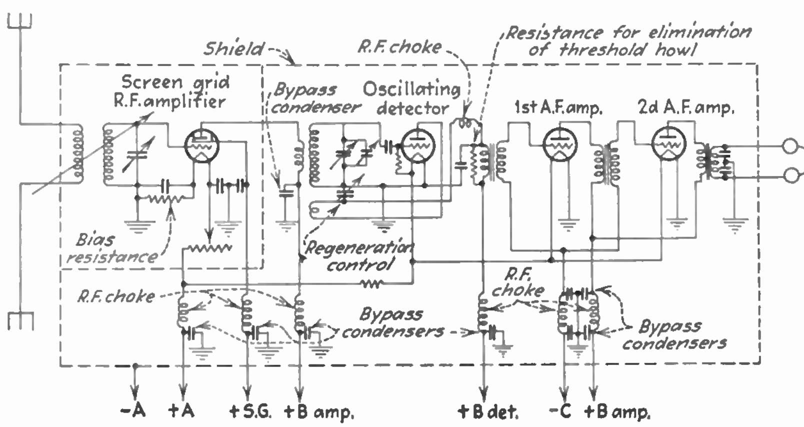


FIG. 261.—Typical regenerative type of code receiver using an oscillating detector.

able condenser in series with this inductance. Changing the condenser capacity alters the magnitude and phase of the current through the inductance, which affects the amount of energy transferred to the grid circuit and hence controls the tendency to oscillate. It is very important that the tuning coils have the lowest possible losses, since the signal voltage developed across the circuit in which the regeneration takes place is proportional to the circuit Q. It is also desirable that the ratio of inducance to capacity in the tuned circuits be high.

In order to obtain satisfactory heterodyne reception it is necessary that the transmitted frequency and the frequency of the local oscillator, or oscillators, be constant. Otherwise the pitch of the audible beat note will vary, and reception will be difficult. In particular the transmitter must be free of frequency shifts caused by keying. The importance of frequency stability increases as the frequency of transmission increases because the same percentage variation then represents a greater number of cycles. At extremely high frequencies, such as 100,000,000 cycles, it is virtually impossible to obtain a steady beat note because of the effect of minute vibrations, etc., on the frequency.

In receivers employing oscillating detectors the input radio-frequency amplifier serves primarily as a buffer tube that prevents the oscillations of the detector from being radiated by the antenna. The contribution which this buffer tube makes to the sensitivity is of secondary importance, particularly at the higher frequencies where the amount of extra amplification that can be obtained under even the most favorable conditions is relatively small. In some receivers, used principally by amateurs, the input radio-frequency amplifier is omitted, thus simplifying the tuning arrangement at the expense of receiver radiation and a slight reduction in sensitivity. Receivers used in commercial point-to-point communication sometimes employ an additional radio-frequency stage before the oscillating detector.

Receivers similar to that shown in Fig. 261 are very satisfactory for all except the very lowest radio frequencies. At these long wave lengths the input circuit of the oscillating detector must be detuned by such a large percentage of the carrier in order to give an audible beat note that a serious reduction of the regenerative amplification results. In the reception of low radio frequencies it is therefore preferable to use a separate heterodyne oscillator, and to operate the detector in a regenerative but non-oscillating condition.

When greater sensitivity is required than can be obtained from the general-purpose receiver of the type shown in Fig. 261 a superheterodyne receiver should be employed. Superheterodyne receivers for code reception differ from those designed for radio telephony only in that the code receiver requires a local oscillator to heterodyne with the output of the intermediate-frequency amplifier for the purpose of giving an audible beat note. This oscillator need not be adjusted during operation and can consist either of an oscillating second detector or of a separate fixed oscillator. The oscillations which are combined with the signal to produce the intermediate frequency are usually generated from a separate oscillator, but when receiving very high frequencies the percentage of detuning that is required to give beat notes in the order of several hundred thousand cycles is so small that it is possible, although not considered best practice, to use an oscillating first detector. While the superheterodyne code receiver represents the most sensitive method that has been devised for the reception of code signals, its advantages in code reception over the very much simpler oscillating detector receiver, such as that of Fig. 261, are small.

The above discussion applies only to continuous-wave code signals, *i.e.*, code signals in which the amplitude is constant throughout the dot and dash interval. Spark signals require receivers of the type used with telephone signals, since the spark transmitter sends out modulated waves. I.C.W. (interrupted continuous wave) signals can be received either in the same way as spark signals, or by heterodyning. The latter method is ordinarily employed where the signals are weak because of the great sensitivity of the oscillating detector, while the former arrangement is often used where the signals are strong since the audio pitch is then determined by the rate of interruption at the transmitter and is independent of receiver adjustments.

116. Receiving Systems for Minimizing Fading.—Fading represents one of the chief factors limiting the usefulness of short-wave radio signals. As is explained in Chap. XV fading is the result of interference between radio waves which reach the receiver along two or more paths of unequal length. Fading is most pronounced at the high frequencies, where a given change in path length represents a larger fraction of a wave length. It is of importance at broadcast frequencies only on night signals from distant stations, and is a negligible factor in the reception of low-frequency radio waves. When fading is present the signal intensity varies widely in a random manner. The rapidity of the fading depends on the conditions existing at the particular time and may be from once every few minutes or every few seconds to many times a second. Occasionally the fading is so rapid as to be at an audible rate, *i.e.*, several hundred times a second.

Automatic Volume Control.—The most obvious means for minimizing fading is to use an automatic volume control that will vary the amplification of the receiver in such a way as to maintain a substantially constant receiver output irrespective of variations in the strength of the received signals. This method is used in a number of commercial broadcast receivers and is also extensively employed in the reception of short-wave telephone signals, but it is not always successful because the signals sometimes fade out so completely as to be lost in the background noise, in which event no amount of amplification will be successful in retrieving them. Furthermore, as the signals fade in and out, the sensitivity of the receiver is varied, and there is a continual variation in the background noise.

Limiting.—Another means of minimizing fading is known as limiting and consists in using a very efficient antenna system that will abstract a large amount of energy from the waves, and then employing a receiver having a non-linear characteristic such that signal inputs above a critical level will give substantially constant output. By using a sensitive receiver the output during ordinary fading-"out" intervals can usually be kept up to the limiting level, and since the stronger signals of the fading-"in" period can do no more, the bad effects of fading are considerably smoothed out. This arrangement fails to function when the fading is so severe that the signal drops below the noise level, and furthermore cannot be employed in the reception of radio-telephone signals because the non-linear characteristic which is required introduces distortion.

Frequency Diversity.—Experiments have shown that frequencies differing by a few hundred cycles do not fade simultaneously, and this makes it possible to minimize fading troubles by transmitting the same message on different frequencies. Many short-wave code transmitters accordingly modulate the transmitter output at some audible frequency, say 500 cycles, and then key this modulated wave. The audio modulation introduces two side-band frequencies and in effect causes the telegraph message to be transmitted simultaneously on three slightly different frequencies which is of considerable help in reducing the number of drop-outs. The disadvantage is that the energy of the transmitter is distributed over a wider frequency range and so is less effective in cutting through interference and noise than would be the case with all the energy concentrated on one frequency.

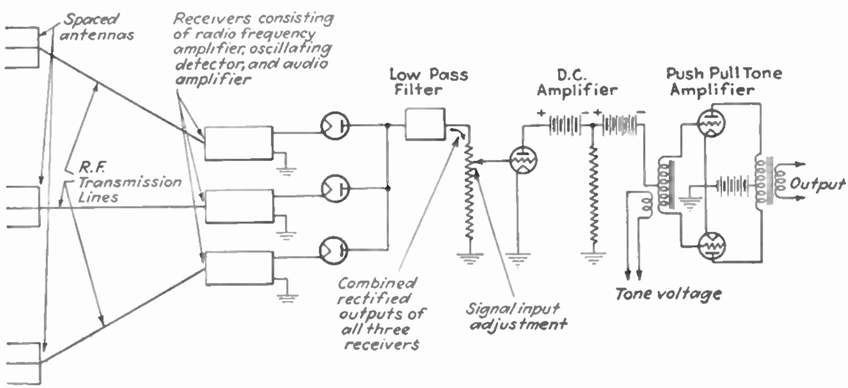


FIG. 262.—Diagram showing operation of diversity receiving system for telegraph signals.

Polarization Diversity.—Experiments have shown that the vertically and horizontally polarized components of a radio wave from the same transmitter do not fade together. This can be taken advantage of to minimize fading effects by using two receivers, one with a vertically polarized and the other with a horizontally polarized antenna, and combining the outputs. This arrangement is seldom used, however, because the spaced antenna system described below is more effective.

Spaced-antenna Diversity System.—The most successful means that has been found for minimizing fading effects makes use of the fact that the signals which are induced in antennas spaced several wave lengths or more apart do not fade simultaneously. When three or more such antennas are employed it is extremely unlikely that the signals will fade out completely on all of them at the same time, so that if the outputs of the different antennas are combined there will be nearly always some signal present.

An example of a commercial installation making use of a spaced-antenna receiving system for minimizing fading of code signals is shown schematically in Fig. 262.[1] This makes use of three short-wave Beverage antennas of the type discussed in Sec. 122, which are spaced approximately 10 wave lengths apart. Each of these antenna systems delivers its output through a radio-frequency transmission line to a separate receiver of the oscillating-detector type. The output of each audio-frequency amplifier is rectified, and the rectified outputs of the three receivers are combined by passing them through a common resistance. The voltage drop across this resistance is amplified by a one-stage direct-current amplifier, the output of which is passed through a resistance that supplies the grid-bias voltage for a push-pull amplifier. The input of this amplifier is excited from an audio-frequency "tone" oscillator. When no signals are being received the current through the resistance common to the rectified outputs of the three receivers is negligible, causing the plate current of the direct-current amplifier tube to be large. This places a rather high negative bias on the push-pull amplifiers, blocking them. If, however, a signal is received from at least one of the antenna systems, current flows through the common output resistance of the three receivers, making the grid of the direct-current amplifier more negative, reducing the amplifier plate current, and lowering the grid bias of the push-pull amplifier. This enables the push-pull amplifier to pass the audio frequency from the tone oscillator. The adjustment on the grid bias of the push-pull amplifier is normally such that when no signal is being received the noise level will just fail to allow the push-pull amplifier to pass tone current.

A diversity arrangement as described will eliminate virtually all drop-outs resulting from fading, since it only rarely happens that the signals will fade out completely on all three receiving systems at the same instant. This is well illustrated by the records of Fig. 263, which compare the combined output of the diversity receiving system with the individual outputs of each antenna. The arrangement by which the output of the receiving system keys a local oscillator gives a final output having an amplitude and frequency independent of the adjustments of the radio receiver. Furthermore by properly adjusting the bias which the direct-current amplifier places on the tone amplifier it is possible to eliminate virtually all background noise. The receiving system therefore has the simplicity and sensitivity of the oscillating detector, the

[1] An arrangement of this sort is known as a "diversity receiving system." For a detailed description of this particular installation see H. H. Beverage and H. O. Peterson, Diversity Receiving Systems of RCA Communications, Inc., for Radio Telegraphy, *Proc. I.R.E.*, vol. 19, p. 531, April, 1931.

selectivity that results from heterodyne detection, and at the same time minimizes fading and noise.

The diversity reception of radio-telephone signals is considerably more difficult than the diversity reception of code signals because the audio-frequency outputs of the separate receiving systems are not always in phase and so if added together may give a highly distorted resultant. Furthermore when the outputs of all receivers are combined directly each receiver continuously contributes its quota of noise. The most satisfactory solution for this situation makes use of the fact that the quality of the telephone signals received at any instant is usually best on the antenna which is delivering the largest signal voltage to its receiver. A considerable improvement in reception can hence be obtained by automatically switching from one receiving system to another in such a way as to be always connected to the receiver which is delivering the largest output.

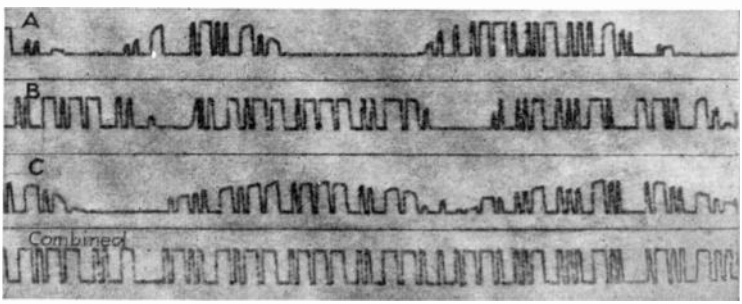


FIG. 263.—Typical records showing output of each receiver of a diversity receiving system and of the combined output. Note the entire absence of fading in the latter.

The impossibility from a practical point of view of using mechanical switches for transferring the output connections from one receiver to another has led to the arrangement shown schematically in Fig. 264.[1] The same three antenna systems are employed here as in the telegraph diversity code receiving system described above. The output of each antenna is delivered to a superheterodyne receiver in which the second detector is so operated as to have a square-law characteristic. The audio-frequency outputs of these three detectors are added directly and represent the final output of the diversity receiving system. The direct-current components of the rectified currents in the three second detectors are also added and are used to operate an automatic volume control that varies the amplification of all three receivers simultaneously.

[1] This is the diversity receiving system used by the RCA Communications, Inc., at their Riverhead Long Island Receiving Station. For a more detailed description see H. O. Peterson, H. H. Beverage, and J. B. Moore, Diversity Telephone Receiving System of RCA Communications, Inc., *Proc. I.R.E.*, vol. 19, p. 562, April, 1931.

This arrangement is virtually equivalent to an automatic switching arrangement because the square-law characteristic of the second detectors causes most of the output of the diversity receiving system to come from the antenna which receives the largest signal. Thus if the signals at the three antennas have relatively values of 10, 5, and 3, the voltage outputs of the three square-law detectors will be in the ratio of 100, 25, and 9, while the power outputs are in the ratio of 10,000, 625, and 81. Thus the square-law characteristic tends to suppress contributions from the weaker signals. Trouble from excessive noise is avoided by making the automatic volume control common to all receivers. When the output of any one receiver is strong the control reduces the amplification of all three receivers and hence keeps down the noise in the common output.

The individual receivers used in a diversity receiving system must be very carefully shielded from each other because otherwise cross-talk

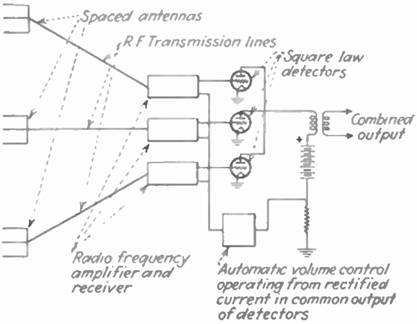


FIG. 264.—Schematic diagram of diversity telephone receiving systems. The outputs of the square-law detectors are combined directly, and an automatic volume control operates on all receivers.

will be introduced between the local oscillators. This shielding is accomplished by the generous use of metal boxes, by filtering all the power leads to the receiver, and by taking care to prevent local oscillator energy from reaching the antenna systems.

117. Reception of Very High Frequency Waves.—The problem of receiving waves having frequencies in excess of 30,000 kc (less than 10 meters in length) has not been satisfactorily solved. While it is possible to use an oscillating detector or superheterodyne receiver in the wavelength range from 1 to 10 meters, the difficulty of maintaining a constant beat note at these extremely high frequencies usually makes the results unsatisfactory. The usual experience is to find that the beat note is merely a ragged hiss because of very slight vibrations and other irregularities in the transmitting and receiving equipment. The most satisfactory receiver for this wave-length range is the superregenerative receiver, which, as explained in Sec. 68, is essentially an oscillating detector which is changed from an oscillating to a non-oscillating condition at a low radio-frequency rate, such as 20,000 times a second. An example of such a short-wave receiver is found in Fig. 265, which shows an ordinary oscillating detector with the plate voltage supplied from an auxiliary oscillator operating at a low frequency.

The ordinary methods of receiving signals fail completely on wave lengths shorter than about 1 meter, and receivers for these very short

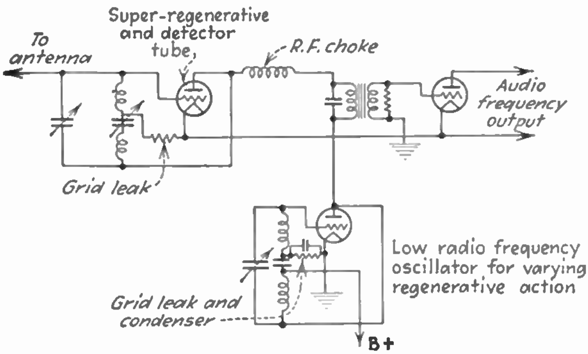


FIG. 265.—Circuit diagram for super-regenerative receiver operating in the range 1 to 10 meters.

waves are still in a very primitive state. Some of the experimenters in this field have made use of crystal rectifiers, such as a galena crystal upon which a pointed wire rests lightly, but such rectifiers are very insensitive. A very promising development is based on the discovery that when a radio signal is impressed between the plate and filament of a

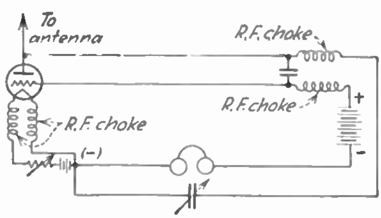


FIG. 266.—Ultra-short wave receiver using regenerative oscillating detector of the Barkhausen type.

Barkhausen oscillator that is operating at approximately the signal frequency it is possible to obtain a large regenerative amplification.[1] The diagram of such a receiver is shown in Fig. 266.

[1] For further information on such electron-oscillating detectors see Kinjiro Okabe, The Amplification and Detection of Ultra-short Electric Waves, *Proc. I.R.E.*, vol. 18, p. 1028, June, 1930; Shintaro Uda, Radiotelegraphy and Radiotelephony on Half-meter Waves, *Proc. I.R.E.*, vol. 18, p. 1047, June, 1930.

CHAPTER XIV

ANTENNAS

118. Fundamental Laws of Radiation.—An understanding of the mechanism by which energy is radiated from a circuit and the derivation of equations for expressing this radiation quantitatively involve conceptions which are unfamiliar to the ordinary engineer.[1] The mathematical formulas which express the results of such an analysis are however quite simple and understandable. Thus the strength of the field radiated from an elementary length of wire δl carrying a current I (see Fig. 267), is given by the formula:

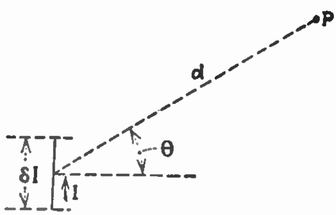


FIG. 267.—Elementary doublet consisting of a length of wire δl carrying a current I.

$$\epsilon = \frac{60\pi}{d\lambda}(\delta l)I \cos \omega\left(t - \frac{d}{c}\right) \cos \theta$$

$$= \frac{60\pi}{dc}f(\delta l)I \cos \omega\left(t - \frac{d}{c}\right) \cos \theta \qquad (154a)$$

where

ϵ = the strength of the wave in volts per meter

δl = the length of wire from which the radiation takes place, measured in the same units as λ

$I \cos (wt + 90°)$ = current flowing in the wire in amperes

d = distance from P to the antenna in meters

θ = angle of elevation of point at which field is desired with respect to a plane perpendicular to the conductor δl

f = frequency of current

$\omega = 2\pi f$

t = time

c = velocity of light = 3×10^8 meters per second

λ = wave length corresponding to frequency f.

The radiated field ϵ varies directly as the current I, the frequency f, the doublet length δl, and the cosine of the angle of elevation, and is inversely

[1] For an elementary mathematical analysis of the radiation phenomenon, see R. R. Ramsey and Robert Driesback, Radiation and Induction, *Proc. I.R.E.*, vol. 16, p. 1118, August, 1928. A more advanced treatment is given by G. W. Pierce, "Electric Oscillations and Electric Waves," McGraw-Hill Book Company, Inc., New York, 1920.

proportional to the distance d. The phase of the field depends on the phase of the current at the instant the wave left the antenna. The strength of the magnetic component H of the wave is related to the electrostatic voltage gradient ϵ by the equation

$$\epsilon = 300H \tag{154b}$$

where H is in lines per square centimeter, and ϵ is in volts per centimeter.

The total field radiated from an antenna is found by adding up the separate fields produced by the elementary lengths of the radiator, taking into account phase relations and planes of polarization in making this addition. When the antenna configuration and the current distribution are known, the radiated field is determined by integrating the contributions which are made by each elementary length.[1]

The wave front of the radiated wave lies in a plane perpendicular to a line drawn toward the antenna, and the waves are polarized in the same direction as the antenna. Thus a plane can be passed through the antenna and an electrostatic flux line of the radiated wave, while the magnetic flux is perpendicular to such a plane.

Current Distribution in an Antenna.—An antenna represents a circuit having distributed constants and so has a current distribution of the type discussed in Sec. 15. Strictly speaking, the inductance and capacity per unit length are not the same for all parts of the antenna, so that an exact solution for the current distribution is extremely complicated. Experiments have shown, however, that for all practical purposes the current in an antenna can be considered as having a sinusoidal distribution in which the phase differs by 180° in the adjacent half-wave-length sections.[2] The current is zero at the open ends of the antenna and approaches zero at all points that are an exact multiple of a half wave length distant from the open end, while the current is maximum at points which are odd quarter wave lengths distant from the open ends. The length of an antenna expressed in wave lengths is very nearly equal to the length between extreme ends, measured in terms of the wave length of a wave traveling with the velocity of light, *i.e.*, a wave length of the radio wave. Examples of current distribution in a number of typical antennas are shown in Fig. 268. The current in each case follows a sinusoidal law, and is zero at the open ends. When the lower end of the antenna is grounded, as in *a* to *f*, the antenna length need not be an exact multiple of a quarter wave length, although the curve of current distribution is always a section of a sine curve. When both ends are ungrounded,

[1] For an example of such an integration, see S. A. Levin and C. J. Young, Field Distribution and Radiation Resistance of a Straight Vertical Unloaded Antenna Radiating at One of Its Harmonics, *Proc. I.R.E.*, vol. 14, p. 675, October, 1926.

[2] This is shown by experimental results published by R. H. Wilmotte, Distribution of Current in a Transmitting Antenna, *Jour. I.E.E.* (*London*), vol. 66, p. 617, June, 1928.

as in g, h, and i, the antenna must be operated at a frequency that makes it an exact multiple of a half wave length long, or the antenna must be loaded in the center as shown in Fig. 268j. Where the antenna is bent, as in Fig. 268e and f, the wave length is very nearly the same value as though there were no bend.

When the antenna is not an exact multiple of a half wave length when ungrounded, or of an odd-quarter wave length when grounded at one end, it is necessary to place a loading reactance (either an inductance or capacity) in series with the antenna in order to obtain resonance. An inductive reactance is equivalent to adding length to the antenna while a capacity has the effect of subtracting length. In either case the current distribution along the actual antenna is a section of a sine curve and is based upon the actual antenna length involved. Thus if a grounded antenna has a length of $\lambda/8$ wave lengths and is loaded up to $\lambda/4$ by an inductance, the current distribution is one-eighth of a sine-wave cycle, as shown in Fig. 268a.

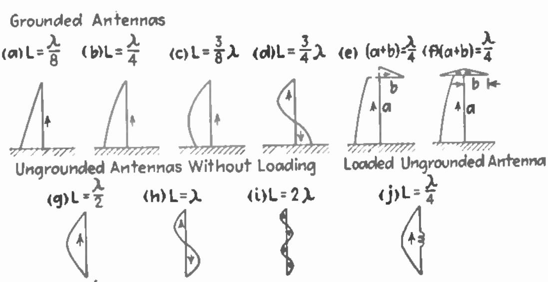


FIG. 268.—Current distribution in typical antennas. In each case the current has a sinusoidal distribution and is zero at the open ends.

The common method of representing antenna current distribution shown in Fig. 268 is not strictly correct, for, as explained in Sec. 15, losses prevent the current from going to zero except at an open end. The resistances of antennas are so low, however, that the errors that are introduced by assuming a sinusoidal current distribution that passes through zero are so small as to be negligible.

Effect of the Ground—Image Antennas.—In all practical antennas it is necessary to take into account the fact that the ground reflects the energy which is radiated in its direction. The magnitude of this effect can ordinarily be determined by assuming that the ground is a perfectly conducting plane and then using the method of images. Consider the situation at Fig. 269a, which shows the electrostatic flux lines that are produced in the immediate vicinity of an antenna grounded at one end. A comparison of this with a similar flux distribution which results when two such antennas are arranged as images of each other, as shown at Fig. 269b, indicates that the flux distribution in the upper half of Fig.

269b is exactly the same as that above the ground in Fig. 269a. Thus as far as the electrostatic fields above the ground plane are concerned it makes no difference whether one deals with the grounded antenna system of Fig. 269a, or with the actual antenna and its image as at Fig. 269b. The effect of the ground can therefore be determined by replacing the ground by an image antenna and then calculating the radiated field produced by the joint action of the actual antenna and its image.

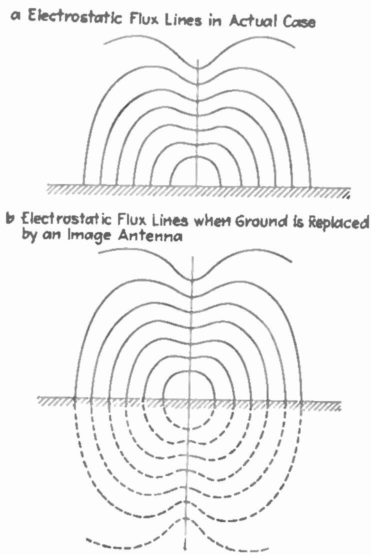


FIG. 269.—Electrostatic field in vicinity of grounded antenna, showing how the effect of the ground can be taken into account by making use of an "image" antenna.

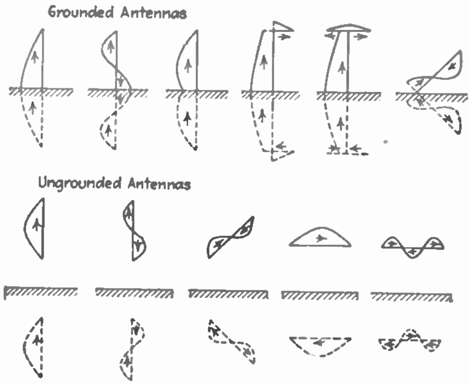


FIG. 270.—Images for common types of antennas.

Examples of image antennas for a number of cases are given in Fig. 270. The general principles for setting up the image antenna are as follows:

1. The image antenna has the same geometric configuration as the actual antenna and is located so as to be a mirror image with the ground surface the plane of symmetry.

2. The current distribution in the image and actual antennas are of identical character, and the magnitudes of the currents are the same in both cases.

3. When one end of the actual antenna is nearer earth than the other, the current in the end of the image nearest the earth flows in the opposite direction with respect to the earth surface from the current in the corresponding part of the actual antenna.

4. When the actual antenna is parallel with the earth the currents in corresponding parts of the actual and image antennas flow in opposite directions.

The only approximation involved in using the image method to take into account the effect of the ground is in considering the earth to be a perfect reflector in which no energy is consumed. While this is never completely realized, the error involved is usually so small that results calculated on the basis of a perfectly reflecting earth agree with the observed behavior well enough for most practical purposes.

Total Radiated Energy.—The rate at which energy passes through 1 sq cm of surface located in the wave front is the average amount of energy contained in 1 cc of wave multiplied by the velocity of light. Thus if the wave at the point in question has a strength ϵ volts effective value per centimeter, the average energy per cubic centimeter is $\epsilon^2 0.08842 \times 10^{-12}$ joule. Multiplying this by the velocity of light in centimeters per second shows that the rate at which energy flows through each square centimeter of wave front is $0.00265\epsilon^2$ joule per second. The energy radiated from the entire antenna system can be determined by imagining that the antenna is at the center of a very large sphere and then adding up the energy that flows through each square centimeter of the spherical surface that is above the ground. While the determination of the total energy radiated from an antenna is entirely straight-forward, the mathematical expressions encountered in carrying out the integration over the spherical surface are difficult to handle and are not familiar to the ordinary engineer.[1]

Radiation Resistance, Loss Resistance, and Antenna Efficiency.—The total amount of energy radiated from a transmitting system can be conveniently measured in terms of a "radiation" resistance, which is the resistance that when inserted in series with the antenna will consume the same amount of power as is actually radiated. While the radiation resistance is a purely fictitious quantity, the antenna acts as though such a resistance is present because the loss of energy by radiation is in its effects equivalent to a like amount of energy dissipated in a resistance. The value of the radiation resistance is determined by the antenna construction, and particularly the size measured in wave lengths, the relation to ground and other conducting objects, such as towers, buildings, and trees, etc., and upon the point on the antenna to which the resistance is referred (*i.e.*, the point at which the radiation resistance is considered as being lumped). Unless specifically stated otherwise, it is customary to refer the radiation resistance to a current loop. An idea of the magnitude of radiation resistance can be gained from the fact that a grounded vertical wire one-quarter wave length long has a radiation resistance of 36.2 ohms, while a half wave-length antenna remote from the ground has a radiation resistance of 72.4 ohms, and a vertical antenna

[1] Examples of such integrations are to be found in the following references: Stuart Ballantine, On the Radiation Resistance of a Simple Vertical Antenna at Wave Lengths below the Fundamental, *Proc. I.R.E.*, vol. 12, p. 823, December, 1924; S. A. Levin and C. J. Young, Field Distribution and Radiation Resistance of a Straight Vertical Unloaded Antenna Radiating at One of Its Harmonics, *Proc. I.R.E.*, vol. 14, p. 675, October, 1926; G. W. Pierce, "Electric Oscillations and Electric Waves," McGraw-Hill Book Company, Inc., New York, 1920.

grounded at the lower end and less than a quarter wave-length long possesses a radiation resistance of approximately $520(l/\lambda)^2$ ohms, where l/λ is the length of the antenna measured in wave lengths.[1] The way in which the radiation resistance of a grounded vertical wire varies with wire length is shown in Fig. 271.

In addition to the radiated energy, energy is also lost in the antenna system as a result of wire and ground resistance, corona, eddy currents induced in neighboring masts, guy wires and other conductors, and dielectric losses arising from imperfect dielectrics, such as trees and insulators, located in the field of the antenna. These losses can be represented in the same way as the radiated energy, *i.e.*, by a resistance which when inserted in series with the antenna will consume the same amount of power as is actually dissipated by the antenna. The magnitude of this antenna resistance depends upon the frequency, the antenna construction, the conditions existing in the immediate vicinity of the antenna, and the point in the antenna system to which the resistance is referred. The total antenna resistance is the sum, $R_r + R_l$, of the radiation resistance R_r and the loss resistance R_l, and determines the amount of energy which must be supplied to the antenna to produce a given current.

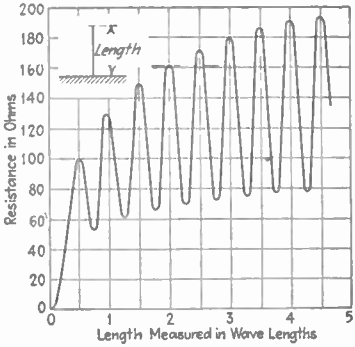


FIG. 271.—Radiation resistance (referred to a current loop) of a vertical wire grounded at the lower end.

The efficiency of the antenna as a radiator is the ratio $R_r/(R_r + R_l)$ of radiation to total resistance. This represents the fraction of the total energy supplied to the antenna which is converted into radio waves. At a given frequency, with other things remaining equal, the radiation efficiency varies with the ratio h/λ of antenna height h to the wave length λ. This is because the radiation resistance is roughly proportional to the square of the antenna height while the antenna losses increase very slowly if at all with added height. Also, if the ratio h/λ is kept constant as the wave length is decreased, it will be found that the radiation efficiency of the antenna will almost invariably increase. The reason for this is that the radiation resistance remains constant if the ratio h/λ is unchanged, but since the higher frequency antenna is smaller its loss

[1] The radiation resistance of most types of short grounded antennas can be determined by the formulas of Fulton Cutting, A Simple Method of Calculating Radiation Resistance, *Proc. I.R.E.*, vol. 10, p. 129, April, 1922. The radiation resistance of longer antennas is taken up by S. Ballantine, *loc. cit.;* S. A. Levin and C. J. Young, *loc. cit.*, and S. Ballantine, *Proc. I.R.E.*, vol. 15, p. 245, March, 1927. This last reference is the source of the information presented in Fig. 271.

resistance is correspondingly less, and a larger proportion of the energy supplied is converted into radio waves. As a result of these relations the radiation efficiency of short-wave antennas is high, often exceeding 90 per cent, while the efficiency tends to decrease as the distance represented by a wave length increases, and may be as low as 5 per cent in antennas designed to radiate the longest waves used in radio communication.

Effective Height.—The term "effective height" is sometimes applied to antenna systems, particularly those in which the height is a quarter wave length or less. As ordinarily defined, the effective height represents the length of elementary radiator of Fig. 267 which when carrying a uniform current equal to the current flowing at the base of the antenna will produce the same field intensity as is actually radiated. When the actual height is not in excess of one-fourth wave length the effective height will usually range between 50 and 90 per cent of the actual antenna height, with the exact value depending on the antenna construction. The method of calculating the effective height is outlined in Sec. 123.

Induction Fields.—The electrostatic and magnetic fields having strengths given by Eqs. (154*a*) and (154*b*) do not include the ordinary magnetic and electrostatic induction fields that are present in the immediate vicinity of the antenna even at low frequencies where the radiation is negligible. The total magnetic field that is produced in the vicinity of an elementary antenna, such as that of Fig. 267, is given by the following equation:

Magnetic field in lines per square centimeter =

$$\frac{2\pi(\delta l)I}{10d\lambda}\cos\omega\left(t-\frac{d}{c}\right)\cos\theta-\left(\frac{(\delta l)I}{10d^2}\right)\sin\omega\left(t-\frac{d}{c}\right)\cos\theta \qquad (155)$$

The notation is the same as in Eq. (154*a*) except that all the units of length are in centimeters. The first term of this equation represents the radiation field given in Eq. (154*a*), while the second term represents the induction field that gives rise to the self-inductance of the antenna system. The induction magnetic field is in phase with the current flowing in the radiator (after making allowance for the time required in propagation) and is inversely proportional to the square of the distance, while the radiation field is in time quadrature, *i.e.*, 90° out of phase, with the current and is inversely proportional to the first power of the distance. Because of the way in which it varies with distance the induction field is of importance only in the immediate vicinity of the antenna, where it is much stronger than the radiation field. At a distance of λ 2π from the antenna the two fields are equal, while at greater distances the radiation field predominates.

Along with the induction magnetic field there is also an induction electrostatic field, which is in time phase with the electrostatic field of the radiated wave and like the magnetic induction field dies out much more rapidly with distance than does the radiated wave.

119. Fundamental Properties of Receiving Antennas and Reciprocal Relations Existing between Transmitting and Receiving Properties.— A receiving antenna is able to abstract energy from a passing radio wave as a result of the voltages which the magnetic flux of the wave induces in the antenna. These induced voltages are distributed along the entire length of the antenna and have a value which per meter of antenna length is $\epsilon \cos \psi \cos \theta$, where ϵ is the field strength of the wave in volts per meter, ψ is the angle between the plane of polarization and the wire in which the voltage is induced, and θ is the angle between the wave front and the direction of the antenna wire. It will be observed that the quantity $\epsilon \cos \psi \cos \theta$ is the component of the field strength that has a wave front parallel to the antenna and is polarized in the same plane as the antenna.

For purposes of analysis it is convenient to replace the distributed voltages actually induced in the antenna by a lumped voltage placed in series with the antenna at the point where the load impedance is placed, and having the proper value to produce the same current that the distributed voltages develop. The magnitude of this lumped voltage is equal to the product of the strength of the radio wave and the effective height of the antenna, where the effective height is defined in terms of the point in the antenna at which the lumped voltage is assumed to act, and for a wave of the angle of incidence and plane of polarization of the passing wave. The effective height of an antenna is the same for reception as for transmission and can be calculated as outlined in Sec. 123.

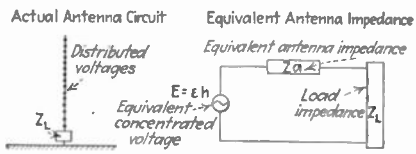


FIG. 272.—Actual receiving antenna with load impedance Z_L and distributed induced voltages, together with equivalent antenna circuit in which the distributed voltages are replaced by a lumped voltage E and the distributed antenna impedance by an equivalent concentrated impedance Z_a.

In order to make use of the energy which an antenna is able to abstract from the radio wave it is necessary to insert a suitable load resistance in the antenna circuit. The proper value of this resistance can be determined by replacing the actual antenna with its distributed constants and distributed induced voltage by the equivalent circuit shown in Fig. 272, in which the distributed voltages are replaced by the lumped voltage E, and the impedance against which the distributed voltages operate is replaced by an equivalent lumped antenna impedance Z_a. The current that flows in this equivalent antenna circuit is exactly the same as the current that flows at the corresponding part of the real antenna. The load impedance that must be inserted in series with the equivalent antenna circuit to consume the maximum amount of energy must meet the following two requirements: (1) the resistance component of the load must equal the resistance component of the impedance Z_a; and (2) the reactive component of the load impedance must be equal in magnitude and opposite in sign to the reactive component of the equivalent antenna impedance Z_a. When the antenna is tuned to resonance with the frequency being received, the antenna impedance Z_a is a pure resistance equal to the total antenna resistance, $R_r + R_l$, referred to the point where the load resistance is to be inserted, and the proper value of load resistance is $R_L = (R_r + R_l)$.

Received Energy.—The total energy which the receiving antenna abstracts from the passing radio wave represents the energy dissipated in the equivalent antenna circuit and so is given by the relation

$$\text{Total power in watts abstracted from radio wave} = \frac{(\epsilon h)^2}{R_L + R_r + R_l} \tag{156}$$

where ϵ is the field strength (r.m.s. value) of the radio wave in volts per meter; h, the effective height of the antenna in meters; R_r, the antenna radiation resistance; R_l, antenna loss resistance; and R_L, the load resistance. The fraction $R_L/(R_r + R_l + R_L)$ of this total energy represents the portion of the abstracted energy which is usefully employed. Of the remainder, part is accounted for by the antenna losses, such as wire and ground resistance, etc., while the rest is reradiated. This reradiation of energy results from the fact that when current flows in an antenna, radiation takes place irrespective of whether the voltage producing the current is derived from a passing radio wave or from a vacuum tube.

The maximum amount of energy which it is theoretically possible for a given antenna to abstract from a passing radio wave occurs when the antenna loss resistance is made negligibly small, and when the load resistance R_L is equal to the radiation resistance. Under these conditions the rate at which energy is abstracted from the wave is $(\epsilon h)^2/2R_r$ watts, and since half of this is reradiated the maximum possible amount of power that can be delivered to a load resistance R_L is $(\epsilon h)^2/4R_L$ watts. Calculations show that a section of wave front extending for only about one-quarter of a wave length on each side of the receiving antenna will be capable of supplying this received energy.

The electromagnetic and electrostatic fields in the vicinity of a receiving antenna are the sum of the fields produced by the radio wave and by the current in the receiving antenna. The result is that the receiving antenna causes a distortion in the field pattern in its immediate vicinity, as seen in Fig. 273, which shows the direction of magnetic flux lines as experimentally determined in a particular case.[1] Analysis shows that the effect of the receiving antenna on the passing wave is, first, abstraction of energy which weakens the main wave, and, second, reflection or reradiation of energy, which redistributes the energy of the passing wave in a

[1] See Henry C. Forbes, Re-radiation from Tuned Antenna Systems, *Proc. I.R.E.*, vol. 13, p. 363, June, 1925; F. A. Kolster and F. W. Dunmore, The Radio Direction Finder and Its Application to Navigation, *Bur. Standards Sci. Paper* 428.

manner depending upon the antenna tuning (see Sec. 124 for a further discussion of this point).

Relations between Receiving and Radiating Properties.—The properties of an antenna when used to abstract energy from a passing radio wave are similar in nearly all respects to the corresponding properties of the same antenna when acting as a radiator. Thus the directional characteristics, the current distribution, the effective height, and the impedance of the antenna, are the same in reception as in transmission. The only difference is in the radiation resistance, which in the case of receiving antennas depends upon the inserted load impedance and tends to be higher than when radiating.[1] These reciprocal relations between the transmission and reception properties are extremely useful because they make it possible to deduce the merits of a receiving antenna from transmission tests and *vice versa.* Thus the directional characteristics of a given antenna can be determined either by operating this antenna as a receiver and moving a portable transmitter about, or by using the antenna as a transmitter and making observations of field strength on a portable receiving set. Similarly the effective height can be calculated by determining the voltage that must be inserted in the antenna to produce the same current as is produced by a wave of known strength, or it can be obtained by determining the field strength which the antenna produces at a known distance when the antenna is transmitting with a known current.

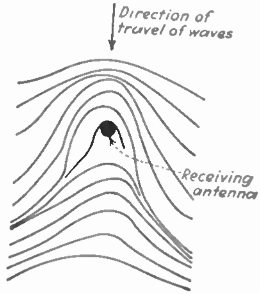


FIG. 273.—Magnetic flux distribution in the immediate vicinity of a receiving antenna, showing how the current induced in the receiving antenna produces radiation and induction fields that distort the flux distribution. In this example the receiving antenna is tuned to a frequency slightly less than that of the wave.

These reciprocal relations between the transmitting and receiving properties of an antenna are incorporated in two reciprocal theorems. The first of these was discovered by Rayleigh and extended to include radio communication by John R. Carson. It is to the effect that if an electromotive force E inserted in antenna 1 causes a current I to flow at a certain point in a second antenna 2, then the voltage E applied at this point in the second antenna will produce a current I at the point in antenna 1 where the voltage E was originally applied. The second theorem is due to Sommerfeld and Pfrang and states that with two antennas A and B of arbitrary orientation the average phase and intensity of the field at B when A is transmitting with a given average power is the same as the average phase and intensity of the field at A when B

[1] See Raymond M. Wilmotte, Generalized Theory of Antennae, *Exp. Wireless and Wireless Eng.*, vol. 5, p. 119, March, 1928.

is transmitting with the same average power.[1] The Rayleigh-Carson theorem fails to be true only when the propagation of the radio waves is appreciably affected by an ionized medium in the presence of a magnetic field, and so holds for all conditions except short-wave transmission over long distances. The Sommerfeld-Pfrang theorem, on the other hand, in addition to having the same limitations as the Rayleigh-Carson theorem, also fails to hold when the antennas are near the earth.

120. Directional Characteristics of Simple Antennas.—The directional characteristics of transmitting antennas are important because

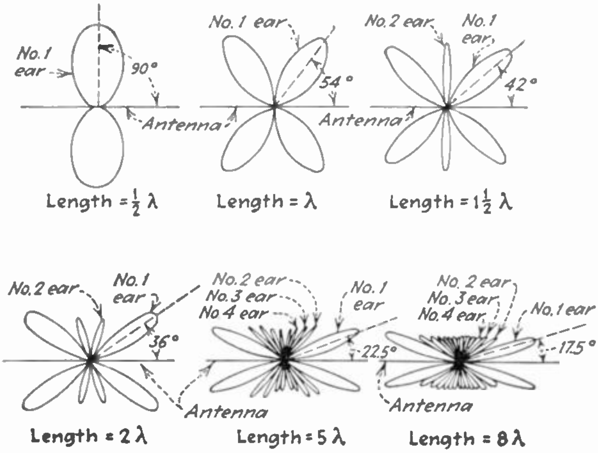


FIG. 274.—Polar diagrams showing strength of field radiated in various directions from an antenna consisting of a wire remote from the ground.

only those waves radiated in certain directions from the transmitter will reach a particular receiving point, while all energy radiated in other directions is wasted as far as transmission to this receiver is concerned. The directional characteristics of receiving antennas are likewise important because an antenna which abstracts much larger quantities of energy from waves coming from the direction of the transmitter than from waves of equal strength arriving in other directions will not pick up interfering signals and static arriving from other directions.

Unloaded Antenna Operated at Resonance.—The directional characteristics of an antenna consisting of a single wire having a length that is an exact multiple of a half wave length (*i.e.*, an unloaded antenna operated at a resonant frequency) and far removed from other objects (particularly the ground) is given by Eq. (161). Polar diagrams giving the directional characteristics of several such antennas in a plane containing the wire are shown in Fig. 274. In these diagrams the length of the radius vector

[1] For a discussion of the Rayleigh-Carson and Sommerfeld-Pfrang theorems see John R. Carson, Reciprocal Theorems in Radio Communication, *Proc. I.R.E.*, vol. 17, p. 952, June, 1929.

in any particular direction is proportional to the strength of the field radiated in that direction. If the directional characteristic is viewed in three-dimensional space it consists of a figure of revolution having a cross section as shown in Fig. 274. An examination of Fig. 274 shows that the directional characteristic contains a number of lobes, the largest of which are always the lobes making the smallest angle with the axis of the wire. Increasing the antenna length as measured in wave lengths has the effect of reducing the angle between the axis of the wire and the direction of the large lobes, and also of increasing the number of lobes present, as is clearly evident in the figure.

The important features of the directional characteristics of a wire antenna in space are incorporated in the curves of Fig. 275, which show the angles with respect to the wire axis at which the radiation is maximum and zero, and also show the relative amplitude of the successive lobes or ears. Thus reference to Fig. 275 shows that a wire five wave lengths long has five lobes in each quadrant, with the maximum of these lobes coming at angles of 22.5°, 46°, 60°, 73°, and 84° with respect to the wire axis, while the relative amplitudes of these lobes are 2.25, 1.40, 1.20, 1.05, and 0.95, respectively.

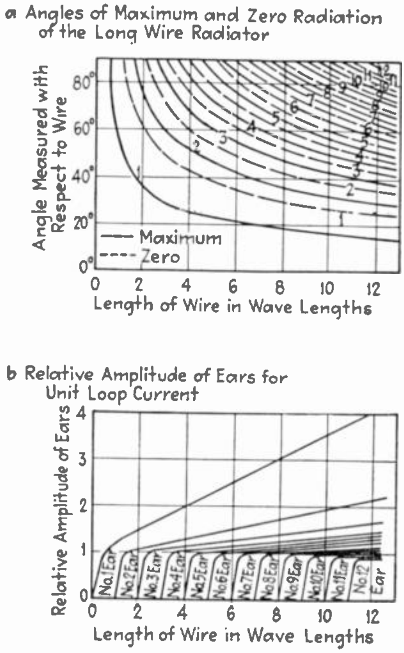


FIG. 275.—Angles at which the radiation from an isolated wire antenna is zero and maximum, together with relative field strength of ears. The ears are numbered so that Ear 1 is the lobe nearest the direction in which the antenna points.

It has already been shown that the radiation from an elementary length of wire is proportional to the cosine of the angle with respect to a plane perpendicular to the wire. The reason that the directional characteristic of a long wire differs from that of a short length is that the current in different parts of the long antenna may not be in the same phase, and the distance from a remote point to various parts of the long antenna will not be the same. The result is that the fields radiated from different portions of the long antenna add together vectorially and give a sum which depends upon the direction and upon the current distribution along the wire.

Effect of Earth.—When a wire antenna is located near the ground the directional characteristic is affected by the presence of the earth. This is because the energy radiated from the antenna toward the earth is reflected with a reversal in the phase of the electrostatic field, and the reflected wave will either add to or subtract from the directly radiated field according to the relative phases.

When an antenna is placed horizontally at a height H above the earth the effect which the earth has on the directional characteristics of the antenna can be obtained by multiplying the directional charac-

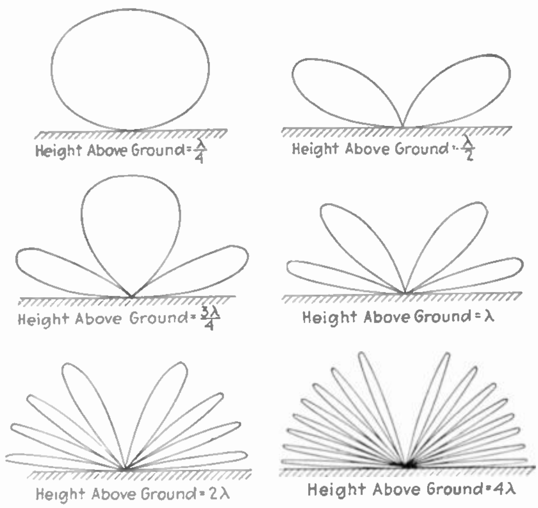


FIG. 276.—Polar diagram of factor $\left|2 \sin\left(2\pi\frac{H}{\lambda} \sin \theta\right)\right|$ for various values of $\frac{H}{\lambda}$ showing how the height of a horizontal antenna above earth affects the directional characteristic of the antenna.

teristic of the antenna when remote from the earth, as given by Eq. (161), by the factor $\left|2 \sin\left(2\pi\frac{H}{\lambda} \sin \theta\right)\right|$, that is

$$\left.\begin{array}{l}\text{Actual radiation from}\\ \text{horizontal antenna in}\\ \text{presence of the ground}\end{array}\right\} = 2 \sin\left(2\pi\frac{H}{\lambda} \sin \theta\right)\left(\begin{array}{l}\text{radiation from}\\ \text{antenna when}\\ \text{in free space}\end{array}\right) \tag{157}$$

where H/λ is the height of the horizontal antenna above ground, measured in wave lengths; and θ is the angle of elevation above the horizontal. The nature of the factor $\left|2 \sin\left(2\pi\frac{H}{\lambda} \sin \theta\right)\right|$, which takes into account the effect of the earth, is indicated by the polar diagrams of Fig. 276. These show that the presence of the earth causes cancellation of radiation along the horizontal and also at certain vertical angles. In between these directions of zero radiation are lobes having a maximum value of 2. The angle of elevation of the first lobe above the horizontal decreases as the height of the antenna above earth is increased, and in order to obtain strong radiation in directions approaching horizontal it is necessary that the height above earth be at least one wave length.

The angles of elevation with respect to the horizontal at which the factor $\left|2 \sin\left(2\pi\frac{H}{\lambda} \sin\theta\right)\right|$ is zero and maximum are shown in Fig. 277 as a function of height above earth measured in wave lengths. Thus when the horizontal antenna is one wave length above ground the effect of the earth causes complete cancellation of radiation in directions which

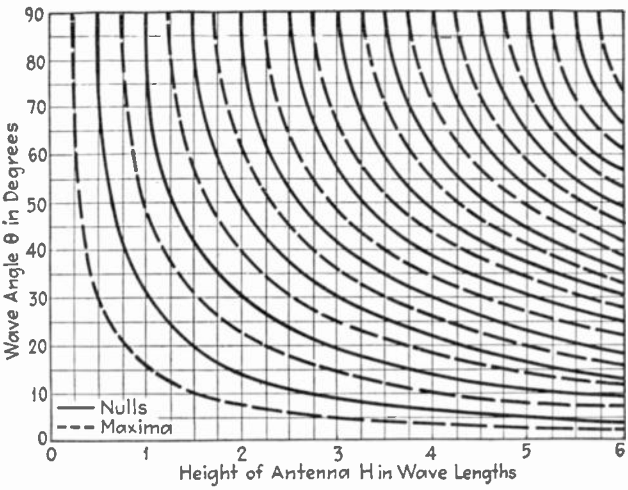


FIG. 277.—Chart showing the vertical angles at which the factor $|2 \sin (2\pi\frac{H}{\lambda} \sin \theta)|$ is maximum and zero. These represent the vertical angles at which the reflections from the ground cause complete reinforcement and complete cancellation, respectively, of the radiation.

make angles of 0°, 30.5°, and 90° with respect to the horizontal, while the presence of the earth reinforces the radiation in directions which make angles of 15.5° and 49° with respect to the horizontal.

When an antenna consisting of a single wire is placed vertically with its center at a height H above the earth the effect of the near-by earth is very much the same as when the wire is horizontal. Thus if the antenna is an exact even number of half wave lengths long the effect of the earth can be taken into account by multiplying the directional characteristic which the antenna would have in free space by the same factor $\left|2 \sin\left(2\pi\frac{H}{\lambda} \sin\theta\right)\right|$ that takes into account the effect of the earth with a horizontal antenna. If on the other hand the vertical antenna is exactly an odd number of half wave lengths long the effect of the earth is taken into account by multiplying the directional characteristic which the antenna would have in space by the factor $\left|2 \cos \left(2\pi \frac{H}{\lambda} \sin \theta\right)\right|$. It will be observed that the only difference between the factors for lengths that are odd and even multiples of a half wave length is that one is a cosine and the other a sine of the same angle. This difference is equivalent to interchanging the positions of the maxima and minima, *i.e.*, an angle of elevation at which the radiation is zero when the length is an even number of half wave lengths is an angle of elevation at which the radiation goes through a maximum when the antenna is an odd number of half wave lengths long.

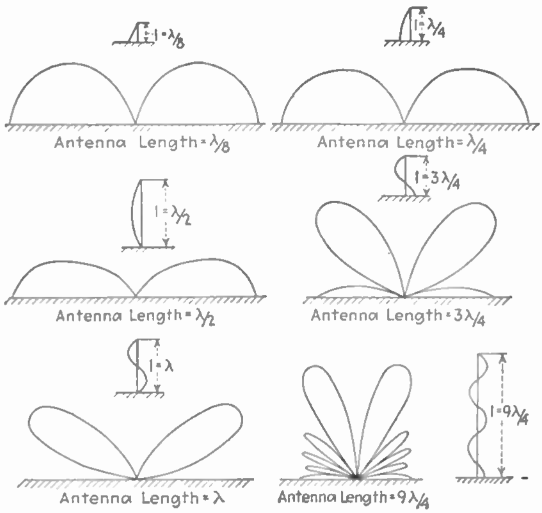


FIG. 278.—Directional characteristics in a vertical plane of field produced by grounded vertical antennas of varying lengths.

When the antenna consists of a vertical wire grounded at the lower end and an odd number of quarter wave lengths long, the directional characteristic is exactly the same as the directional characteristic of a wire in free space having a length twice the length of the grounded wire. The directional characteristics of such a grounded antenna can therefore be calculated with the aid of Eq. (161) by considering that the l in the equation is twice the antenna length. If the length of the grounded wire is not exactly an odd multiple of a quarter wave length the resulting directional characteristic must be calculated with the aid of Eq. (162). Examples of the directional characteristic of grounded antennas of varying lengths are shown in Fig. 278. When the length is small compared with a quarter wave length the radiation is almost exactly proportional to the cosine of the angle of elevation, but as the length is increased up to slightly more than one-half wave length the effect is to concentrate the radiation more and more sharply along the horizontal. As the antenna length is still further increased new lobes appear at the vertical. With increasing antenna length these rotate toward the horizontal, first increasing and later decreasing in size. The vertical angle at the maximum becomes greater as the length is increased.

The above discussion has been based on the assumption that the earth is a perfect reflector, *i.e.*, has either infinite conductivity or an infinite dielectric constant. Under most conditions the results calculated on this supposition are in satisfactory agreement with the observed behavior, but this is not necessarily the case with all earths. The effect of an imperfect earth is to cause the wave reflected by the ground to have a magnitude smaller than the incident wave because of earth losses, and to have its phase dependent upon the power-factor angle of the earth.[1]

Flat-top Antennas.—Antennas used at long waves are usually short compared with a quarter wave length and are provided with a flat top. The radiation from the vertical portion of such an antenna is vertically polarized and varies with the cosine of the angle of elevation, as in the case of the elementary antenna of Fig. 267. The radiation from the flat top is horizontally polarized, but is small because the flat top is only a small fraction of a wave length above earth, and can ordinarily be neglected without introducing appreciable error. The directional characteristic of long-wave antennas in a horizontal plane is circular as far as the vertically polarized radiation is concerned. The horizontally polarized radiation from the flat top is decidedly directional, however, and in some long-wave transmitters in which the flat top is very long compared with the height it is claimed that rather pronounced directivity is obtained.

The directional characteristics of loop antennas are discussed in Chap. XVI in connection with radio aids to navigation.

121. Fundamental Principles of Antenna Arrays.[2]—An antenna array consists of a number of antennas separated in space and electrically connected in a definite phase relation. While there are an infinite number of possible antenna arrays the only ones that are used commercially to any extent are the broadside array, which consists of a number of

[1] The relation between the wave reflected from an imperfect earth and the incident wave can be calculated by methods described by P. O. Pedersen, "The Propagation of Radio Waves," Chap. VIII, G. E. C. Gad, Copenhagen, Denmark (in English).

[2] The discussion in this section is based to a large extent upon the following references: Ronald M. Foster, Directive Diagrams of Antenna Arrays, *Bell System Tech. Jour.*, vol. 5, p. 292, April, 1926; G. C. Southworth, Certain Factors Affecting the Gain of Directive Antennas, *Proc. I.R.E.*, vol. 18, p. 1502, September, 1930; E. J. Sterba, Theoretical and Practical Aspects of Directional Transmitting Systems, *Proc. I.R.E.*, vol. 19, p. 1184, July, 1931.

conductors arranged along a line and excited in phase; the end-fire array, which consists of a similar line of conductors uniformly spaced but with a progressive phase difference between adjacent radiators; and the co-linear array, which consists of antennas arranged co-linearly with each other and in phase; or some simple combinations of these fundamental array types.

Antenna arrays make possible highly directional antenna systems. Thus the broadside array of Fig. 279, which consists of a series of elementary radiators uniformly spaced along a line and excited in phase, concentrates the radiation in a direction at right angles to the line of the array and gives very little radiation in other directions. This result is achieved because the array elements are so placed that the component radiations are in the same phase in the direction of the array maximun but are more or less out of phase in all other directions, giving partial or complete cancellation.

The relative intensity of the radiation in different directions from an array consisting of elementary radiators spaced at uniform intervals along the X-, Y-, and Z-axes and with a uniform progressive phase difference between adjacent radiators in the directions of the X-, Y-, and Z-axes can be calculated with the aid of Eq. (165). This equation gives the directional characteristic of an array composed of radiators which send out their radiation uniformly in all directions. The directional characteristic of an array composed of antennas which do not radiate uniformly in all directions can be obtained by multiplying the values calculated from Eq. (165) for the case of spherical radiators by the directional characteristics of the individual antenna which goes to make up the array.

The effect which the presence of the earth has on the directional characteristics of an antenna array can ordinarily be taken into account by considering that the earth modifies the directional characteristics of the individual antennas of which the array is composed. The only case in which difficulty is encountered in applying this method of taking into account the earth effects is when the antenna array consists of several elements stacked one above the other. When the currents in the different tiers of such an array are in phase the correct results can be obtained by assuming that the directional characteristic of the individual antenna of the array is the directional characteristic that would be obtained when the center of such an antenna is placed above ground at a height corresponding to the distance above the earth of the mid-point of the array. If the currents in the different tiers are not in phase this method falls down, and the analysis becomes much more complicated.

The merit of an antenna array is most conveniently measured in terms of the array "gain," which is defined as the power which must be supplied to a standard comparison antenna to lay down a given field strength in the desired direction divided by the power that must be supplied to the antenna array to accomplish the same result. The comparison antenna is usually taken either as a wire one-half wave length long and at arbitrary height above earth and orientation, or as one of the individual antennas of which the array is composed. The latter method of expressing gain has the merit of giving a result which is nearly independent of earth effects and of the directive characteristics of the comparison antenna, except when the comparison antenna is sharply directional in the same planes as is the array. The merit of an array is calculated by integrating the energy radiated over a spherical surface and comparing the value thus obtained with the results of a similar integration about the comparison antenna carrying sufficient current to produce the same field strength in the desired direction.

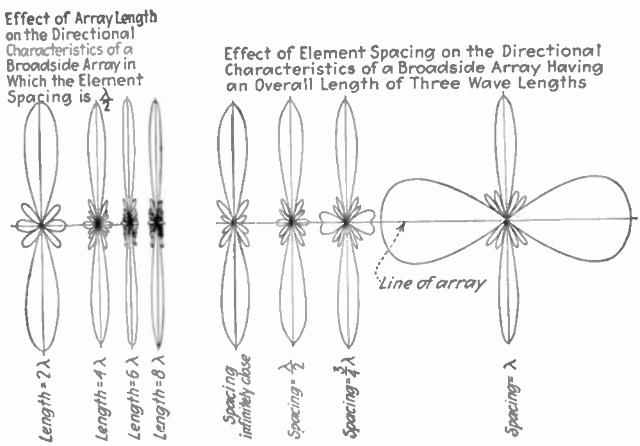


FIG. 279.—Effect of array length and element spacing on the directional characteristics of field radiated from a broadside array in a horizontal plane.

The Broadside Array.—The simple broadside antenna array consists of a number of antennas spaced at uniform distances along a horizontal line and connected so as to be excited in phase. Typical directional characteristics for such arrays are shown in Fig. 279 and consist of a beam sharply defined in the horizontal plane and having its maximum at right angles to the line of the array. The sharpness of the main beam in a horizontal plane depends primarily upon the over-all length of the array and is substantially independent of the spacing between adjacent antennas provided this spacing does not exceed about three-quarters of a wave length. With greater spacings parasitic lobes of large magnitude develop and spoil the directional characteristics. These effects of array length and element spacing on the directional characteristics of the field radiated in a horizontal plane are shown by the series of diagrams of Fig. 279, which have been calculated with the aid of Eq.

(165) on the assumption that the individual antennas of the array are vertical wires radiating uniformly in all horizontal directions. It will be observed that increasing the array length while maintaining the spacing between antennas constant causes a proportional increase in the sharpness of the main lobes, and increases the number but not the magnitude of the minor lobes. Changing the spacing of the individual antennas (and hence the number of antennas in a given length while maintaining the over-all length constant) varies the size of the minor lobes in the array pattern. The area of these minor lobes is, however, inappreciable until the spacing exceeds about $3\lambda/4$, when two or more of the minor lobes suddenly expand to a size comparable with the main lobes.

The directional characteristic of a linear broadside array in a vertical plane at right angles to the physical line of the array is exactly the same as the corresponding directional characteristic of the individual antennas of which the array is composed. The vertical radiation in other planes, such as the vertical plane parallel to the line of the array, is small because the array radiates most of its energy in a plane at right angles to its length.

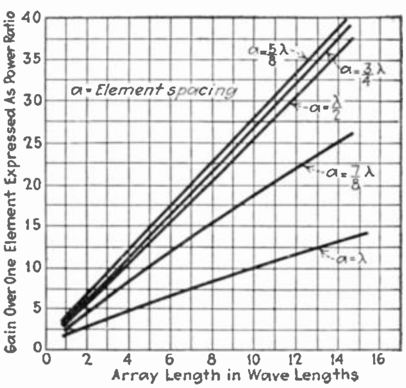


FIG. 280.—Gain of broadside array as a function of array length for various element spacings. These curves are for the case where radiation from the individual antenna is proportional to the cosine of the angle of elevation and give the gain over a single element of the array.

The effect of array length and antenna spacing on the array gain is shown in Fig. 280. The gain is almost exactly proportional to the over-all length of the array and is practically independent of spacing of the antennas composing the array provided this spacing does not exceed about $3\lambda/4$. These gain curves are closely related to the directional characteristics of Fig. 279, since the energy radiated in any direction is proportional to the square of the field strength in that direction.

Simple End-fire Arrays.—An end-fire differs from a broadside array in that there is a progressive phase difference between adjacent antennas in the array, this difference in phase ordinarily being a fraction of a cycle equal to the spacing between adjacent antennas measured in wave lengths. Thus if the adjacent antennas are one-quarter wave length apart the individual antennas of the array are so connected that as one travels from one end of the array to the other there is a progressive phase shift of 90° between adjacent antennas. The effect of phasing the array elements in this way is to cause the radiation in a horizontal plane to be concentrated in a unidirectional beam directed along the line of the antenna array and pointed toward the end in which the phase lags. The

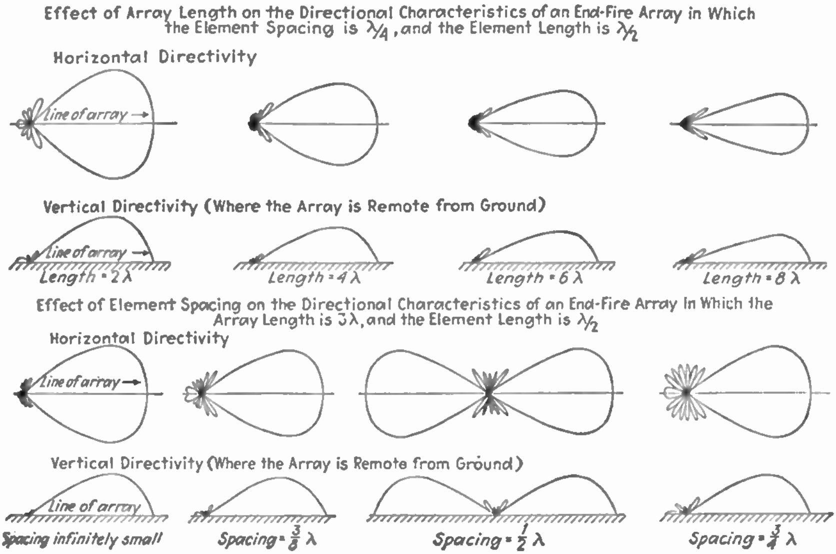


FIG. 281.—Effect of array length and element spacing on the directional characteristic of field radiated from an end-fire array.

radiation in a vertical plane parallel to the line of the array, *i.e.*, in a vertical plane passing through the direction of maximum horizontal radiation, is more or less concentrated along the horizontal.

These effects are illustrated by the directional characteristics shown in Fig. 281. It will be observed that, as in the case of the broadside array, the directivity depends almost solely upon the over-all length of the array, and is virtually independent of the element spacing provided this spacing does not exceed a certain critical value which in the case of the end-fire array is approximately $3\lambda/8$. The effect of array length and element spacing on the gain of end-fire arrays is shown in Fig. 282. The gain is almost proportional to the over-all length of the array and is almost independent of the element spacing provided this spacing does not exceed the critical value of $3\lambda/8$, at

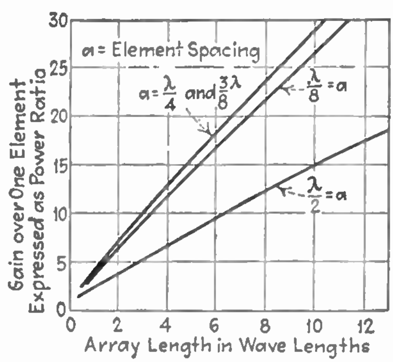


FIG. 282.—Gain of end-fire array as a function of array length for various element spacings. These curves are calculated on the assumption that the radiation from the individual antenna is proportional to the cosine of the angle of elevation.

which parasitic minor lobes of large amplitude develop. A comparison of Figs. 282 and 280 shows that the gain of broadside and end-fire arrays for the same over-all length is approximately the same.

Simple Co-linear Arrays.—The co-linear array consists of a series of antennas stacked coaxially one above the other, with the individual

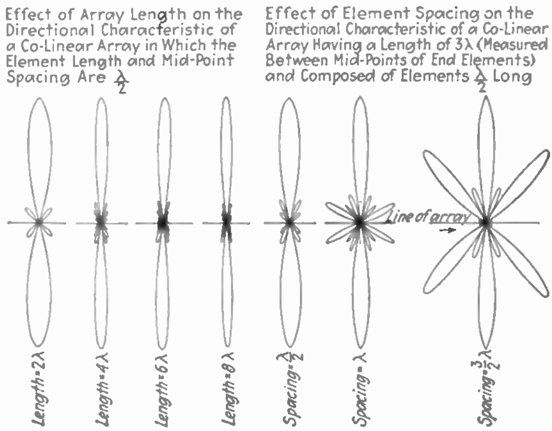


FIG. 283.—Effect of array length and element spacing on the directional characteristic of field radiated from a co-linear array in a plane containing the array.

antennas excited in the same phase. Stacking antennas in this way causes the radiation to be concentrated in a direction at right angles to the axis of the array (*i.e.*, gives vertical directivity) without affecting the radiation in a plane perpendicular to the axis of the array. The

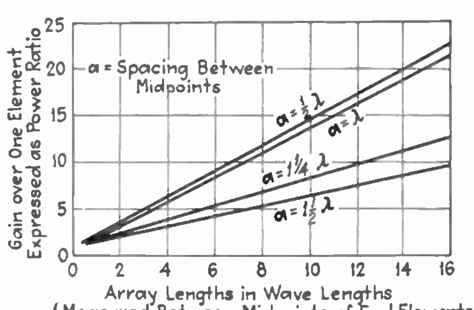


FIG. 284.—Gain of co-linear array as a function of array length for various element spacings. These curves are for the same elementary radiators as in the case of the broadside array.

directional characteristics and gain of co-linear arrays under different conditions are shown in Figs. 283 and 284 and are of the same general character as the corresponding properties of the broadside and end-fire arrays. Thus the sharpness of the beam along the horizon and the array gain are both determined primarily by the over-all length of the array as measured between centers of the end antennas and are substantially independent of the spacing of the array antennas provided this spacing does not exceed a critical value, which depends on the directional characteristics of the individual radiator and is never less than $3\lambda/4$. The gain obtained from stacking is however less than the gain obtained in broadside and end-fire arrays of equal length because of the fact that the antennas composing the co-linear array already possess a certain amount of directivity in a vertical plane, which has the effect of reducing the gain that can be obtained from stacking.

Array of Arrays.—The antenna arrays used in practice are nearly always combinations which can be considered as an array, each element of which is itself an antenna array. The directional characteristic of such an array of arrays is the group effect resulting from the combination

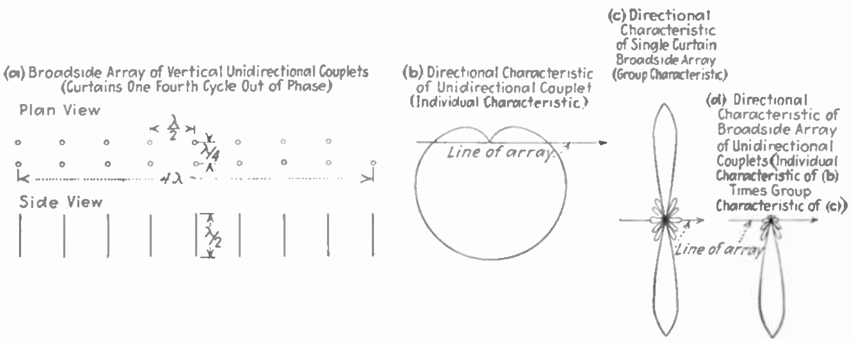


FIG. 285.—Broadside array, each element of which is composed of unidirectional couplets, together with polar diagrams showing individual and group characteristics, and the actual directional characteristic of the field radiated from the array of arrays.

of arrays, multiplied by the directional characteristic of the individual arrays. The directional characteristic of the individual elementary array can be calculated by Eq. (165), while the group effect depends upon the arrangement of the elementary arrays and is the directional characteristic of an array composed of elements located at the center of each individual elementary array of the array of arrays, and having uniform radiation in all directions.

The commonest example of an array of arrays consists of two parallel broadside arrays one-quarter wave length apart and excited in phase quadrature. Such an arrangement is illustrated in Fig. 285*a* and can be considered as a broadside array, each element of which consists of two radiators spaced one-quarter of a wave length apart and connected so as to be one-quarter of a cycle out of phase. The directional characteristic in a horizontal plane of this two-element elementary antenna array is unidirectional, as shown in Fig. 285*b*, while the directional characteristic of a simple broadside array is shown in Fig. 285*c*. The characteristic of the broadside array, each element of which consists of a unidirectional couplet, is obtained by multiplying the individual characteristic of Fig. 285*b* by the group directional characteristic of Fig. 285*c*, which results in the unidirectional broadside beam of Fig. 285*d*. The directional characteristic in a vertical plane is determined in a similar manner, but is not of great importance because the presence of the second broadside array does not materially affect the vertical radiation. Unidirectional characteristics in a horizontal plane, similar to those illustrated in Fig. 285, are obtained whenever the spacing between the two broadside arrays is an odd multiple of a quarter wave length, and the phase difference between the two arrays is a fraction of a cycle equal to the spacing in wave lengths.

Many short-wave antenna arrays consist of several arrays of broadside couplets stacked one above the other in order to give directivity in both vertical and horizontal planes. Such an antenna is illustrated in Fig. 286 and can be considered as consisting of a broadside array of

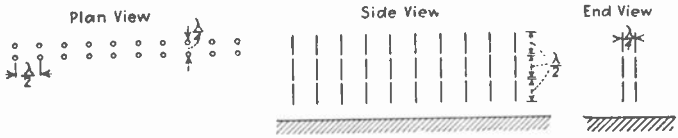


FIG. 286.—Typical antenna array consisting of a broadside array of stacked unidirectional couplets.

elementary antenna arrays, each element of which consists of an array of stacked unidirectional couplets. Each elementary array of such an antenna is in itself an array of arrays, and the directional characteristic is obtained by multiplying the directional characteristics of the unidirectional couplet of Fig. 285*b* by the directional characteristics of a stacked co-linear antenna and then again multiplying by the directional characteristic of a simple broadside array. It will be observed that the principal effect of the stacking is to increase the directivity in the vertical plane, that the effect of the second curtain is largely confined to eliminating the back-end radiation, and that the broadside arrangement gives high directivity in a horizontal plane. The result is a very intense radiation confined to a small cone directed along the horizontal and aimed broadside to the antenna array. The approximate directional characteristic of such an array in three-dimensional space is illustrated in Fig. 287.

The supporting towers of antenna arrays usually divide the main array into smaller arrays, or bays, between the ends of which there is a small spacing to accommodate the tower. Such an arrangement is essentially an array of arrays in which each bay represents an elementary array, and the group effect corresponds to a linear array of elementary antennas with a spacing equal to the distance between bay centers. If the spacing between the adjacent bays is at all appreciable, parasitic lobes of considerable size appear in the directional characteristic and seriously reduce the gain of the array.[1]

The gain of an antenna array composed of an array of arrays depends upon the individual and group directional characteristics and follows a very complicated mathematical law. It is possible however to give certain general principles by which the gain of such antenna systems can be readily estimated under many practical conditions. Thus an array of unidirectional couplets spaced $\lambda/4$ apart as in Fig. 285 has exactly twice as much gain as the corresponding single-curtain broadside array. If the spacing of the unidirectional couplets is some other odd multiple of $\lambda/4$ the gain from the second broadside curtain is approxi-

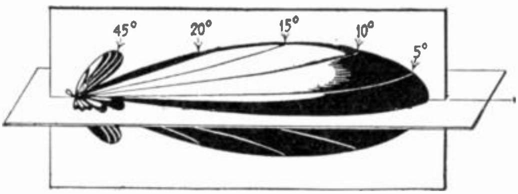


FIG. 287.—Directional characteristic in three-dimensional space of field radiated from a broadside array of stacked unidirectional couplets. (*From Southworth.*)

mately, but not exactly, two times. Again, when the elementary array tends to concentrate the radiation in one plane, while the group effect is to concentrate the radiation in another plane, the gain of the array of arrays is then very nearly equal to the gain of the elementary array multiplied by the gain of the group array. Thus when several broadside arrays of couplets are stacked the total gain is very nearly the gain obtained by stacking elements in a co-linear array, multiplied by two to take into account the effect of the second broadside array, and finally multiplied by the gain of a linear broadside array corresponding to the group array.

Arrays of Horizontally Polarized Elements.—While the above discussion has been applied specifically to arrays in which the array elements are vertical, the same principles apply to antenna arrays composed of horizontal elements. The only essential difference between the two is a 90° rotation in space. The horizontally polarized array thus gains directivity in a horizontal plane by arranging in-phase array elements co-linearly, obtains vertical directivity by placing horizontal radiators above one another and exciting them in the same phase, and is made unidirectional by means of a reflecting array placed behind the main array. A

[1] See G. C. Southworth, Certain Factors Affecting the Gain of Directive Antennas, *Proc. I.R.E.*, vol. 18, p. 1502, September, 1930.

typical example of an array composed of horizontal radiators is shown in Fig. 288.

The effect of the earth on a horizontally polarized array is exactly the same as in the case of any horizontal antenna and so is obtained by multiplying the directional characteristic in free space by the factor $\left|2 \sin\left(2\pi\frac{H}{\lambda} \sin \theta\right)\right|$. The earth causes cancellation of radiation along the horizontal, and if the main part of the radiation is to be directed at a low angle of elevation the center of the array must be at least one wave length high.

Agreement between Calculated and Observed Antenna Characteristics.—Experience has shown that the gain and the directional characteristics of antenna arrays calculated on the basis of a perfectly conducting earth agree satisfactorily with observed results under nearly all conditions. Comparisons of theoretical and observed directional characteristics for typical vertically and horizontally polarized antenna arrays[1] are given in Fig. 289 and show excellent agreement.

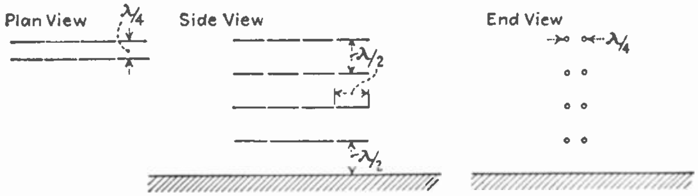


FIG. 288.—Views of a typical array composed of horizontally polarized antennas.

The imperfections of the earth have little effect upon the observed and calculated gain of an antenna array when the gain is defined in such a way as to make the effect of the earth substantially the same on both the array and the comparison antenna. The data on this point are relatively meager, however, but indicate in general that the calculated gains are realized approximately in practice. Thus Sterba cites a case in which the calculated and observed gains were 131.8 and 147.9, respectively, as compared with the individual antenna of the array.[2]

Radiation Resistance of Antenna Arrays.—The radiation resistance of an antenna array is rather difficult to calculate because the couplings between array elements cause energy to be radiated back and forth, with the result that the radiation resistance of each individual element in the array depends not only upon the characteristics of the individual

[1] See G. C. Southworth, *loc. cit.;* see also M. Bäumler, K. Krüger, H. Pendl, and W. Pfitzer, Radiation Measurements of a Short-wave Directive Antenna at the Nauen High-power Radio Station, *Proc. I.R.E.*, vol. 19, p. 812, May, 1931.

[2] See E. J. Sterba, Theoretical and Practical Aspects of Directional Transmitting Systems, *Proc. I.R.E.*, vol. 19, p. 1184, July, 1931.

element but also to an appreciable extent upon its position within the array. Methods have been devised[1] for calculating the radiation resistance, but these involve such complicated mathematical expressions as to be of little practical use. The radiation efficiency of short-wave antenna arrays is so high that it is permissible to assume that most of the energy delivered to the antenna system is radiated, with the result that the most important features are the amount of power that is avail-

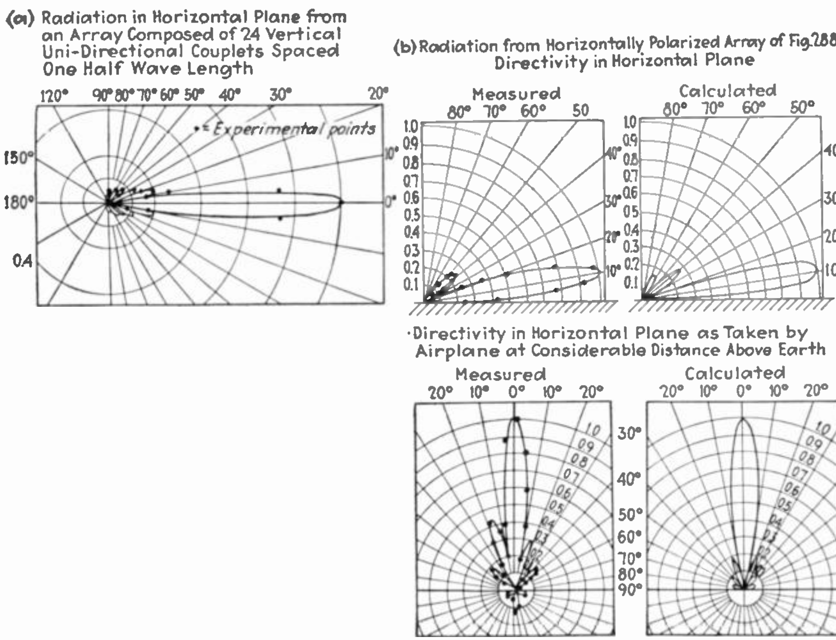


FIG. 289.—Comparison of measured and calculated distribution of radiated field about arrays composed of horizontal and vertical elements. The calculated results assume a perfect earth.

able and the direction in which the antenna radiates this power, rather than the radiation resistance itself.

122. Directive Antennas Employing Long Wires. *The Beverage or Wave Antenna for the Reception of Low-frequency (Long-wave) Signals.*— The Beverage or wave antenna consists of a wire ranging from one-half to several wave lengths long, pointed in the direction of the desired transmitting station, and mounted at a convenient height (usually 10 to 20 ft.) above earth. The end toward the transmitting station is grounded through an impedance approximating the characteristic impedances of the antenna when considered as a one-wire transmission line with ground return, while the energy abstracted from the radio waves is

[1] See A. A. Pistolpors, The Radiation Resistance of Beam Antennas, *Proc. I.R.E.*, vol. 17, p. 562, March, 1929.

delivered to a radio receiver at the end of the antenna farthest from the transmitter. An antenna of this type discriminates very strongly against waves which arrive from directions other than that in which the grounded end points, and in particular virtually eliminates signals arriving from angles which differ by more than 90° from the optimum direction, as is illustrated by the typical directional characteristic of a wave antenna shown in Fig. 291.

The wave antenna makes use of the tilt which is assumed by the wave front of vertically polarized waves in the vicinity of the earth as a result of the earth losses. This vertical tilt is caused by the energy which flows downward from the upper parts of the wave in order to supply the earth losses. For purposes of analysis the wave can be resolved into vertically and horizontally traveling components, of which the former cuts across the wave antenna and thereby induces on the wire a wave that immediately divides into equal parts that run toward the opposite ends of the wave antenna with a velocity approximating that of light. A radio wave arriving from the direction in which the wave antenna

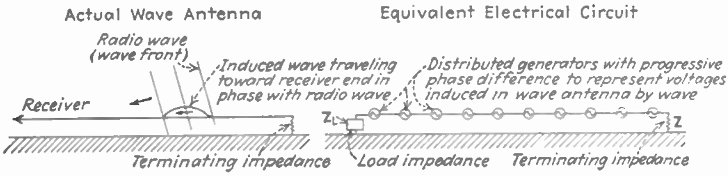


FIG. 290.—Details of action taking place in the Beverage wave antenna when the incident wave arrives from the direction in which the response is maximum.

points travels along the antenna at approximately the same speed as the induced wave traveling toward the receiver and is continually adding energy to this wave, with the result that a strong signal is delivered to a receiving set at the end of the wave antenna farthest from the transmitter. The part of the induced wave that travels along the wave antenna toward the grounded end is of no importance, being largely canceled out by the fact that it travels against the radio wave. Furthermore what small wave does arrive at the far end of the antenna is absorbed in the grounding resistance.

Consider now the response of the wave antenna to a transmitter located at right angles to its length. The radio waves from such a transmitter induce waves in the antenna which are everywhere in phase but, being at different distances from the receiver, reach the receiver more or less out of phase and largely cancel each other. Again, if the radio transmitter is located in the direction in which the receiver end of the wave antenna points, there will be little or no energy delivered to the receiver because the principal induced wave travels with the radio wave and is absorbed by the terminating resistance at the far end of the antenna.

The behavior of a Beverage antenna can be analyzed mathematically by considering it to be a long transmission line grounded at one end through an impedance Z and grounded at the other end through a load impedance. The action of the wave on the antenna can be represented by a series of uniformly distributed generators having a progressive change in phase of $(2\pi/\lambda) \cos \theta$ radians per unit length, where λ is the wave length of the radio wave, and θ is the bearing angle of the radio waves measured with respect to the direction in which the grounded end of the transmission line points. The voltage developed by these distributed generators per unit length is equal to the strength of the radio wave multiplied by the sine of the angle of tilt of the wave front away from the vertical. This basis for the mathematical analysis of the wave antenna is illustrated in Fig. 290.[1]

The directional characteristics of the wave antenna depend upon the length of the antenna, the velocity of phase propagation of the wave induced in the antenna, and the grounding impedance at the end pointing in the desired direction of reception. The length of the antenna in general determines the width of the zone of good reception, and for best results the antenna length should be not less than one-half wave length. The reception in the backward direction is determined partially by the antenna length and to a much greater extent by the grounding impedance of the far end. By making this grounding impedance differ slightly from the characteristic impedance it is possible to cancel out back-end reception completely for any desired direction and to reduce materially the back-end reception from other directions. The directivity of a wave antenna can be increased by combining several separate wave antennas into an array. Arranging the wave antennas parallel with each other and side by side narrows the lobe of maximum reception, while placing them coaxially one behind the other reduces the back-end reception. Typical directional characteristics of a single wave antenna and of an array consisting of four such antennas are shown in Fig. 291.

The operation of the wave antenna requires that there be earth losses because without these losses the wave front is vertical and the wave antenna picks up no energy. The wave antenna therefore abstracts maximum energy when located over poorly conducting earth and is least satisfactory when over moist earth or water.

The simple wave antenna that has been described can be modified in many details. Thus it is possible to use a two-wire line in which the two wires are effectively in parallel as far as the wave antenna is concerned

[1] The complete mathematical solution of the wave antenna is given by Harold H. Beverage, Chester W. Rice, and Edward W. Kellogg, The Wave Antenna, *Trans. A.I.E.E.*, vol. 42, p. 215, 1923; also see Austin Bailey, S. W. Dean, and W. T. Wintringham, The Receiving System for Long-wave Trans-atlantic Radio Telephony, *Proc. I.R.E.*, vol. 16, p. 1645, December, 1928.

but at the same time serve as a two-wire transmission line to transmit the energy delivered to the receiving end of the wave antenna back to the other end in order to permit the receiver and the grounding resistance to be located at the same place. It is also possible artificially to load the wave antenna with series or shunt condensers, or inductances. This makes the phase velocity of the induced wave greater or less than the velocity of light and thus alters the directional characteristics.

The wave antenna is the most satisfactory means that has been devised for the commercial reception of long-wave signals in point-to-point communication. Its directivity is sufficient to reduce materially static disturbances and interference from other stations, it is relatively inexpensive to construct because of the low height required, and as a result of its great length is able to abstract considerable energy from the radio wave. In addition to these advantages, the wave antenna is non-

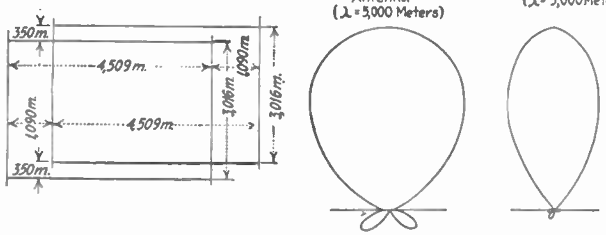


FIG. 291.—Directional characteristics expressed in terms of field strength of single Beverage wave antenna and of an array of such antennas.

resonant and so can be used to receive simultaneously from several transmitters operating on different frequencies, provided these stations all lie in the same direction.

The long-wave Beverage antenna is used only in the reception of radio signals and does not give satisfactory results in transmitting. This is because the wave antenna is inherently inefficient, requiring for its operation large energy losses in the adjacent soil. The effectiveness of the wave antenna results from its directivity and its great length, which is sometimes in excess of 5 miles.

Short-wave Beverage Antenna.[1]—The simple wave antenna that has been described is not satisfactory for the reception of high-frequency short-wave signals. If the antenna is only several wave lengths long the length in meters is small, and little energy is abstracted from the waves, while if the length is great (such as several hundred wave lengths) the various induced waves do not add up in phase. These considerations have caused the wave antenna to be modified as shown in Fig. 292*a* when

[1] See H. H. Beverage and H. O. Peterson, Diversity Receiving System of RCA Communications, Inc., for Radio Telegraphy, *Proc. I.R.E.*, vol. 19, p. 531, April, 1931

not RCA

H. err

applied to the reception of short-wave signals. Co-linear pairs of horizontally polarized antennas approximately one-half wave length long are loosely coupled to a two-wire transmission line, one end of which is terminated in an impedance approximating the characteristic impedance, while the other end goes to the radio receiver. The operation

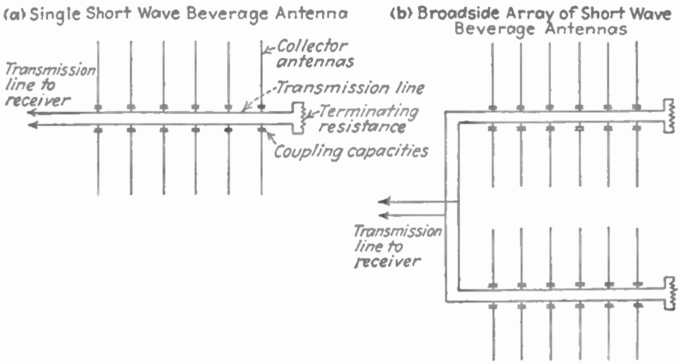


FIG. 292.—Plan view of single short-wave Beverage antenna and of an array consisting of two such antennas in broadside.

of this short-wave Beverage antenna is fundamentally the same as that of the long-wave antenna but differs in that the energy is collected by the horizontal antennas from the horizontally polarized component of the radio wave. The action thus does not depend upon the tilt of the wave front.

The short-wave Beverage antenna can be considered as an end-fire horizontally polarized antenna array, each element of which consists of two co-linear half-wave horizontal antennas, with the proper phase relations between the adjacent antenna elements maintained by the transmission line with its terminating impedance. The directional characteristics of the antenna in vertical and horizontal planes are therefore similar in all respects to those of ordinary horizontally polarized end-fire arrays and have already been discussed. A single Beverage short-wave antenna gives a rather broad beam in a horizontal plane, as do all end-fire arrays, but this can be sharpened by placing two Beverage antennas side by side as shown in Fig. 292*b*. The directional characteristic of a typical short-wave Beverage antenna is shown in Fig. 293.

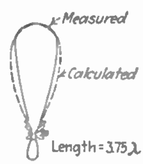


FIG. 293.—Directional characteristics in a horizontal plane of typical short-wave Beverage antenna.

The short-wave Beverage antenna has a number of advantages compared with other types of antenna arrays. First and most important it will give satisfactory reception of all signals lying within a frequency range extending 30 to 50 per cent above and below the frequency which makes the antenna elements a half wave length long. Thus the same antenna can be used for simultaneous reception of a number of signals of different frequencies or can receive signals of different frequencies at different times. The short-wave Beverage antenna furthermore has sufficient directivity for most practical purposes, is relatively inexpensive to construct, and is capable of abstracting large amounts of energy from the passing waves.

The horizontally polarized collecting elements of a short-wave Beverage antenna should have a length that is approximately one-half wave length at a frequency in the center of the band to be received. These collectors are then loosely coupled to the transmission line by means of small capacities usually supplied by the mounting insulators. The spacing between collectors as measured along the transmission line should not exceed $3\lambda/8$ for the highest frequency to be received on the antenna. The maximum allowable array length is determined by the velocity of phase propagation along the transmission line, which is always less than the velocity of light unless the loading effect of the coupled collector antennas is eliminated by series condensers or shunt inductances. In most cases the array is made three to five wave lengths long. When the collector elements are horizontally polarized the entire antenna array should be placed at least one-half wave length above ground in order to avoid the region where the ground reflections cancel the horizontal component of the wave. If the collector elements are vertically polarized however the height above ground is unimportant. The back-end reception is controlled by the terminating impedance and can be made very small, with a "blind" spot in any desired backward direction, by using proper termination.

The short-wave Beverage antenna was developed primarily for reception but can be used for transmission and has the great advantage of being only slightly resonant so that the same antenna can be operated over an appreciable frequency range without making any antenna adjustments.

Tilted-wire Antennas.[1]—The tilted-wire antenna is a modified wave antenna which is tilted with respect to the direction of the desired signals, as shown in Fig. 294*a*, so that a voltage is induced in the antenna wire without the necessity of depending upon the tilt of the wave front. The angle which the tilted wire makes with the wave front is very important, since if this angle approaches 90°, as shown in Fig. 294*b*, the waves induced in different parts of the tilted wire reach the receiver end at the same time as does the radio wave and add up in phase at the receiver, but the magnitude of the resulting voltage is small because the antenna intercepts only a small part of the passing wave. On the other hand if the tilt angle is small, as in Fig. 294*c*, the voltages induced in the wire are large,

[1] This section is based largely upon the discussion by E. Bruce, Developments in Short-wave Directive Antennas. *Proc. I.R.E.*, vol. 19, p. 1406, August, 1931.

Bell [illegible]

but the induced waves must travel farther to reach the receiver than does the radio wave because of the wire tilt, and the waves induced in various parts of the antenna arrive at the receiver more or less out of phase, thus tending to cancel each other. The tilt angle for which the energy delivered to the receiver is maximum is that which makes the

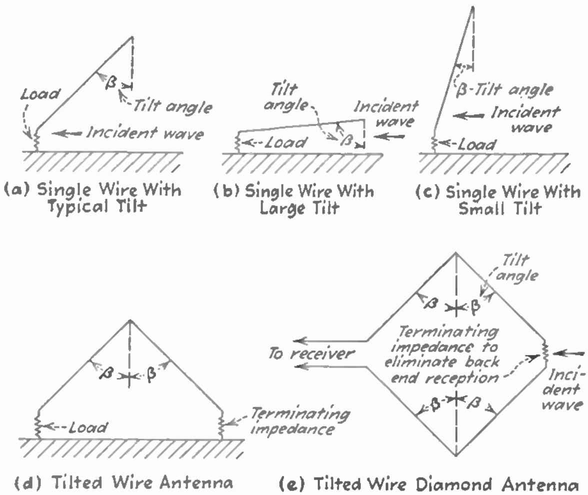


FIG. 294.—Examples of tilted-wire antennas.

projection of the tilted wire in the direction of wave propagation exactly one-half wave length less than the wire length, and is given in Fig. 295. The optimum tilt angle increases as the length of the tilted wire is increased and, when the wire is very long, changes very slowly with variations in the wave length of the antenna.

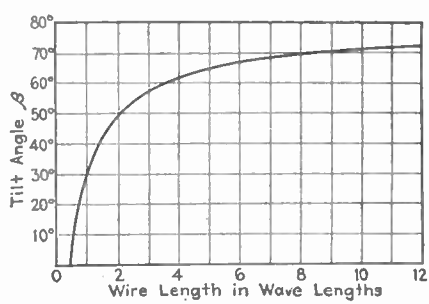


FIG. 295.—Optimum tilt angle of tilted-wire antennas.

As a consequence a long tilted wire will give satisfactory response over an appreciable range of frequencies.

It is possible to place two tilted-wire antennas end to end to form an inverted V, as shown in Fig. 294d. This arrangement requires no more supporting towers than the single vertically polarized tilted wire and makes the far end of the antenna accessible to the ground for purposes of termination. Furthermore the optimum tilt angle of the two legs of the V varies in opposite directions as changes in frequency cause the length of the legs expressed in wave lengths to vary, with the result that the V antenna gives substantially optimum response over a wide frequency range.

The most satisfactory form of tilted-wire antenna consists of two horizontally polarized V tilted wires arranged as shown in Fig. 294*e*. Such an antenna increases the directivity in a horizontal plane, avoids the necessity of obtaining a constant-resistance ground connection for the far-end termination, and can be supported on four wooden telephone poles of moderate height. The directivity is excellent, the gain compares favorably with that of much more complicated antenna arrays, and the antenna can be used for either transmission or reception at any frequency within a two-to-one range without any readjustments whatsoever and without losing its directive characteristics. In transmitting, the far-

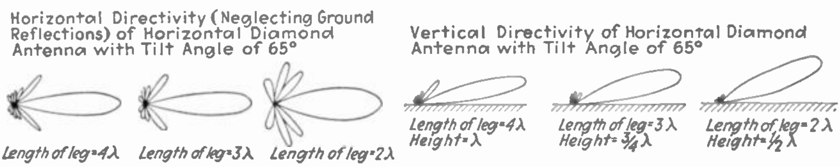


FIG. 296.—Polar diagrams showing direction characteristics expressed in terms of field strength of the same horizontal diamond-shaped antenna for three different frequencies.

end termination is required to dissipate the energy that would be otherwise radiated in a backward direction, and can conveniently consist of a long two-wire transmission line made of iron wires having high attenuation.

The directional characteristic of the horizontal diamond-shaped antenna is determined by the length of each leg, by the tilt angle, and by the height above ground, and increases as the antenna dimensions when measured in wave lengths increase. The angle of elevation at which the response in a vertical plane is maximum is determined by the height above earth, as is the case in all horizontal antennas. If this angle is to be kept small the diamond must be at least one wave length above ground. In order to obtain a true picture of the directivity it is necessary to view the solid diagram of the antenna because with a perfect ground there is no radiation whatsoever in a horizontal plane. The horizontal width of the beam as seen from a plan view can, however, be accurately determined by ignoring the effects of the earth and assuming that the antenna is in free space. The directional characteristics of diamond-shaped tilted-wire antennas having the optimum tilt angle given in Fig.

295 can be calculated by means of Eq. (163). Typical polar curves giving field strength patterns in vertical and horizontal planes are given in Fig. 296.

When the double V (diamond) or single V antennas have legs that are at least several wave lengths long and have a tilt angle that approximates the optimum value as given by Fig. 295, the directivity compares favorably with that of other types of antenna arrays and is maintained over a wide range of frequencies. This is illustrated by the curves of Fig. 296, which show the horizontal and vertical directional characteristics of the same horizontally polarized diamond antenna for three different frequencies. The only essential difference is that as the frequency is lowered, the antenna dimensions when measured in wave lengths are less, and the directivity is accordingly reduced.

In practical tilted-wire antennas it is desirable to make up each conductor of two wires connected in parallel and spaced a small distance apart. The characteristic impedance of the antenna can then be controlled by the spacing, which may be made such as to assist in matching the radio-frequency transmission line. Furthermore with an ordinary wire the characteristic impedance of different parts of the tilted wire would not be the same because the varying spacing between opposite legs as one moves along the diamond causes a reduction in directivity. With the spaced-wire conductor it is possible to compensate for this effect by varying the spacing between parallel conductors in such a way as to keep the characteristic impedance constant. This means a greater spacing as the wire recedes from ground, or in the case of the diamond, as the two sides separate.

The tilted-wire antenna, particularly the horizontal diamond form, has many advantages over ordinary antenna arrays. The cost is much less than for other antennas of equal directivity, the construction is extremely simple, and no delicate adjustments are required to place the antenna in operation. Satisfactory transmission and reception can be obtained over a frequency range of at least two to one without any change in adjustments, which permits simultaneous operation on a number of frequencies if this is desirable. It is probable that the highly satisfactory performance and low cost of the tilted-wire antenna will cause it to displace the more complicated antenna arrays under many conditions.

Long-wire Antenna Arrays.[1]—The long-wire antenna array is an ingenious combination of long-wire antennas arranged to give a sharply defined beam in the desired direction with a relatively simple antenna structure. A long single wire remote from earth and of a length that is an exact multiple of one-half wave length tends to concentrate the

[1] The material in this section is based upon the paper of P. S. Carter, C. W. Hansell, and N. E. Lindenblad, Development of Directive Transmitting Antennas, RCA Communications, Inc., *Proc. I.R.E.*, vol. 19, p. 1773, October, 1931.

radiation along two conical or funnel-shaped surfaces which are coaxial with the wire, as has been explained in Sec. 120.

Two parallel spaced long-wire antennas excited in opposite phase cancel each other's radiation in a direction at right angles to the plane containing the wires but do not cancel in the plane of the wires because of the spacing between the radiators in that direction. The radiation from the two-wire combination is therefore concentrated in four ears which have their maximum length in the plane of the wires. Two of these ears can be eliminated by staggering the parallel wires so that the perpendicular to the line joining the wire extremities is in the direction of the undesired radiation, as shown in Fig. 297*b*. This cancellation is obtained because the fields from corresponding parts of the two wires so staggered are in phase opposition in the direction of the undesired lobes. At the same time, with a spacing between wires such that the distances to corresponding parts of the two wires in a direction parallel to the remaining two lobes differ by one-half wave length, as shown in

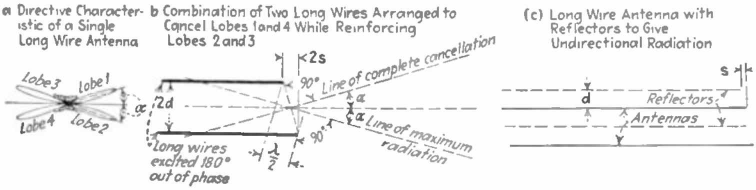


FIG. 297.—Long-wire antenna array showing how the long wires are combined to give a sharply defined beam.

Fig. 297*b*, the radiation in the direction of these lobes is very intense. The back-end radiation can now be eliminated by adding a second pair of long-wire antennas, as shown in Fig. 297*c* and exciting these with currents that lead the currents in the corresponding parts of the first pair of ears by 90°. The result is a cancellation of the back ear for exactly the same reason that adding a second curtain to a broadside array makes the characteristic unidirectional.

An analysis of the geometry of the long-wire antenna shows that the distance d between adjacent wires (see Fig. 297*c*) is given by the formula

$$\text{Distance } d \text{ between wires} = \frac{\lambda}{4} \frac{\cos \alpha}{\sin 2\alpha} \tag{158}$$

where

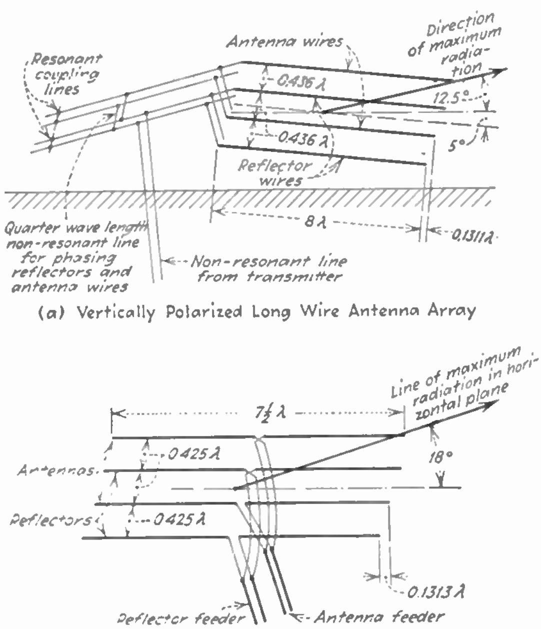
λ = wave length of transmitted wave

α = angle of maximum radiation with respect to the wire axis as given by Fig. 275.

The stagger distance s (see Fig. 297*c*) between adjacent wires is similarly obtained by the equation

$$\text{Stagger distances } s = \frac{\lambda}{4} \frac{\sin \alpha}{\sin 2\alpha} \tag{159}$$

The long-wire antenna array of Fig. 297*c* will be horizontally or vertically polarized, depending on whether the plane of the wires is horizontal or vertical, respectively. When the vertical configuration is employed it is customary to tilt the wires with respect to the earth's surface, as shown in Fig. 298*a*, in order to reduce the angle which the direction of maximum radiation makes with the horizontal. The height of the low end of the antenna with respect to the ground is relatively unimportant inasmuch as this array produces negligible radiation in the direction of the earth. The characteristics of the horizontally polarized



(a) Vertically Polarized Long Wire Antenna Array

(b) Horizontally Polarized Long Wire Antenna Array (Plan View)

FIG. 298.—Examples of commercial long-wire antenna arrays, together with feeder circuits for the vertically polarized array. (Feeder arrangement for horizontal array is similar.)

long-wire array depend to a considerable extent upon the height above earth, which is the factor determining the angle of elevation at which the radiation is concentrated.

The directional characteristics of long-wire antenna arrays can be analyzed by considering that these antennas are end-fire arrays composed of four antennas arranged as shown in Fig. 297*c*, with the individual element having the directional characteristic of a single long-wire radiator. The spacing between adjacent elements of the end-fire array is the distance between the centers of the long wires and is $\lambda/(4 \sin 2\alpha)$, while the phase difference between adjacent wires is 90°. The effect of the earth in the case of the horizontally polarized array can be taken into account by the use of Fig. 276, while the earth can be ignored in the vertical array.

The long-wire antenna array is inexpensive to construct, is readily adjusted, and at the same time has excellent directive characteristics. Thus the gain of a single vertically polarized array composed of wires eight wave lengths long is approximately sixteen times over a standard half-wave comparison antenna, while with a broadside of two such arrays spaced two wave lengths apart and excited in phase the gain figured on the same basis is approximately thirty-five times. These values compare favorably with those obtained from much more complicated antenna arrays.

Partially Folded Long-wire Antenna.[1]—This type of long-wire antenna is illustrated in Fig. 299 and consists of a long wire folded at the middle to form a V and excited with a voltage that is applied in series with the wire at the point of folding. The angle between the two legs of the V is twice the angle of maximum radiation from a long-wire antenna having

"Vee"

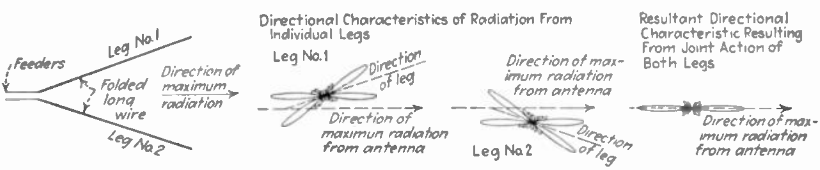


FIG. 299.—Partially folded-wire antenna, showing how the radiation from the two legs combines to give a well-defined beam.

a length equal to that of one leg of the V. Each leg of the folded V antenna has a directional characteristic such as shown in Fig. 274, but when the radiation from the two sides of the V is combined the result is an addition along the center line of the fold, and a partial cancellation in all other directions, giving the result illustrated in Fig. 299, which has excellent directivity in the plane of the V (the horizontal plane). The directivity in the plane at right angles to the plane of the V (the vertical plane), can be improved by stacking two folded radiators approximately one-half wave length apart as shown in Fig. 300*a* and excited in phase with each other. The back-end radiation is eliminated by placing a second pair of folded radiators an odd number of quarter wave lengths back of the first pair and making the phase of this second pair of folded wires differ from the phase of the first pair by a fraction of a cycle equal to the spacing measured in wave lengths. The directivity in the plane of the V can be increased if necessary by placing radiators in broadside to form a W, as shown in Fig. 300*b* and exciting them in phase.

[1] See P. S. Carter, C. W. Hansell, and L. E. Lindenblad, *loc. cit.*

The calculation of the complete directional characteristic of a folded-wire radiator is complicated by the fact that the two legs of the folded wire are at an angle with respect to each other, which causes the wave radiated in certain directions to be elliptically polarized. The radiation of a simple folded-wire antenna in the plane of the antenna, and at right angles to this plane, is horizontally polarized however and can be calculated by Eq. (164). The directional characteristic of an array of folded wires, such as shown in Fig. 300, can then be calculated by the methods that have been outlined for arrays. The directivity is determined largely by the length of the folded wires and the height above earth, assuming that this length and height are measured in wave lengths.

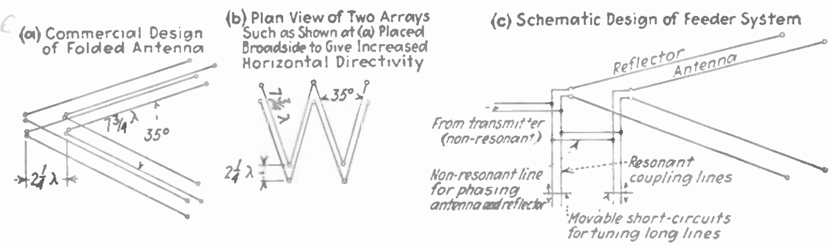


FIG. 300.—Commercial design of partially folded long-wire antenna array, together with feeder circuits.

The long-wire folded-antenna array possesses the same advantages of the long-wire antenna array but to an even greater degree. The construction is extremely simple, the necessary phase adjustments can be readily made, and the gain is excellent. The gain of a single bay of the type illustrated in Fig. 300*a* and composed of wires eight wave lengths long at an average height above earth of one wave length is approximately thirty-nine times over a standard half-wave comparison antenna, while the use of additional bays gives a proportionate increase in gain (*i.e.*, two bays give a gain of 78, three a gain of 117, etc.).

123. Antenna Formulas.—1. Radiation from an elementary wire antenna: in absence of ground

$$\epsilon = \frac{60\pi}{d\lambda} I(\delta l) \cos \theta \qquad (160)$$

where

ϵ = field strength in volts per meter
d = distance to antenna in meters
θ = angle of elevation with respect to plane perpendicular to antenna wire
λ = wave length of radiated wave in meters
δl = length of elementary antenna in meters
I = current in antenna.

2. Radiation from a wire in space.[1]

a. When the wire is an odd number of half wave lengths long:

$$\epsilon = \frac{60I}{d} \frac{\cos\left(\pi \frac{L}{\lambda} \cos \theta\right)}{\sin \theta} \tag{161a}$$

b. When the wire is an even number of half wave lengths long:

$$\epsilon = \frac{60I}{d} \frac{\sin\left(\pi \frac{L}{\lambda} \cos \theta\right)}{\sin \theta} \tag{161b}$$

where

ϵ = field strength in volts per meter
d = distance to antenna in meters
I = current in amperes at a current loop
L = length of antenna in meters
λ = wave length in meters
θ = angle of elevation measured with respect to wire axis.

3. Radiation from vertical grounded wire:[2]

$$\epsilon = \frac{60I}{d}\left[\frac{\cos 2\pi\frac{L}{\lambda} - \cos\left(2\pi\frac{L}{\lambda}\sin\theta\right)}{\cos\theta}\right] \tag{162}$$

where θ is the angle of elevation with respect to ground, and the remainder of the notation is as in Eq. (161).

4. Radiation from a tilted-wire diamond antenna in space when legs are an exact multiple of a half wave length long:[3]

a. Radiation in the plane of antenna (ground reflection neglected): Relative field strength =

$$\left[\frac{1 + \cos \Phi}{\cos^2 \beta - \sin^2 \Phi}\right] \cos\left\{\pi\frac{l}{\lambda}\sin(\Phi + \beta) + n\frac{\pi}{2}\right\} \cos\left\{\pi\frac{l}{\lambda}\sin(\beta - \Phi) + n\frac{\pi}{2}\right\} \tag{163a}$$

where

Φ = bearing angle with respect to line passing through the feeder and terminating impedance apexes of the diamond
β = tilt angle of antenna
l = length of leg of antenna
λ = wave length
n = factor that is zero when l is an odd number of half wave lengths and is unity when l is an even number of half wave lengths.

[1] See Carter, Hansell, and Lindenblad, *loc. cit.*

[2] See Stuart Ballantine, On the Radiation Resistance of a Simple Vertical Antenna at Wave Lengths below the Fundamental, *Proc. I.R.E.*, vol. 12, p. 823, December, 1924.

[3] See E. Bruce, *loc. cit.*

b. Radiation in plane perpendicular to plane of antenna and parallel to the direction of maximum radiation (antenna remote from ground):

$$\text{Relative field strength} = \left|\left[\frac{1 + \cos\theta}{1 - \sin^2\beta\cos^2\theta}\right]\cos^2\left(\pi\frac{l}{\lambda}\sin\beta\cos\theta + n\frac{\pi}{2}\right)\right| \quad (163b)$$

where θ is the angle of elevation with respect to plane of antenna, and the remainder of notation as in Eq. (163*a*). Equation (163*b*) assumes that the terminating impedance is the characteristic impedance multiplied by the sine of the tilt angle, and that the leg length is an exact multiple of $\lambda/2$.

5. Radiation from a single partially folded antenna remote from ground when the length of each leg is an even multiple of a half wave length:[1]

Field strength in plane of fold =

$$\sqrt{E_a^2 + E_b^2 - 2E_aE_b\cos\left(2\pi\frac{l}{\lambda}\sin\alpha\sin\Phi\right)} \quad (164a)$$

where

E_a and E_b = radiation in desired direction from individual legs of folded antenna as given by Eq. (161)
l = length of leg of fold
λ = length corresponding to one wave length
α = half of angle of fold
Φ = angle with respect to bisector of fold.

Radiation in vertical plane passing through bisector of fold =

$$\frac{120I}{d}\left[\frac{\sin\left(\frac{n\pi}{2}\cos\alpha\cos\theta\right)\sin\alpha}{1 - \cos^2\theta\cos^2\alpha}\right] \quad (164b)$$

where

n = number of half wave lengths in each leg of antenna
α = half of angle of fold
θ = angle of elevation with respect to plane of fold
I = current at current loop
d = distance to antenna in meters.

The radiation from an array of folded wires, such as illustrated in Fig. 300, can be calculated on the basis of an array composed of individual antennas having directional characteristics of Eqs. (164*a*) and (164*b*).

6. Radiation from an antenna array consisting of $\mathfrak{N}$ parallel planes each made up of N parallel columns where each column is made up of n individual radiating elements radiating uniformly in all directions:[2]

[1] See Carter, Hansell, and Lindenblad, *loc. cit.*

[2] See G. C. Southworth, Certain Factors Affecting the Gain of Directive Antennas, *Proc. I.R.E.*, vol. 18, p. 1502, September, 1930.

$$\text{Relative field strength} = \left| \frac{\sin n\pi(a \cos \Phi \cos \theta + b)}{n \sin \pi(a \cos \Phi \cos \theta + b)} \frac{\sin N\pi(A \sin \Phi \cos \theta + B)}{N \sin \pi(A \sin \Phi \cos \theta + B)} \frac{\sin \mathfrak{N}\, \pi(\mathfrak{A} \sin \theta + \mathfrak{B})}{\mathfrak{N} \sin \pi(\mathfrak{A} \sin \theta + \mathfrak{B})} \right| \tag{165}$$

where

n, N, and $\mathfrak{N}$ = number of radiators along x-, y-, and z-axes, respectively

a, A, and $\mathfrak{A}$ = spacing of adjacent radiators along x-, y-, and z-axes, respectively, measured in fractions of a wave length

b, B, and $\mathfrak{B}$ = phase displacement between adjacent radiators along x-, y-, and z-axes, respectively, measured in fractions of a cycle

θ = angle with respect to xy plane (angle of elevation)

Φ = angle with respect to xz plane (bearing angle).

7. Radiation from an antenna array in which the individual antennas are not spherical radiators is obtained by calculating the directional characteristic for spherical radiators, using Eq. (165) and then multiplying the result by the actual directional characteristic of the individual antenna involved.

8. Radiation from an array of arrays is determined by calculating the individual directional characteristic of the elementary array as outlined in Eq. (165) above, and then multiplying this individual effect by the group effect calculated by determining the directional characteristic of an array consisting of spherical radiators located at the centers of the elementary arrays.

9. Effect of ground—horizontal antennas. The effect of a perfect ground is obtained by multiplying the directional characteristic of the antenna in free space by the factor $\left|2 \sin\left(2\pi\frac{H}{\lambda} \sin \theta\right)\right|$, where $\frac{H}{\lambda}$ = height of horizontal antenna above earth measured in wave lengths, and θ is the angle of elevation with respect to earth.

10. Effect of ground—vertical antennas. When the vertical antenna is even number of half wave lengths long, the effect of a perfect earth is obtained by multiplying the directional characteristic of the antenna in free space by the factor $\left|2 \sin\left(2\pi\frac{H}{\lambda} \sin \theta\right)\right|$, where $\frac{H}{\lambda}$ is now defined as the height of the center of the antenna above earth, as measured in wave lengths.

When the vertical antenna is an odd number of half wave lengths long the multiplying factor that takes into account the presence of a perfect earth is $\left|2 \cos\left(2\pi\frac{H}{\lambda} \sin \theta\right)\right|$.

11. Formulas for calculating the gain of antennas and antenna arrays under certain conditions are given by G. C. Southworth, E. J. Sterba, and P. S. Carter, C. W. Hansell, and N. E. Lindenblad in the references already cited. These formulas are too complicated and too specialized to warrant inclusion in a work of this type, and there is unfortunately no simple general equation available for calculating array gain.

12. The effective height of an antenna for radiation in any given direction can be determined by dividing the equation of field strength radiated in the desired direction by $(60\pi I/d\lambda)$. This is equivalent to equating Eq. (160) for $\theta = 0$ to the field strength of the radiator in question, and solving for (δl).

13. Formulas for loop antennas are given in Chap. XVI.

124. Wave Reflectors and Directors. *Parasitic Antennas.*—An important method of obtaining directivity consists in placing parasitic antennas in the vicinity of the actual transmitting or receiving antenna, as the case may be, and tuning these parasitic antennas so that the energy which they pick up from the passing radio waves is reradiated in the proper phase to produce a marked directional action.

The fundamental principles involved in the production of directional effects by parasitic antennas can be understood by considering a simple example, such as shown in Fig. 301, involving a single transmitting and a single receiving antenna. The waves radiated from the transmitting antenna induce a voltage in the parasitic antenna and cause a current to flow. The current causes the parasitic antenna to act as a secondary radiator. The phase of this secondary radiation relative to the phase of the passing wave depends upon the tuning of the parasitic antenna. In particular if the secondary radiator offers an inductive reactance to the induced voltage (secondary radiator tuned to a lower frequency) the reradiated energy will be in such a phase as to neutralize more or less the resultant field in the direction marked "receiving antenna" in Fig. 301. On the other hand when the secondary antenna offers a capacitive reactance to the induced voltage (secondary radiator tuned to a higher frequency) the reradiation is in such a phase as to cause the resultant field to converge in the direction of the secondary radiator. The secondary or parasitic antenna thus acts as a "wave reflector" when tuned to a lower frequency than that of the wave, and as a "wave director" when tuned to a higher frequency. These effects are well illustrated by the experimental curve of Fig. 301 showing current in receiver output as a function of the wave length of the parasitic antenna relative to the transmitter frequency.

When the parasitic antenna is close to the transmitting antenna the induction field will induce an appreciable voltage that is 90° out of phase with the voltage induced by the radiation field. The effect of this is to cause the tuning adjustment of the parasitic antenna required to give a particular type of directivity to depend somewhat upon the distance to the transmitting antenna.

Antenna arrays of the type illustrated in Fig. 309, which involve two curtains of conductors, are frequently operated with the rear curtain functioning as a parasitic antenna so adjusted as to cause cancellation of radiation in the backward direction. This arrangement avoids the necessity of providing means for exciting the rear (or reflecting) curtain in the proper phase to give a unidirectional characteristic and yet gives results that are substantially the same as though the rear curtain was excited directly from the transmitter instead of parasitically from the front curtain. In order to obtain the most satisfactory results the parasitic reflector must be spaced the proper distance from the radiating curtain and must be tuned so that its radiation is in the proper phase. Under practical conditions the best spacing between curtains is approximately, but not exactly, a quarter of a wave length.[1]

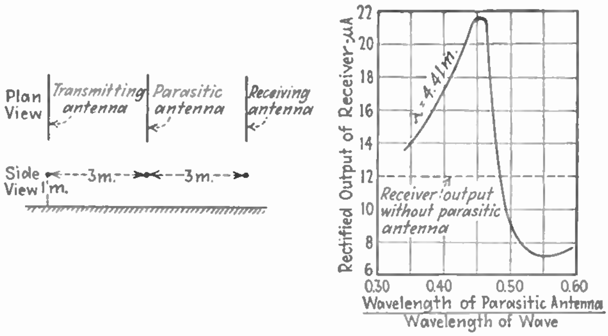


FIG. 301.—Simple arrangement for illustrating action of wave reflectors and directors, together with experimental curve showing effect of tuning the parasitic antenna.

A simple reflecting system consisting of three parasitic antennas arranged as shown in Fig. 302 is also sometimes employed to obtain a certain amount of directivity with simple and otherwise non-directional transmitting antennas. By properly adjusting the lengths of the parasitic antennas most of the radiation is concentrated in the direction of the open side. By combining such a reflecting system with a number of wave directors it is possible to produce a beam having very great directivity in both vertical and horizontal planes.[2]

Reflectors.—Another method of obtaining directivity is to surround the antenna with a reflector, which in the case of transmission concen-

[1] Raymond M. Wilmotte and J. S. McPetrie, A Theoretical Investigation of the Phase Relations in Beam Systems, *Jour. I.E.E.* (*London*), vol. 66, p. 949, September, 1928.

[2] See Hidetsugu Yagi, Beam Transmission of Ultra-short Waves, *Proc. I.R.E.*, vol. 16, p. 715, June, 1928. This paper is a classic upon the subject.

trates the radiation in the desired direction in much the same manner as a mirror reflects light, and in the case of reception gathers in the energy from a considerable portion of the passing wave and focuses it upon the receiving antenna. The principles involved in the reflection of radio waves can be understood by considering a simple example, such as illustrated in Fig. 303, in which a radio wave is assumed to impinge upon a perfectly conducting wall. The effect of this wall is to reflect the arriving wave backward with a reversal of the electrostatic field but with no change in the direction of the magnetic flux. The reflected wave is of exactly the same amplitude as the incident wave, and the angle of reflection is equal to the angle of incidence. In other words the entire phenomenon of the reflection of a radio wave by a perfect conductor is strictly analogous to the reflection of light waves by a mirror.

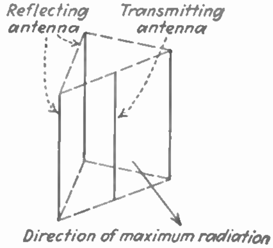


FIG. 302.—Antenna provided with simple arrangement of three parasitic reflecting antennas for concentrating the radiation.

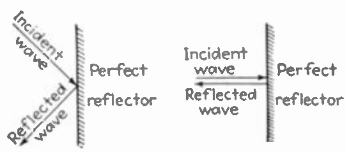


FIG. 303.—Relation of incident and reflected waves in the case of a perfect reflector. The two waves are of equal magnitude and opposite phase at the reflector, and the angle of incidence equals the angle of reflection.

The mechanism of reflection is as follows. The incident wave in striking the reflector induces currents which reradiate a wave that can be considered as the reflected wave. The magnitude and phase of the reflected wave must be such that at the surface of the reflector the vector sum of electrostatic fields of the incident and reflected waves is exactly equal to the voltage drop developed in the reflector by the induced currents. In the case where the reflector is a perfect conductor the voltage drop in the reflector is zero, and the incident and reflected waves are of the same amplitude but the electrostatic flux is reversed. When the reflector is a good, but not perfect, conductor the result is similar, but the reflected wave is smaller than the incident wave as a result of the voltage drop in the conductor. If the reflecting surface is a combined dielectric and imperfect conductor, such as is the surface of the earth, the reflection is not only incomplete (*i.e.*, the reflected wave is not only smaller than the incident wave) but also has the phase of the electrostatic flux shifted by an angle other than 180°.[1]

[1] For further information on the reflection of radio waves, see G. W. Pierce, "Electric Oscillations and Electric Waves," McGraw-Hill Book Company, Inc.,

Directional antennas making use of metallic reflectors usually employ a paraboloid of revolution, as shown in Fig. 304, or a cylindrical surface having a parabolic cross section, as also illustrated in Fig. 304. In either case the antenna is placed at the focus.[1] The reflectors are made of copper, which is near enough a perfect conductor to give substantially complete reflection. Directive antenna systems of these types are particularly well suited to transmission at ultra-high frequencies, since the dimensions of the required reflector are then within reason. At lower radio frequencies the size of reflector required is so great as to make it preferable to use other types of directive systems.

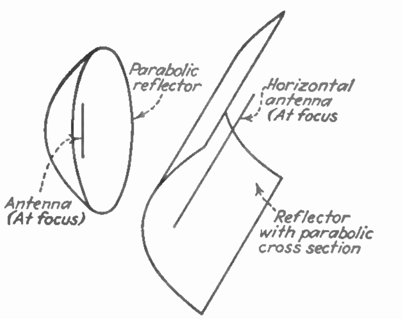


FIG. 304.—Examples of parabolic reflectors for producing directive characteristics.

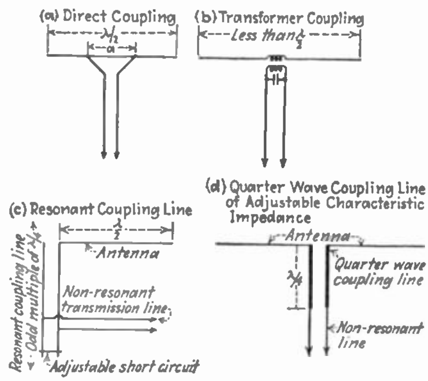


FIG. 305.—Typical methods of coupling an antenna to a non-resonant transmission line.

125. Radio-frequency Transmission Lines.—The output of a radio transmitter is ordinarily delivered to the antenna by means of a two-wire transmission line. This makes the antenna more or less independent of the location of the building housing the transmitter, facilitates the placing of several transmitters in the same building, and enables the same transmitter to excite different antennas at different times. The loss of energy in transmission to distances as great as ½ mile is relatively small provided the transmission line is constructed in such a manner as to radiate no energy, and resonances are avoided by terminating the antenna end of the line in a load impedance equal to the characteristic impedance as explained in Sec. 15.

The usual radio-frequency transmission line consists of a two-wire open-air line symmetrically arranged with respect to the ground. Radia-

New York, 1920; and P. O. Pedersen, "The Propagation of Radio Waves," G. E. C. Gad, Copenhagen, Denmark (in English).

[1] For examples of such reflectors see Franklin, Short-wave Directional Wireless Telegraphy, *Jour., I.E.E.* (London), vol. 60, p. 930, August, 1922; A. Meissner, Directional Radiation with Horizontal Antennas, *Proc. I.R.E.*, vol. 15, p. 928, November, 1927.

tion is avoided by placing the two wires close to each other and by avoiding capacity unbalances to ground. The characteristic impedance of a two-wire line at high frequencies is very nearly a resistance of $\sqrt{L/C}$ ohms, when L is the line inductance in henrys, and C is the line capacity in farads, per unit length. The characteristic impedance of a two-wire line is given with good accuracy by the formula

$$\text{Characteristic impedance} = 276 \log_{10} \frac{2D}{d} \text{ ohms} \qquad (166)$$

where D is the spacing between wire centers, d is the wire diameter, and D and d are in the same units. Since a tuned antenna offers a resistance load the antenna automatically supplies a load of the correct power factor, and the only problem of termination is that of obtaining the proper magnitude of load impedance.

Coupling the Antenna to the Transmission Line.—The commonly used methods of coupling the antenna to the transmission line are shown in Fig. 305. At Fig. 305*a* the proper coupling is obtained by connecting the two ends of the transmission line symmetrically to the half-wave antenna at suitable distances from the center. At Fig. 305*b* a tuned transformer is used to control the load impedance which the antenna couples into the transmission line. At Fig. 305*c* the antenna is connected to a short resonant transmission line an odd quarter wave length long, and the non-resonant transmission line is then connected to the proper point on the resonant line to give the required load impedance. In these arrangements the short resonant line acts as an impedance transformer in which the radio of impedance transformation depends on the point at which the non-resonant line is connected. When the load impedance which the antenna offers differs by only a moderate amount from the characteristic impedance of the line it is possible to transform the load impedance offered by the antenna to the characteristic impedance of the line by connecting the two with a quarter wave-length non-resonant transmission line having a characteristic impedance of $\sqrt{Z_r Z_o}$, where Z_r is the load impedance offered by the antenna, and Z_o is the characteristic impedance of the transmission line which goes to the transmitter. This method of coupling an antenna to a non-resonant transmission line is shown at Fig. 305*d*.

The adjustment of the impedance match between the transmission line and the antenna is most satisfactorily carried out by cut-and-try methods. The usual procedure is to connect one terminal of a thermocouple meter to the transmission line by a sliding contact, leaving the other end coupled to the line by stray capacities, and then adjusting the terminating impedance until the meter reading remains constant as the meter is slid along the line. This adjustment is independent of the length of the line and depends only upon the characteristic impedance.

Transmitters located in close proximity to the antenna can be conveniently coupled to the antenna by use of resonant transmission lines, such as illustrated in Fig. 306. These arrangements are very popular where simple antenna systems are employed, and have the great advantage that the entire transmission line and radiating system can be adjusted at the transmitter. The arrangement shown at Fig. 306*a* is known as a current-feed system and makes use of a transmission line that has a length approximately an integral multiple of a half wave length. Maximum radiation is obtained when the current at the transmitter end of the transmission line is maximum. The arrangement shown at Fig. 306*b* is commonly called a "Zeppelin" antenna and consists of a transmission line an odd number of quarter wave lengths long, one wire of which is connected to the end of the antenna while the other wire is capacitively coupled to the antenna. A modification of the Zeppelin antenna is shown at Fig. 306*c*. Maximum radiation occurs when the current at the transmitter end of the line is maximum.

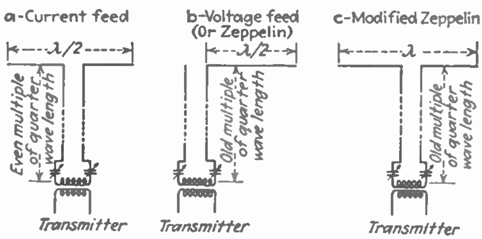


FIG. 306.—Representative types of resonant transmission lines for delivering energy to antennas.

Resonant transmission lines are usually adjusted by making their length slightly greater than the required value and then placing series condensers in each side of the line, as shown in Fig. 306. The reactance of these condensers has the effect of shortening the equivalent length of the line to a value less than the physical length and also neutralizes the reactance of the coil that couples the transmission line to the radio transmitter.

Resonant transmission lines have the disadvantage that the average power factor of the line is low, which makes the transmission-line losses high in proportion to the length. If the power is large, the resonant rise of voltage in such a line is also great enough to introduce problems of insulation.

Transmission Lines Used with Receiving Antennas.—Transmission lines are used in reception to connect directive receiving antennas with the radio receiver. Extreme care must be taken with receiving lines, however, in order to prevent the line from acting as an antenna and abstracting energy from passing radio waves. If this happens the main objects of the directional receiving antenna array will be lost, since the antenna effect of the transmission line will not be directional and disturbing waves and noises arriving from all directions will be picked up. Transmission lines commonly used to connect antennas with radio receivers are shown in Fig. 307. The open-air untransposed line is not entirely satisfactory unless it is enclosed in a metal pipe that serves as a shield. The buried concentric-pipe transmission line is practically immune from interfering signals and weather conditions but is rather expensive. The four-wire line with diagonally opposed wires in parallel is both simple and effective. Such a line is balanced against extraneous disturbances and can be run long distances without trouble from undesired pick-up.

Resonant transmission lines are seldom employed in reception because they must be made with extremely low losses if the antenna is located some distance from the receiver, while if the antenna can be located close to the receiver the receiving antenna usually consists of a long wire run direct to the receiver.

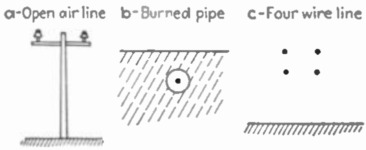


FIG. 307.—Types of transmission lines employed with receiving antennas.

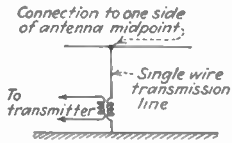


FIG. 308.—Single-wire transmission line.

Miscellaneous Comments.—In some cases a single-wire transmission-line arrangement such as shown in Fig. 308 is employed.[1] Resonances can be avoided in this line by connecting the wire at a suitable distance to one side of the mid-point of the antenna. The single-wire transmission line is not so efficient as other types, however, since the return path is supplied by the earth, and this causes the losses to be high.

In the construction of radio-frequency transmission lines care must be taken to avoid irregularities at corners, and at mounting insulators. The transmission line should also approach the antenna at right angles so that there will be a minimum of coupling between antenna and line.

In a non-resonant line there is a uniform progressive advance in phase as one goes along the line away from the transmitter end. This amounts to 2π radians for each wave length and can be used to give different parts of an antenna array any desired phase difference.

126. Typical Short-wave Transmitting Antennas.—When it is desired to avoid marked directional effects it is customary to use one of the arrangements shown in Figs. 305 or 306, with an antenna one-half wave length long and connected to the transmitter by means of a transmission

[1] See W. L. Everitt and J. F. Byrne, Single-wire Transmission Lines for Short-Wave Antennas, *Proc. I.R.E.* vol. 17, p. 1840, October, 1929.

line. Resonant lines are preferred when the antenna is within a few hundred feet of the transmitter; otherwise it is better to use non-resonant lines because of the greater transmission-line efficiency.

The antenna may be orientated either vertically or horizontally, depending whether vertically or horizontally polarized radiation is desired. In the latter case the height above ground determines the angle of elevation of maximum radiation, which is nearer the horizontal as the height is increased, as indicated by Fig. 277.

The procedure to be followed in tuning such an antenna system depends upon whether a resonant or non-resonant feeder system is employed. With a resonant feeder the antenna wire is cut to a length approximately 5 per cent less than a half wave length of the radio wave in space,[1] and the feeder wires are cut to have a length equal to or slightly greater than the proper multiple of a quarter wave length based on a radio wave in space. After erection the antenna and feeder system are coupled to the transmitter by means of an inductance coil and a series condenser in each feed wire. The series condensers are adjusted until maximum current flows in the coupling coil, which indicates that the antenna system is in tune with the transmitter frequency.

Antennas which are matched to a non-resonant feeder by the use of a short section of resonant line are most conveniently tuned up by disconnecting the non-resonant line and exciting the antenna system with a near-by transmitter. The position of the short circuit on the resonant-coupling line is then adjusted until maximum current is obtained at the short-circuited point. When this condition is realized the antenna system is in tune with the desired frequency. Power is now applied to the non-resonant feeder line, and the point at which this line connects to the resonant coupling line is varied until the non-resonant condition is obtained. In the antenna system shown at Fig. 305*a*, in which the non-resonant transmission line is directly coupled to the antenna, it is first necessary to excite the antenna with the feeder disconnected and adjust the antenna length until resonance is obtained. The feed-wire connections are then made and varied until no resonance effects are observed on the transmission line.

Directional antennas are used whenever possible, since they make it possible to concentrate the radiation in the direction of the receiving station. Directional antennas are limited to communication between fixed points, however, since it is impracticable to attempt to vary the direction of maximum radiation without completely rebuilding or readjusting the antenna.

[1] This is because of the end effect which causes the wave length of a wave on the wire to be slightly less than the wave length of the same wave traveling in space. See C. R. Englund, The Natural Period of Linear Conductors, *Bell System Tech. Jour.*, vol. 7, p. 404, July, 1928.

The horizontal diamond antenna illustrated in Fig. 294 probably comes nearer to the ideal directional antenna than any other type. It will accept power from the transmitter over a wide frequency range, is simple and economical to construct, involves no complicated adjustments, and gives satisfactory directivity.

The long-wire and folded long-wire antenna arrays possess the same simplicity of construction and adjustment as does the tilted-wire antenna but differ in that these arrays must be designed for a particular frequency and require readjustment when the transmitter frequency is changed. Commercial designs of long-wire and folded long-wire antennas and associated feeder systems as used by RCA Communications, Inc., are shown in Figs. 298 and 300. Each individual long wire is tuned to resonance by adjusting the short-circuited point on the resonant coupling line. The antenna and reflectors are then coupled together by a nonresonant line having the proper length to give the required phase difference between the antenna and reflector. The point at which the line from the radio transmitter is then connected to the resonant coupling line is finally adjusted until resonance is eliminated. The process of tuning up the antenna system and adjusting the feeder lines is rather a lengthy one, since the couplings between antenna elements make the adjustments interdependent.

Directional antenna systems consisting of an array of individual antennas, such as shown in Figs. 286 and 288, are used in many transmitting stations but are gradually giving way to simpler but almost as effective directive antennas of the tilted-wire, long-wire, and partially folded long-wire types. Practical antennas composed of arrays of individual radiators are nearly always arranged in two curtains with all the radiating elements in the same curtain in phase, but with proper phasing and spacing between curtains to give a unidirectional characteristic. The directivity in a horizontal plane is determined by the horizontal length of the curtain, while the vertical directivity is controlled by the vertical length of the curtain and, in the case of horizontal antennas, also by the height of the antenna center above earth.

Typical Commercial Antenna Arrays.—Examples of commercial designs of antenna arrays composed of numerous radiators are shown in Fig. 309. These differ primarily in plane of polarization and in the methods used to obtain the proper phase of the elementary radiators. The horizontally polarized antenna of Fig. 309*d* is used in the transmitting station at Nauen, Germany, and maintains the radiators of each section in the same phase by spacing them one-half wave length apart along a resonant transmission line and connecting adjacent radiators to opposite sides of the line. Figure 309*e* shows this same array rotated 90°, resulting in a two-tier broadside array that is vertically polarized. The antenna of Fig. 309*a* consists of a wire folded in such a way as to provide vertically polarized radiators which are connected together by horizontal resonant half-wave transmission lines. The horizontally polarized radiation from the top and bottom members is small because the current flows in opposite directions in the two halves of each top and each bottom member. This type of antenna array is used at Lawrenceville, N. J., in the short-wave transatlantic telephone transmitter. The vertically polarized array of Fig. 309*f* maintains the radiators in phase by using a non-resonant transmission line in which the velocity of phase propagation is made infinite by means of shunt inductances, thus maintaining all parts of the transmission line in the same phase. This array is in use in many of the transmitting stations of RCA Communications, Inc. The array of Fig. 309*b* was developed by Chireix and Mesny and is used extensively in the French short-wave radio stations.[1] It makes use of wires so bent that the currents flowing along any diagonal are all in phase, giving a result equivalent to a vertically polarized array. The part of the array marked "parasitic radiator" is excited parasitically from the active portion of the array and has the effect of increasing the height of the array and hence the directivity. The array shown at Fig. 309*c* was devised by T. Walmsley and consists of two curtains of vertically polarized radiators spaced one-half wave length apart. Radiation from the horizontal members is small because the phase relations are such that the different horizontal portions largely cancel each others' effects. This type of antenna is used in the short-wave stations of the British Postal Service. The array of Fig. 309*g* consists of a series of half-wave co-linear radiators connected by wave coils having sufficient inductance and distributed capacity to represent a half wave length. These coils shift the phase 180° without producing appreciable radiation and so permit the co-linear antennas to radiate in phase with each other. For the sake of simplicity the drawings of Fig. 309 do not show the reflecting curtain that is always used to give a unidirectional characteristic. This reflecting curtain is ordinarily structurally similar to the radiating array except in the TW array, where it is customary to employ a parasitic reflector consisting of a series of half-wave-length antennas behind each tier of the array.

Miscellaneous Features of Antenna Arrays.—The process of tuning up an antenna array is a long and complicated procedure because of the couplings which exist between different parts of the array and which cause a wave length as measured along the array wires to differ by as much as 10 per cent from the wave length of the same wave in space. The usual procedure is to excite the array at the desired frequency and then adjust the lengths of the various elements until the current loops are located at the desired points. In the case of the reflector curtain it is further necessary to insure that the currents in the reflector are in the

[1] H. Chireix, French System of Directional Aerials for Transmission on Short Waves, *Exp. Wireless and Wireless Eng.*, vol. 6, p. 235, May, 1929.

proper phase to avoid backward radiation. This is accomplished by placing a receiver some distance behind the reflector and adjusting the reflector for minimum backward signal strength.

Steel towers, guy wires, and other conductors in the vicinity of an antenna array tend to alter the directive characteristics and so must be

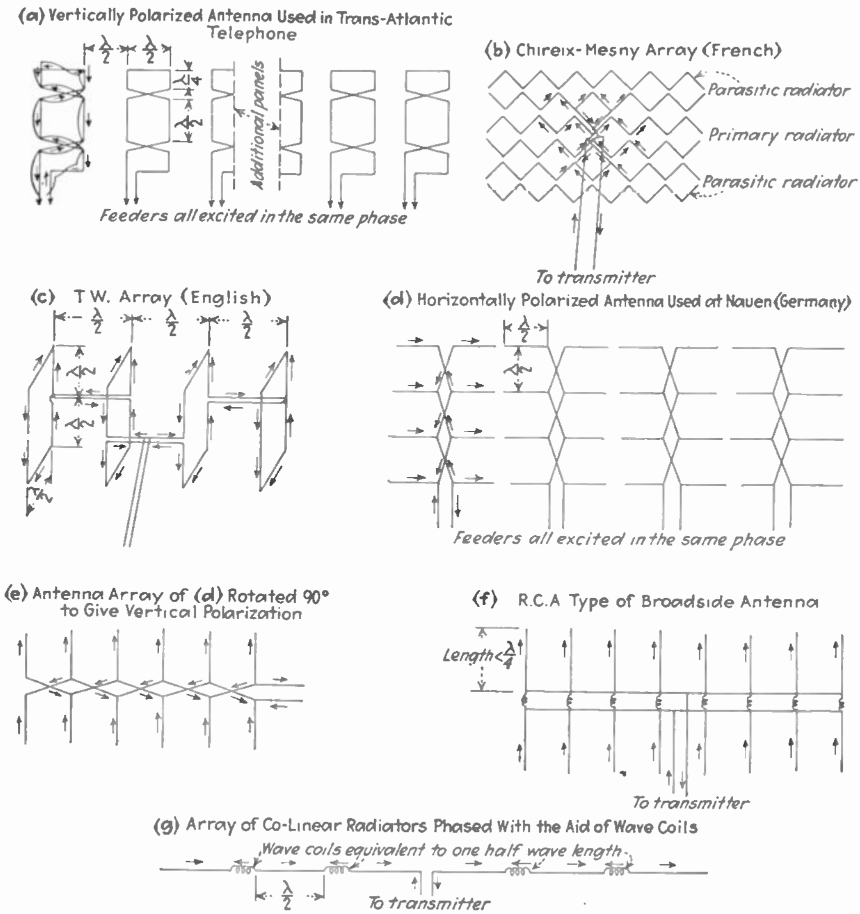


FIG. 309.—Commercial designs of antenna arrays. The reflector antenna that is always used to give a unidirectional characteristic is omitted in the figures for the sake of clearness.

located with considerable care. Steel towers should not have a length that makes them resonant to the frequency being transmitted and should furthermore be located at points where the coupling to the antenna system is minimum. Guy and supporting wires should be reduced to the smallest possible number and should be broken up into lengths much shorter than a half wave length by means of insulators.

Commercial antennas used in regions where severe winters are experienced are usually arranged so that a 60-cycle current can be passed through the antenna system without interfering with the normal functions. In this way the antenna wires can be warmed sufficiently to prevent the formation of sleet.

The amount of directivity that can be usefully employed in an antenna array is limited by the fact that the best direction in which to concentrate the radiation varies somewhat with atmospheric conditions. Thus it has been found that the width of the beam in a horizontal plane should be not less than about 10°, for otherwise there are times when the beam will be deflected sufficiently from the great circle path to miss the receiving point entirely. In general it appears that the amount of directivity in a horizontal plane that can be usefully employed is somewhat greater than is desirable in a vertical plane, and that when the gain of the array is greater than 50 to 100 times as compared with a half-wave antenna the average strength of the signals received at a distant point will be less than would be obtained from the same power delivered to a less directional array.

Short-wave communication over great distances is carried on most effectively when the transmitting antenna directs the maximum radiation at an angle very close to the horizontal. It is generally considered that the lowest possible angle of elevation is best, but at the same time the loss that results from directing the main radiation at an angle in the order of 10° above the horizon is not great, and cases are on record where angles as high as 80° gave better results than a low-angle radiation[1]. It is therefore desirable to direct most of the radiated energy as near the horizontal as is practicable but at the same time to avoid excessive directivity in a vertical plane.

Horizontally and vertically polarized transmitting antenna arrays can be used with almost equal success. The horizontally polarized array is frequently less expensive to construct since it usually does not require as high supporting structures as does the vertical array, but has the disadvantage of directing the maximum radiation at an angle with respect to the horizon, while the vertical array gives maximum radiation along the horizontal. The maximum radiation from a horizontally polarized array is at a low angle of elevation however if the average antenna height is one wave length or more above earth, and it is questionable as to whether there is any material loss in signal strength at the receiver under these conditions.

127. Transmitting Antennas for Broadcast and Lower Frequencies.— Antennas used in radiating broadcast signals are nearly always grounded at the lower end and are provided with a small or moderate-sized flat

[1] See A. Meissner, Directional Radiation from Horizontal Antennas, *Proc. I.R.E.*, vol. 15, p. 928, November, 1927.

top suspended above earth at a height approaching a quarter of a wave length. The principal problems involved in the design of such an antenna are concerned with keeping down the loss resistance. If this is not done the radiation efficiency will be low, since at broadcast frequencies a quarter wave length represents considerable distance, which makes the radiation resistance small in proportion to the antenna size.

The most important component of the antenna loss is the earth resistance, which tends to be high because the earth is a poor conductor and is so large that considerable skin effect is present at high frequencies. The most satisfactory method of obtaining a low-resistance ground connection is to use a "counterpoise" to supply the earth connection. The counterpoise consists of a network of conductors placed a few feet above ground and more or less completely shielding the earth from the electrostatic field of the antenna. The counterpoise thus serves all the functions of an ordinary earth connection, but being composed of copper conductors has a very low resistance. In cases where the earth is an especially good conductor it is possible to obtain a satisfactory ground connection by burying a network of wires a few inches below the surface of the earth.

The service range of a broadcast station is determined by the field strength radiated in a horizontal direction. The waves radiated skyward, while ultimately returning to earth after reflection by the Kennelly-Heaviside layer, cannot be depended upon to give satisfactory reception since they are nearly always accompanied by fading. The ideal broadcasting antenna would therefore be one which would concentrate all of the radiation along the horizontal and send little or no energy upward at an appreciable angle of elevation. Such an antenna would require such extremely high towers as to be out of the question, but serious proposals have been made to employ vertical antennas approximately one-half wave length long instead of the quarter-wave-length antennas now customary.[1] Such an increase in height would enable a given amount of power supplied to the antenna to have a service area about 50 per cent greater than obtained with a quarter-wave-length antenna.

Transmitting antennas for frequencies in the range 100 to 500 kc are ordinarily grounded antennas of the flat-top type. The height of the flat top is made as great as possible but is always such a small fraction of a wave length that the radiation efficiency of these antennas is rather low unless extreme precautions are taken to provide a low-resistance ground connection. The ground is usually of the buried-wire type, since at the lower frequencies the skin effect is not so pronounced, and

[1] For example, see A. Meissner, Transmitting Antennas for Broadcasting, *Proc. I.R.E.*, vol. 17, p. 1178, July, 1929; P. P. Eckersley, T. L. Eckersley, and H. L. Kirke, Design of Transmitting Aerials for Broadcasting Stations, *Jour. I.E.E.* (*London*), vol. 67, p. 507, April, 1929.

the counterpoise area required to intercept all of the electrostatic flux from the antenna would be enormous.

Long-wave Antennas.—Antennas used at frequencies below 100 kc are provided with large flat tops suspended at a height of 300 to 600 ft. above earth. Principal considerations in the design of such antennas are the height, which determines the radiation resistance, the antenna-loss resistance, which controls the radiation efficiency, and the flat-top capacity, which determines the maximum antenna voltage developed by a given antenna current.

The height above earth must be as great as possible, and even then the radiation resistance is only a fraction of 1 ohm. This low radiation resistance makes it necessary to take extreme care in keeping the loss resistance down. In particular, the tuning coils employed in the antenna circuit must have a high Q, corona must be entirely absent, and the resistance of the earth connection must be as low as possible. The most effective way of keeping the ground resistance low is to ground the antenna at a number of points, so that current can reach a ground connection without traveling very far through the high-resistance earth. A method of providing such earth connections is the multiple-tuned antenna shown in Fig. 310 in which the antenna is provided with a series of tuned down leads spaced at regular intervals. The tuning is so adjusted that the currents in the various down leads are all in phase, in which case the action is equivalent to placing the earth connections in parallel.[1]

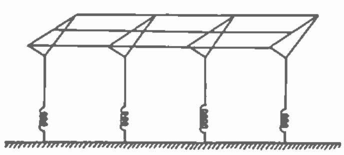


FIG. 310.—Multiple-tuned long-wave antenna. The several down leads have the effect of providing parallel ground connections, giving a low total effective ground resistance.

The voltage to which the flat top of a long-wave transmitting antenna is charged is equal to $I/2\pi fC$. Since the antenna current I is commonly in the order of hundreds of amperes, and since the frequency is low, the flat-top capacity must be large if excessive voltages and consequent corona losses are to be avoided. In order to provide this capacity, flat-topped structures over a mile long and 500 to 600 ft. wide are not at all uncommon.

128. Receiving Antennas.—When it is desired to avoid marked directional effects in the reception of broadcast and lower frequencies the receiving antenna ordinarily consists of a single wire grounded at one end, with the other end raised to the highest elevation obtainable without undue expense. The principal requirements that such a non-directional receiving antenna must meet are: first, a height sufficient to enable vertically polarized waves having a strength between 0.1 and 1.0 μv per meter to deliver to the radio receiver sufficient signal energy to be audible

[1] For further information on the design of long-wave transmitting antennas see N. Lindenblad and W. W. Brown, Main Considerations in Antenna Design, *Proc. I.R.E.*, vol. 14, p. 291, June, 1926.

above the receiver noise level; and, second, dimensions sufficiently small compared with a wave length to avoid marked directional effects. The energy abstracted from waves of broadcast and lower frequencies is determined largely by the antenna height, since such waves are always vertically polarized. If this energy is sufficient to overcome the receiver noises, its exact amount and the antenna efficiency (*i.e.*, the fraction of the total abstracted energy that is delivered to the radio receiver) are unimportant, since radio frequency amplification is relatively inexpensive.

Non-directional antennas for the reception of short-wave signals present problems not encountered in the case of lower frequency signals because the antenna dimensions must not be appreciably greater than one-half wave length. At high frequencies this is such a small size that the amount of energy abstracted from passing waves is relatively small and may easily be insufficient to overcome the noise level inherent in the radio receiver. The short-wave receiving antenna commonly employed when directivity is not important consists of a wire in the order of 50 to 100 ft. long coupled to the receiver through a small inductance or capacity. While such an arrangement possesses some directivity at the shorter wave lengths it is ordinarily found satisfactory. Non-directional receiving antennas for short waves are commonly run from a convenient supporting structure directly to the radio receiver and may be either horizontal, vertical, or inclined at an angle, since short-wave signals arriving from distant transmitters nearly always have both vertically and horizontally polarized components.

Directional Receiving Antennas.—In point-to-point radio communication it is of considerable advantage to employ directional receiving antennas, since this reduces the interference from static and unwanted radio transmitters and hence effects material improvement in the signal-to-noise ratio. This is equivalent to increasing the power of the transmitter, and in cases where the bulk of the interference arrives from directions other than that in which the transmitter is located, benefits obtained from a directional receiving antenna are very great.

The Beverage wave antenna is the most satisfactory directional antenna that has been devised for the reception of signals of broadcast and lower frequencies. Such antennas are used extensively in long-wave point-to-point communication, and have excellent directional characteristics when the length is in the order of one wave length, particularly when several such antennas are arranged in an array as shown in Fig. 291.

Directional antennas for the reception of high-frequency radio signals make use of the same principles that are employed in directional transmitting antennas. It has been found, however, that while the receiving antenna can make use of high directivity in a horizontal plane, the directivity in a vertical plane should be relatively small, since the angle of elevation at which the major part of the energy is received from a distant transmitting station varies greatly. At times nearly all the energy will arrive along the horizontal, while at other times it approaches the receiver from a direction that is nearly vertical. The result is that a receiving antenna possessing marked directivity in a vertical plane, such as would be obtained by stacking antennas one above the other, will on the average be less effective than a simple non-directional antenna.

There is relatively little choice between horizontally and vertically polarized receiving antennas if the horizontally polarized antenna is at a height of one wave length or more above earth. This is because the plane of polarization of short-wave signals arriving from a distant transmitter is always independent of the plane of polarization of the transmitting antenna. There is a tendency, however, to favor horizontally polarized receiving antennas since experience has shown that man-made interference of local origin is in the main vertically polarized and so creates less disturbance when a horizontal antenna is employed.

The most satisfactory directional short-wave antennas are the Bruce antenna, the short-wave Beverage antenna, and the horizontal-diamond antenna. The Bruce antenna is shown in Fig. 311 and consists of a bent

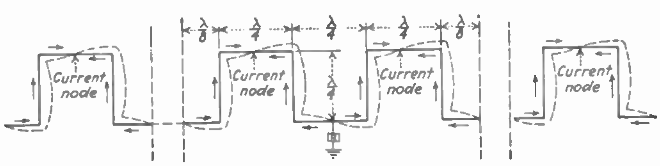


Fig. 311.—Bruce type of receiving antenna, together with current distribution.

wire so arranged that the quarter-wave-length vertical sections are all in phase, while the two halves of each horizontal section carry currents of opposite phase and so act primarily as a transmission line. This type of antenna possesses very little vertical directivity because of its small vertical height, but will give excellent horizontal directivity provided the over-all length is in the order of five to ten wave lengths. A unidirectional characteristic is obtained by placing a similar reflecting antenna one-quarter wave length behind the first antenna. The receiving antenna in the original short-wave transatlantic telephone circuit is of the Bruce type. The short-wave Beverage antenna is usually employed in a horizontally polarized broadside array consisting of two such antennas side by side about one wave length above the earth, as illustrated in Fig. 292. This antenna possesses satisfactory directive characteristics and has the desirable property of functioning satisfactorily over a considerable frequency range without requiring readjustment. The horizontal-diamond antenna is illustrated in Fig. 294 and, like the Beverage short-wave antenna, will function satisfactorily over a wide frequency range without any adjustments. This type of antenna is the simplest of all the directional antennas and yet possesses excellent characteristics.

Other types of directional short-wave antennas, such as the broadside and end-fire arrays, and the long-wire and partially folded long-wire antennas, can also be used in the reception of short-wave signals. These arrays function satisfactorily at only a single frequency, however, and furthermore most of them possess other disadvantages which make their use in reception desirable only under special circumstances.

In the construction of directional short-wave receiving antennas it is necessary to take special precautions to prevent the directional characteristics from being partially destroyed by the action of near-by conductors such as towers, guy wires and power lines. Such objects, if of the righ,

Fig. 312.—Airplane equipped with antenna such as used in short-wave communication between plane and ground.

dimensions, will abstract appreciable energy from waves that arrive in unwanted directions and will reradiate this energy to the receiving antenna. If the interference arriving from these unwanted directions is large this action will largely balance out the other advantages of the directional antenna. It is hence customary to locate short-wave directive receiving antennas in spaces that are clear of trees, houses, power and telephone wires, etc., and considerable attention is also paid to minimizing the effect of towers and other metal structures that must necessarily be near the antenna. In the Bruce antenna used in the short-wave transatlantic telephone the antenna is supported by a wooden framework rather than steel towers in order to avoid all distortion of the directional characteristic of the antenna.

129. Antennas for Use on Aircraft.—The simplest type of aircraft antenna consists of a wire weighted at the lower end and trailing from the under side of the plane. Such antennas can be given a length that will lead to a high efficiency at any desired wave length but are not viewed with favor in commercial operation because of the hazard to the plane, and because the trailing weight is often snapped off.

Commercial aircraft equipped with radio transmitters ordinarily employ a small antenna consisting of wires running from the top of a short pole to the wing tips or to the tail of the plane. A typical plane antenna is shown in Fig. 312. Antennas of this type, while small, are sufficiently effective at wave lengths of 100 meters and less to permit telephone communication from plane to ground to be carried on over distances in the order of 100 miles using the 50-watt transmitter illustrated in Fig. 237, and when sensitive receivers are employed are satisfactory for reception of both long- and short-wave signals.

CHAPTER XV

PROPAGATION OF RADIO WAVES

130. Factors Affecting the Propagation of Radio Waves.—The energy radiated from a transmitting antenna can be divided into a ground wave which is in immediate contact with the surface of the earth, and a sky wave which is propagated in the atmosphere above the earth. The ground wave is vertically polarized since the conducting earth would short-circuit and hence wipe out the electrostatic flux of a horizontally polarized ground wave. At the point where the vertical electrostatic flux of the ground wave meets the earth there is an induced charge which travels with the wave along the surface of the ground. The energy loss occasioned by this induced charge flowing through the ground resistance is supplied by the wave and causes a continual flow of energy downward toward the ground that gives rise to a slight forward tilt of the wave front.

The strength of the ground wave that starts away from the transmitting antenna is determined by the field which the antenna radiates along the horizontal, and so depends upon the directional characteristics of the transmitting antenna. In particular, a horizontally polarized transmitting antenna produces no ground wave whatsoever, while an array consisting of radiators stacked vertically and excited in phase produces a very powerful ground wave. As the ground wave travels away from the transmitting antenna it becomes attenuated as a result of spreading out to cover a greater area and because of energy abstracted from the wave. The spreading out effect depends only on distance, being independent of the frequency or character of terrain over which the wave passes, while the energy absorbed from the wave depends upon the earth resistance; irregularities of the earth's surface, such as hills and valleys; imperfect dielectrics, such as trees and buildings on the surface of the earth; and the presence of conducting objects, such as metal buildings and radio antennas, in which the wave induces voltages.

Ground-wave Propagation.—At the lower radio frequencies the energy abstracted from the ground wave is very small, making the principal attenuating factor that of distance. The result is that the ground wave of high-power long-wave radio transmitters is of importance at distances as great as 1000 to 2000 km and is affected only to a small extent by the character of the surface over which the wave travels. As the frequency is increased, the energy abstracted from the ground wave becomes greater, until at broadcast frequencies the range of the ground wave from a powerful transmitter seldom exceeds 100 to 150 km. Furthermore at broadcast frequencies the range of the ground wave is affected to a marked extent by the character of the earth's surface, being very much greater over sea than over land and over level prairie country than over forested, rough, or urban areas. At very high frequencies (wave lengths of 50 meters or less) the rate at which energy is abstracted from the ground wave by the earth is so great that the range of the ground wave is in the order of a few kilometers.

An important factor contributing to the reduction in ground-wave range that occurs as the frequency is raised is that the only energy available for replenishing ground losses is within several wave lengths of the earth and so becomes less as the wave length becomes less (*i.e.*, as the frequency is raised). The situation can be summarized by stating that since the high-frequency waves tend to travel in straight lines the energy in the wave at some distance above the surface of the earth will not spread downward to replace earth losses, whereas the low-frequency waves do not possess this tendency toward straight-line propagation to the same extent.

The propagation of the ground wave is substantially independent of factors such as season and time of day, which are of fundamental importance in determining the characteristics of the sky wave. Moisture conditions of the surface of earth have some slight effect, but in general the ground wave can be depended upon to behave substantially the same under practically all circumstances that are encountered in practice.

Action of an Electron under the Influence of a Radio Wave.—The rapid attenuation of the ground wave at all except the lowest frequencies makes it necessary to rely upon the sky wave in long-distance communication. The sky wave is able to reach distant points as a result of ionized regions in the upper atmosphere which refract the wave back to earth and permit propagation of energy around the curvature of the earth. The effect which an ionized region has on a radio wave can be understood by considering the behavior of a single ion or electron when under the influence of a passing radio wave. Take first the case of an electron in a vacuum with no magnetic field present other than the weak magnetic field of the wave. The wave's electrostatic field exerts forces on the electron which vary sinusoidally with time and cause the electron to vibrate sinusoidally along a path parallel with the flux lines of the wave. The amplitude and average velocity of vibration are greater the lower the frequency, and the velocity lags 90° behind the electric field of the radio wave because the moving electron offers an inertia reactance to the forces acting upon it. Since a moving charge is an electrical current, the vibrating electron acts as a small antenna which abstracts energy from the radio waves and then reradiates this energy in a different phase. The resulting effect is exactly as though the vibrating charge was a parasitic antenna tuned to offer an inductive (*i.e.*, inertia) reactance to the wave frequency, and alters the direction in which the resultant energy flows (see Sec. 124). The magnitude of this effect varies with the amplitude and average velocity of the electron vibration and therefore becomes increasingly great as the wave frequency is lowered. Ions in the path of a wave act in much the same way as electrons, but because of their heavier mass ions move enormously slower than electrons under the same force and so in comparison have negligible effect.

Effect of the Earth's Magnetic Field.—The effect which electrons in the atmosphere have on radio waves is influenced by the fact that these electrons are in the presence of the earth's magnetic field, which exerts a deflecting force on the moving electron as explained in Sec. 24. The magnitude of this deflecting force is proportional to the instantaneous velocity of the electron in a direction at right angles to the lines of magnetic flux. At very high radio frequencies, where the maximum velocity attained by an electron in the path of the wave is small, the deflecting forces which are exerted by the earth's magnetic field are correspondingly slight, with the result that the path followed by the vibrating electron is a narrow ellipse (see Fig. 313*a*). The ratio of minor to major axes of the ellipse depends upon the orientation of the electrostatic flux lines of the radio wave with respect to the earth's magnetic field. The motion along the minor axis causes some of the energy reradiated by the vibrating electron to have a component polarized at 90° with respect to the polarization of the passing wave. Hence a radio wave in passing through an ionized region in the upper atmosphere has its plane of polarization affected by the earth's magnetic field.

As the frequency of the wave is reduced the maximum velocity attained by an electron under the influence of the wave is greater, and the minor axis of the elliptical path becomes larger in proportion to the major axis (see Fig. 313*b*), thus increasing the fraction of the absorbed energy that is reradiated in a different plane of polarization. This tendency continues as the frequency is lowered until at about 1400 kc the deflecting force of the magnetic field is so related to the frequency that the direction of the electron is reversed at exactly the same instant the electrostatic flux of the passing wave changes polarity, causing the electron to follow the spiral path shown at Fig. 313*c*. Since the electron velocity in the spiral increases without limit the electron abstracts more energy from the passing wave than is reradiated, causing the sky-wave attentuation at 1400 kc to be greater than at higher or lower frequencies. At frequencies lower than this critical or resonant frequency the path followed by an electron vibrating in the presence of the earth's magnetic field is approximately as shown at Fig. 313*d* and *e*. The principal effect of the magnetic field under these conditions is to lower the maximum electron velocity, thus reducing the amount of energy which is absorbed and reradiated from a low-frequency radio wave. The effect which an electron path such as shown at Fig. 313*d* and *e* has on the plane of polarization is relatively small since the radiation from the two sides of each reversal loop is of opposite phase and tends to cancel.

Effect of Atmospheric Pressure on Electron Behavior.—The discussion that has been given of the effect which an electron has on a passing radio wave has assumed that the electron was in a vacuum. Actually, however, there is always a certain amount of gas present in the atmosphere even at high elevations, and from time to time the vibrating electron will collide with a gas molecule. In such a collision the kinetic energy which the electron has acquired from the radio wave is partly transferred to the gas molecule and partly radiated in the form of a disordered radio wave which contributes nothing to the transmission. The net result is therefore an absorption of energy from the passing wave. The magnitude of the energy thus absorbed depends upon the gas pressure (*i.e.*, upon the likelihood of a vibrating electron colliding with a gas molecule) and upon the velocity which the electron acquires in its vibration (*i.e.*, upon the energy lost per collision). The absorption of energy from a radio wave passing through an ionized region is hence less for a wave of high frequency than for a wave of low frequency, and in the case of low-frequency waves is reduced by the presence of a magnetic field.

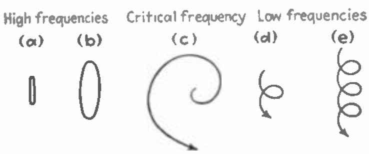


FIG. 313.—Paths followed by an electron in the earth's magnetic field when vibrating under the influence of a radio wave. To avoid confusion the paths for the low-frequency cases represent only a half-cycle of the vibration, which repeats cyclically.

The Kennelly-Heaviside Layer.—Studies of the propagation characteristics of radio waves have established beyond any possible doubt that as a result of ultra-violet radiation from the sun there is an ionized region in the upper atmosphere at an elevation in the order of several hundred kilometers. This is called the Kennelly-Heaviside layer, after the names of the two scientists who independently and almost simultaneously pointed out that the existence of such a region in the upper atmosphere would account for many of the properties of radio waves. The way in which the electrons are distributed in the Kennelly-Heaviside layer is not known with certainty, but it appears that the lower boundary of the ionized region is rather sharply defined, and that the electron density increases rapidly above this lower boundary and reaches a maximum above which the density becomes less. Whether the electron density increases steadily to a single maximum, as shown at Fig. 314*a*, or whether there is a minor maximum near the lower boundary, as shown at Fig. 314*b*, has not been definitely determined, since the experimental data on this point are not yet conclusive.[1] The height and electron density

[1] Thus see G. W. O. Howe, How Many Ionized Layers, *Exp. Wireless and Wireless Eng.*, vol. 8, p. 463, September, 1931.

of the Kennelly-Heaviside layer show a marked diurnal and a lesser seasonal variation, the layer being lowest and the electron density greatest in the daytime and in summer.

The effect which the Kennelly-Heaviside layer has on the propagation of the sky wave sent out from a radio station is relatively complicated because the energy which each individual electron in this layer abstracts from the passing radio-wave energy is reradiated in a new phase and new plane of polarization, and this reradiated energy is in part reabsorbed by other electrons which cause a further change in the phase and plane of polarization. The end result is equivalent to a refraction, *i.e.*, bending, of the radio wave away from the regions of high electronic density toward regions of lower density, as indicated in Fig. 315. The exact path followed by a ray leaving the transmitting antenna is determined by the extent to which the refractive index of the upper atmosphere departs from unity as a result of the ionized Kennelly-Heaviside layer.

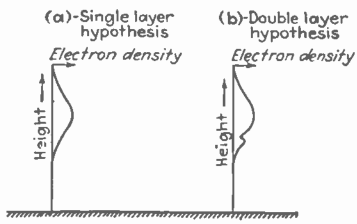


FIG. 314.—Hypothetical distributions of electrons in the upper atmosphere that have been suggested by different investigators.

The refractive index of a medium is defined as the ratio of the velocity of light to the phase velocity of a wave traveling through the medium. When the magnetic permeability is unity, as is the case in the Kennelly-Heaviside layer, the refractive index is equal to the square root of the dielectric constant of the medium. When the gas pressure is so low that collisions between the free electrons and the gas molecules are relatively infrequent, and when no magnetic field is present, the dielectric constant and refractive index of an ionized region are:[1]

$$\text{Dielectric constant } k = 1 - 81\frac{N}{f^2} \tag{167}$$

$$\text{Refractive index } u = \sqrt{k} = \sqrt{1 - \frac{81N}{f^2}} \tag{168}$$

where

N = electron density in electrons per cubic centimeter

f = frequency in kilocycles.

It is apparent from Eq. (167) that the dielectric constant and hence the refractive index become lower the greater the electron density and the lower the frequency. When the gas pressure is sufficient to cause frequent collisions between the vibrating electrons and the gas molecules the effect is to lower the average velocity which the electrons acquire

[1] A simple derivation of Eq. (167) is given by William G. Baker and Chester W. Rice, Refraction of Short Radio Waves in the Upper Atmosphere, *Trans. A.I.E.E.*, vol. 45, p. 302, 1926. See Eckes - IRE Sept. 1931 for better form of results

under the influence of a passing wave, with the result that the refractive index is nearer unity than given by Eq. (167). The exact effect of such collisions follows a rather complicated law but can be calculated.[1] At low radio frequencies the earth's magnetic field has much the same effect on the refractive index of the Kennelly-Heaviside layer as does the presence of gas of relatively high pressure since, as has already been explained, the magnetic field reduces the average velocity of the vibrating electrons. At frequencies above about 6000 kc the magnetic field has negligible effect on the refractive index.

Path of Radio Wave in Kennelly-Heaviside Layer.—The path followed by a radio wave in traveling through an ionized region in the absence of a magnetic field is determined by the refractive index of the medium

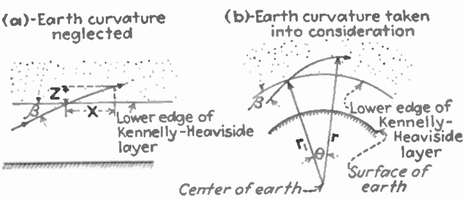


FIG. 315.—Diagrams illustrating notation Eqs. (169) and (170).

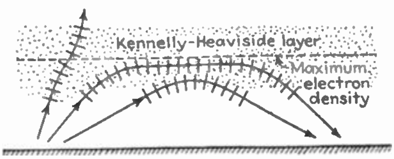


FIG. 316.—Diagram illustrating the change that takes place in the direction of the wave front when a wave travels through the Kennelly-Heaviside layer. The part of the wave front in the region of greatest electron density advances faster than the rest of the wave and so causes a refraction.

and the angle of incidence with which the wave enters the medium. Upon the assumption that the refractive index is a function only of height above earth, and that the earth's surface is flat, a wave entering the Kennelly-Heaviside layer follows a path that is given by the following differential equation:[2]

$$dx = \frac{\cos \beta dz}{\sqrt{\mu^2 - \cos^2 \beta}} \tag{169}$$

where

β = angle of incidence of wave entering the Kennelly-Heaviside layer

μ = refractive index of Kennelly-Heaviside layer, and is a function only of z

z = height of ray above lower edge of layer

x = distance parallel to earth's surface, measured from the point at which the wave enters the ionized region.

This notation is shown graphically at Fig. 315*a*.

[1] See Appendix I of Baker and Rice, *loc. cit.*

[2] The derivation of this equation is to be found in a number of places in the literature. Thus see Baker and Rice, *loc. cit.*

When the distance along the earth's surface is so great that it is not permissible to consider the earth as flat Eq. (169) must be modified as follows:[1]

$$d\theta = \frac{r_1 \cos \beta dr}{r\sqrt{r^2\mu^2 - r_1^2 \cos^2 \beta}} \tag{170}$$

where

θ = polar angle (in radians) intercepted by wave path

r = distance from wave to center of earth

r_1 = distance from lower edge of Kennelly-Heaviside layer to center of earth

β = angle of incidence of wave entering the Kennelly-Heaviside layer.

This notation is shown graphically at Fig. 315*b*.

A physical picture of the relations expressed in Eqs. (169) and (170) can be gained with the aid of Fig. 316, which shows small sections of wave fronts traveling in different parts of the Kennelly-Heaviside layer. The curvature of the path results from the fact that each part of the wave front travels with a velocity equal to the phase velocity in that part of the medium. The edge of the wave front where the electron density is greatest (phase velocity greatest) advances faster than the rest of the front, thus causing the wave to follow a curved path in which the bending is away from the region of maximum electron density. The curvature depends upon the rate at which the electron density changes with height and is greatest when this change is maximum. At the point of greatest electron density the gradient is zero, and the path is straight. Once past this point the waves bend away from the earth and will not return except as a result of a reflection from outside the earth's atmosphere, while below the point of maximum electron density the bending is always earthward.[2] Equations (169) and (170) neglect the effect of the earth's

[1] See Appendix II of Baker and Rice, *loc. cit.*

[2] It is well at this point to note the distinction between phase velocity on the one hand, and signal or group velocity on the other. The phase velocity is determined by the rapidity with which the phase changes along the path of a wave and is equal to the velocity of light divided by the refractive index. In the case of waves traveling in the Kennelly-Heaviside layer the phase velocity is hence greater than the velocity of light and may even approach infinity under some conditions.

The signal or group velocity represents the velocity with which the energy of the wave travels rather than the rate at which the phase changes. The signal velocity cannot exceed the velocity of light and may even approach zero in certain limiting cases. The phase and signal velocities are related to each other by the equation

$$\frac{\text{Group velocity}}{\text{Phase velocity}} = \frac{1}{1 - \dfrac{\omega dv}{v d\omega}} \tag{171}$$

where $\omega = 2\pi f$, and v = phase velocity. See P. O. Pedersen, "The Propagation of Radio Waves" p. 169, G. E. C. Gad, Copenhagen, Denmark (in English).

magnetic field on the polarization of the radio waves, but the error this involves at frequencies above about 6000 kc is small. The presence of the earth's magnetic field changes the plane of polarization of the waves and also splits the wave into two components that travel with different velocities along slightly different paths.[1]

In the above discussion it has been tacitly assumed that the change in refractive index in a distance corresponding to one wave length is negligibly small. If this is not true, reflection as well as refraction takes place. At the lowest radio frequencies where the wave length is long a small amount of reflection may be present, but at the higher radio frequencies this is negligible.

131. Propagation of Low-frequency Radio Waves.—The behavior of the signals received from a transmitter radiating low-frequency radio waves depends upon the distance between receiver and transmitter. When this distance does not exceed 500 to 1000 km most of the received energy is carried by the ground wave, and time of day and season have hence very little effect. On the other hand when the distance is 5000 to 10,000 km practically all of the signal energy has traveled *via* the sky wave, and the strength of the received signal shows pronounced seasonal and diurnal variations as a result of corresponding changes in the Kennelly-Heaviside layer. At intermediate distances the sky and ground waves are both of appreciable magnitude, causing the observed field strength to be the vector sum of two waves which have traveled over paths of unequal lengths and so in general are not of the same phase. Since this phase difference depends upon the distance to the transmitter there is tendency for the field strength to alternately increase and decrease as the distance from receiver to the transmitter is increased.[2] The relative phase of the two component waves at any particular location depends upon the height of the Kennelly-Heaviside layer as well as the distance to the transmitter, and so shows diurnal and seasonal variations. The result is that while in summer the signals received from a not too distant transmitter may be stronger during the day than at night, the opposite may be true in winter, while at another receiving location these actions may be entirely reversed. A schematic illustration of the different receiving zones in the case of low-frequency waves is shown in Fig. 317.

Diurnal and Seasonal Variations in Signals Received at Great Distance.—When the entire transmission path is in daylight the field strength of low-frequency waves radiated from a distant transmitting station is normally moderately low but relatively constant. As soon as the sunset line falls

[1] An analysis of the effects produced by the earth's magnetic field is given by H. W. Nichols and J. C. Schelleng, Propagation of Electric Waves over the Earth, *Bell System Tech. Jour.*, vol. 4, p. 215, April, 1925.

[2] See J. Hollingworth, Propagation of Radio Waves, *Jour. I.E.E.* (*London*), vol. 64, p. 579, May, 1926.

across the transmission path there is a sudden reduction in field strength, after which the intensity rapidly increases to a high value which is maintained until sunrise, when the signals return to the normal daytime level. Typical curves showing diurnal variation in field strength of long-wave signals at different times of year are shown in Fig. 318.[1] The principal seasonal effect is longer night and shorter day transmission conditions in winter. The seasonal and diurnal variations become more pronounced as the frequency is raised, as is brought out by Fig. 318. While low-frequency radio signals normally behave in a fairly regular manner, neither the daily nor yearly cycles repeat exactly. There is also evidence to show that low-frequency radio waves propagate in a north-south direction with less attenuation than when traveling in an east-west direction.[2]

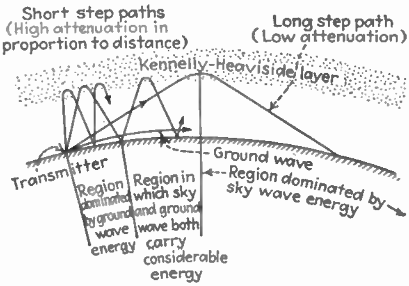


FIG. 317.—Schematic diagram illustrating paths followed by energy radiated from a long-wave transmitting antenna.

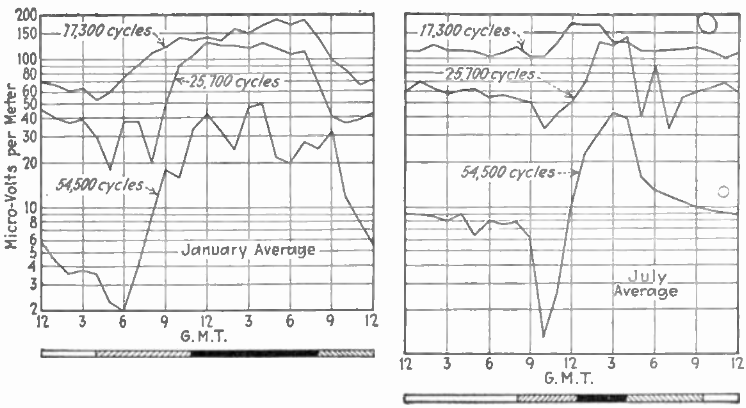


FIG. 318.—Curves showing average diurnal variation in strength of long-wave signals of different frequencies propagated across the north Atlantic during mid-winter and mid-summer months. (Note that signal strengths at different frequencies cannot be compared because the radiated power was not the same at all frequencies.) The solid and clear strips at the bottom of the figure indicate periods when the entire transmission path is in darkness and light, respectively, while the shaded strips indicate part of the path in darkness and part in light.

Austin-Cohen Formula.—The normal daytime signal that a long-wave radio transmitter delivers to a distant receiver when the transmission path

[1] These curves are taken from Lloyd Espenschied, C. N. Anderson, and Austin Bailey, Transatlantic Radio Telephone Transmission, *Proc. I.R.E.*, vol. 14, p. 7, February, 1926.

[2] See Eitaro Yokoyama and Tomozo Nakai, East-west and North-south Attenuations of Long Radio Waves on the Pacific, *Proc. I.R.E.*, vol. 17, p. 1240, July, 1929.

is over water is given with fair accuracy by the revised Austin-Cohen formula, which is as follows:[1]

See IRE July 1931 p. 1152 for inverse distance law

$$\epsilon = 377\frac{hI}{\lambda d}\sqrt{\frac{\theta}{\sin\theta}}\,e^{\frac{-0.0014d}{\lambda^{0.6}}} \times 10^3 \tag{172}$$

where

e = base of Napierian logarithms
ϵ = field strength in microvolts per meter
h = effective height of transmitting antenna in kilometers
λ = wave-length in kilometers
d = distance in kilometers to transmitter
θ = angle at center of earth intercepted by transmission path in radians
I = current flowing in the vertical part of the transmitting antenna.

Ray Paths.—The way in which low-frequency waves are propagated to the receiver can be understood by considering the paths actually followed by the waves, and the relation of these paths to the earth and the Kennelly-Heaviside layer.[2] Figure 317 shows the route of several such hypothetical rays leaving the transmitting antenna at different vertical angles. At low frequencies, such as are here involved, the Kennelly-Heaviside layer is a very efficient refracting medium which bends the waves that strike it backward toward the earth in very much the same manner as would a mirror. The sky wave thus refracted in the upper atmosphere travels downward and strikes the earth, where it is reflected with the angle of reflection equaling the angle of incidence. Each time the wave is reflected at the surface of the earth or refracted by the Kennelly-Heaviside layer a certain amount of energy is lost by absorption and scattering, while propagation in the intervening space is substantially without attenuation other than that caused by spreading. The result is that most of the energy reaching a distant receiving point travels along the path involving the least number of steps and so follows the ray leaving the transmitting antenna at a low vertical angle. It is thus apparent that the most desirable type of long-wave transmitting antenna would be one concentrating the radiation along the horizontal, but unfortunately such an antenna is not practicable because of the enormous height required to give vertical directivity at long wave lengths.

The diurnal and seasonal variations in signal strength that are observed at a great distance from a long-wave transmitter result from corresponding variations in the height and electron density of the Kennelly-Heaviside layer. The attenuation will in general be greater the lower the layer and

[1] L. W. Austin, Preliminary Note on Proposed Changes on the Constants of the Austin-Cohen Transmission Formula, *Proc. I.R.E.*, vol. 14, p. 377, June, 1926.

[2] This discussion closely follows P. O. Pedersen, "The Propagation of Radio Waves," G. E. C. Gad, Copenhagen, Denmark (in English).

the greater the electron density, and so is less at night than in the day, and less in winter than in the summer. These seasonal and diurnal variations become more pronounced as the frequency of the wave is increased, until, at frequencies in the order of 500 kc, communication can be carried on only at night even when the transmission distance is only moderate, and even then the received signal strength is seldom as great as that developed by a like amount of energy radiated at a lower frequency.

Low-frequency radio waves are always vertically polarized when observed near the earth's surface and normally have a wave front that is tilted somewhat forward. The vertical polarization results from cancellation of the horizontally polarized waves at the earth's surface as a result of reflection with reversal of the electrostatic field. As the frequency is increased slight deviations from the vertical may appear in the angle of polarization, but these effects are usually small at reasonable distances above the ground.

Fading, *i.e.*, rapid change in signal strength, is seldom observed at low frequencies except occasionally during the sunset period or at night, and even then the fading is relatively slow compared with fading of broadcast and higher frequencies.

132. The Propagation of Waves of Broadcast Frequencies (Frequency Range 550 to 1500 Kc).—The primary object of a broadcast transmitter is to deliver a high-quality signal continuously to listeners within a limited distance of the transmitter rather than to produce an understandable signal at great distances, as is the case in other types of radio communication. The region around the transmitter in which reception free from objectionable disturbances or distortion can be obtained at all times is called the service area of the broadcast station and represents the region in which the ground wave is powerful enough to override all ordinary interference (either natural or man made). Within the service area the signals will show little or no seasonal and diurnal variation in strength and will have an entertainment value that is not reduced by interference. The strength of ground wave required to overcome local interference varies with the location of the receiver and is much higher in cities, particularly in industrialized districts, than in rural regions. Experience has shown that satisfactory service in metropolitan areas requires a signal strength in the order of 5–30 mv/m, while in rural regions fairly satisfactory service is frequently obtained with signal strengths as low as 0.1 mv/m.[1]

The size of the service area obtained with a given amount of energy radiated from the transmitting antenna depends primarily upon the transmitter frequency and the character of the country over which the ground

[1] See Lloyd Espenschied, Radio Broadcast Coverage of City Areas, *Trans. A.I.E.E.*, vol. 45, p. 1278, 1926; C. M. Jansky, Jr., Some Studies of Radio Broadcast Coverage in the Mid-west, *Proc. I.R.E.*, vol. 16, p. 1356, October, 1928.

wave passes. The attenuation increases very rapidly with frequency and is also much greater over broken country than over water. In

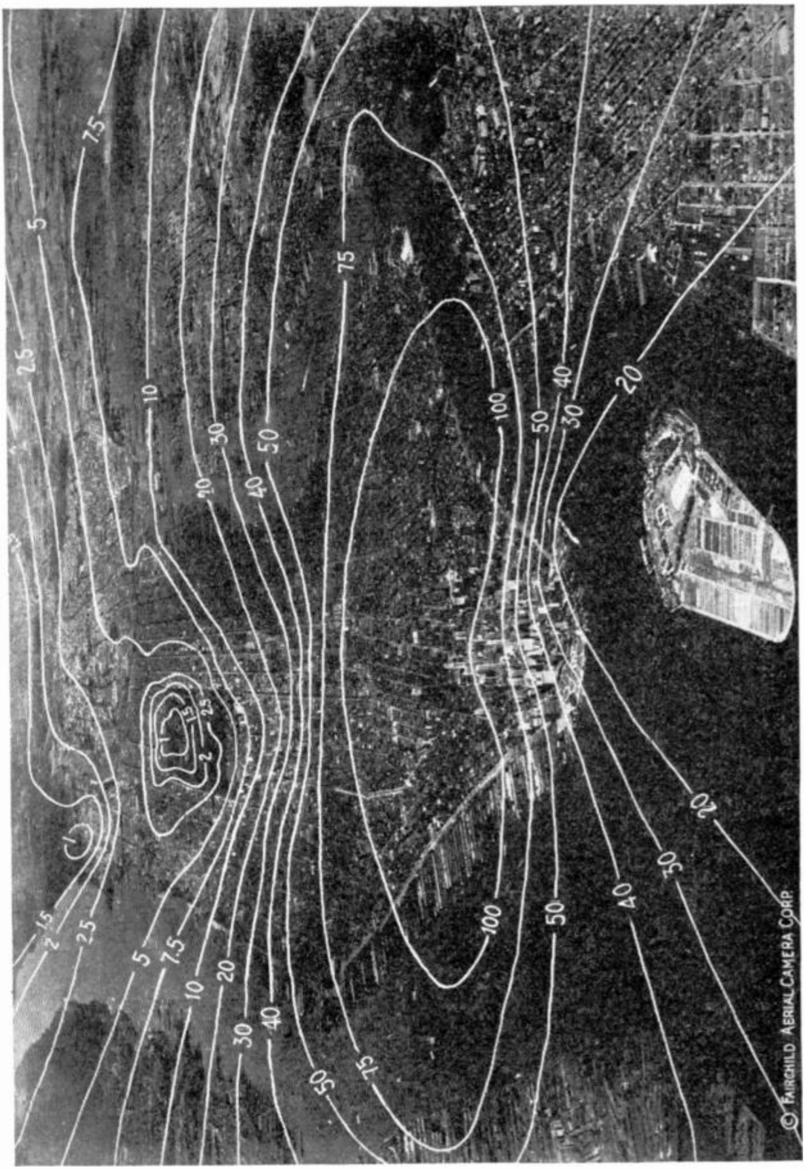


FIG. 319.—Radio field intensity in millivolts per meter for a transmitter located on the top of a building in lower Manhattan. New York City. Note the high rate of attenuation over the built-up districts, particularly the sky-scraper section, as compared with the attenuation over the water and suburban areas. (*Courtesy Fairchild Aerial Camera Corporation and the Institute of Radio Engineers.*)

particular, cities with tall steel buildings and extensive built-up areas attenuate the ground wave very severely, while forested and hilly districts have more attenuation than level prairie country. An example of ground-wave attenuation in a metropolitan region is given by Fig. 319. It will be observed that the attenuation is greatest for waves traveling over the sky-scraper district of lower Manhattan Island; is less, but still large, for waves traveling over the built-up region of upper Manhattan Island; and is least for the waves traveling over the water and the less densely populated areas to the right and left.

Fading—Quality Distortion.—Satisfactory service from a broadcast station requires not only that the ground wave be of sufficient intensity to override the interference, but also that the ground wave be considerably stronger than the sky wave. At receiving points where the ground and sky waves are of approximately equal strength the received signal is the vector sum of two nearly equal waves which travel paths of unequal length and hence are not necessarily in phase. As a consequence there will be places where the two waves are in phase and will combine to give a large resultant signal, while at other points there will be phase opposition and hence low signal strength. This results in the formation of an interference pattern the exact character of which depends upon the frequency and upon the height and make-up of the Kennelly-Heaviside layer. This interference pattern is extremely sensitive to conditions in the Kennelly-Heaviside layer and continually flits about over the surface of the earth in response to minor variations in the ionization of the upper atmosphere. As a consequence the two component waves may at one moment add up in phase at a particular location and yet cancel each other a few moments later, with the result that the received signal will vary in intensity (*i.e.*, will fade). The rapidity with which the received signal fades in and out depends upon the extent to which the conditions in the Kennelly-Heaviside layer fail to be constant, and at broadcast frequencies is commonly several times a minute. The fading tends to take place more rapidly as the frequency increases, since at a higher frequency the wave length is less, and this reduces the movement of the Kennelly-Heaviside layer that is required to change the length of the sky-wave path by one-half wave length.

The interference pattern produced by the combination of ground and sky waves is very sensitive to frequency since it requires only a small change in frequency to increase or decrease the difference in the length of path by one-half wave length, and hence to change a reinforcement in the interference pattern to a point of cancellation. Investigations on broadcast signals have shown that frequencies differing by as little as 250 cycles often fade in and out quite independently of each other. Since a modulated broadcast signal consists of a carrier and double side band that normally extend over a frequency range approximately 10,000 cycles wide, it is apparent that the different frequency components of a modulated wave will fade in and out quite independently of each other. This phenomenon is termed "selective fading" and causes the signal as received to differ from the transmitted signal, causing quality distortion.[1] In order to minimize selective fading it is essential that the carrier frequency of a broadcast transmitter be constant during modulation. Any frequency modulation present in the radiated wave causes additional fading and greatly increases the distortion of the received signal. All modulation systems now used in broadcast transmitters are hence of the modulated-amplifier type in order to give a carrier frequency that is independent of the modulation.

Broadcast signals fade somewhat even within the service area since some sky-wave energy returns to earth within a few miles of the transmitter.[2] The extent of this fading is small, however, and causes no serious effects where the ground wave is at least several times as strong as the sky wave.

At considerable distances from the transmitter the ground wave is practically extinguished by attenuation, so in order that fading may occur there must either be two sky waves reaching the receiver by different paths, or there must be a focusing action in the Kennelly-Heaviside layer which causes the waves to alternately concentrate at, and then to miss, the receiving point. It is probable that both of these effects are present to a certain extent, but the exact phenomena involved are not completely understood. The observed fact is that fading is nearly always present on broadcast signals that have traveled great distances, but that the most severe fading and greatest quality distortion occur at moderate distances from the transmitter.

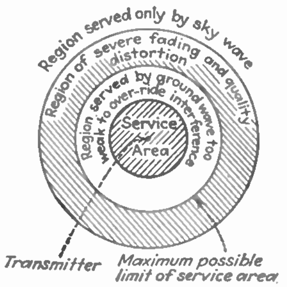


FIG. 320.—Schematic diagram illustrating receiving conditions existing around a broadcast transmitter.

Summary of Broadcast Characteristics.—The behavior of broadcast signals is summarized in the schematic diagram of Fig. 320. The service area surrounding the transmitter represents the region where the ground wave is much stronger than the sky wave and is powerful enough to override all ordinary interference. In this area there are negligible fading and quality distortion, and the signal strength shows no diurnal or seasonal variation. Immediately surrounding the service area there is a region in which the ground wave, while much stronger than the sky wave, does not override all noise. In this region the signals are free

[1] A very thorough study of fading phenomena on broadcast frequencies is to be found in the classic article of Ralph Bown, De Loss K. Martin, and Ralph K. Potter, Some Studies in Radio Broadcast Transmission, *Proc. I.R.E.*, vol. 14, p. 57, February, 1926.

[2] See Greenleaf W. Pickard, Short Period Variations in Radio Reception, *Proc. I.R.E.*, vol. 12, p. 119, April, 1924.

of fading and quality distortion and are of substantially constant strength at all times, but the reception is not always so satisfactory as in the service area. Next comes a region where the ground and sky waves are of nearly equal strength. This area is characterized by quality distortion, severe fading, and signals too weak to give satisfactory service. Finally there is the great outer region where reception, if possible at all, depends upon energy carried by the sky wave. Here the signals are characterized by moderate fading and quality distortion, and a strength that is never great and having large diurnal and seasonal variations. Reception in this outer region is usually impossible in the daytime and is most satisfactory during winter nights.

The size of the service area depends on the transmitter power, the frequency, and the character of the surface over which the ground wave travels, as has already been stated. As the power is increased the service area enlarges until it extends to the region where severe fading and quality distortion occur. This represents the greatest possible service area, since further increases in power, while giving stronger signals, do not alter the *relative* amplitudes of sky and ground waves. The only way in which the maximum possible service area can be increased is by the use of directive antennas that produce a strong ground wave and a weak sky wave.

The power required to make the actual service area approach the maximum possible service area is, to first approximation, independent of the frequency and is in the order of 500 kw; so it is much greater than the most powerful broadcast transmitter in operation.[1]

The attenuation of the sky wave depends upon the conditions in the Kennelly-Heaviside layer and is much greater by day than at night, and in summer than in winter. During winter nights the sky-wave attenuation is often almost negligible while in summer days it is quite common for the sky wave to attenuate more rapidly than the ground wave. As a result of these variations in the sky-wave attenuation, the location of the region of severe fading and quality distortion is constantly changing, and during summer days when the sky-wave attenuation is unusually high this zone may entirely disappear.

133. Propagation Characteristics of Short Waves (Frequency Range 1500 to 30,000 Kc).—At frequencies above 1500 kc the ground wave attenuates so rapidly as to be of no importance except for transmission over very short distances. Short-wave communication therefore depends upon the ability of the Kennelly-Heaviside layer to refract the high-frequency sky wave back to earth at the receiving point. As a result the strength of the signal received from a distant transmitter depends on the transmitted frequency, the height and electron density

[1] See P. P. Eckersley, The Calculation of the Service Area of Broadcast Stations, *Proc. I.R.E.*, vol. 18, p. 1160, July, 1930.

of the Kennelly-Heaviside layer, and the angle at which the transmitted waves enter the ionized region.

The most important features involved in the propagation of high-frequency sky waves can be understood with the aid of Fig. 321, which may be considered as showing hypothetical ray paths for sky waves of different frequencies leaving the transmitting antenna at different vertical angles. These diagrams assume a single ionized layer and while illustrating the general nature of short-wave propagation as now understood must not be taken too literally. Case A in Fig. 321 can be considered as representing the situation that exists when the frequency is low enough to be easily refracted by the Kennelly-Heaviside layer. It will be observed that energy radiated almost directly upward from the transmitting antenna is not refracted sufficiently to be bent earthward and so passes on through the ionized region to the space beyond. At a

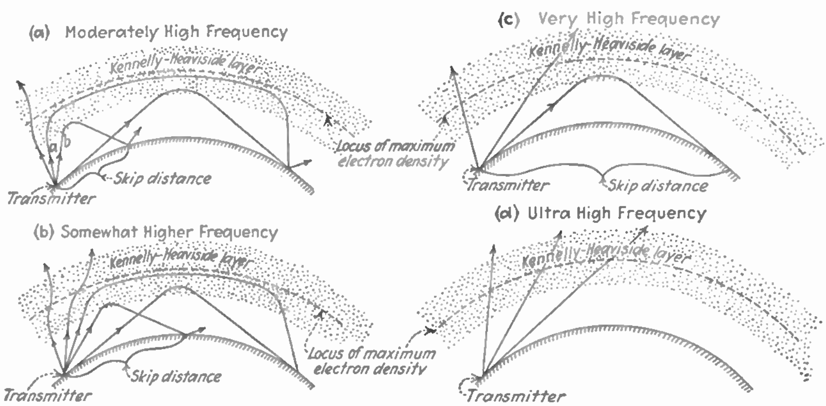


FIG. 321.—Hypothetical ray paths followed by sky waves of different frequencies. For the sake of clearness the height of the Kennelly-Heaviside layer is shown greatly exaggerated.

somewhat lower vertical angle one finds a ray (marked *a*) that is bent parallel with the earth's surface just before reaching the region of maximum ionization of the Kennelly-Heaviside layer. This ray travels a considerable distance almost parallel to the earth's surface because of the low rate of change of ionization with height along its path, but is ultimately bent downward into the lower part of the layer where the ionization gradient is greater, and is then directed earthward. A ray (marked *b* in Fig. 321*a*) leaving the transmitter at a still smaller vertical angle does not reach the region where the density of ionization changes slowly with height and so returns to the earth moderately close to the transmitter. All other sky waves, if they return to earth at all, do so at greater distances from the transmitter than ray *b*. The result is that in the region between the transmitter and the point at which ray *b* strikes the earth there will be a silent zone in which no signals will be received unless one is

Schelling - IRE June 1930 says 10 $\frac{\mu v}{m}$ ok for short wave transmitters on average
1 $\frac{\mu v}{m}$ ok sometimes

within the range of the rapidly attenuated ground wave. The distance from the transmitter to the point at which the first sky wave returns to earth is called the "skip distance," because the sky waves skip over it.

Case *B* in Fig. 321 can be considered as representing the conditions when the transmitted frequency is somewhat higher than in Case A. Since the refracting power of the Kennelly-Heaviside layer is less at the higher frequency, the critical angle of elevation at which the radiated rays first return to earth is less, and the skip distance is increased correspondingly. In this case the skip distance may be as great as several thousand kilometers.

As the frequency is increased still more (Case C) the refracting power of the Kennelly-Heaviside layer is so reduced that only rays striking the layer at a very glancing angle of incidence can be refracted sufficiently to be returned earthward, with the result that the skip distance is very great and the only part of the radiated energy that has any chance of being useful in communication is that which is radiated at a very low angle with respect to the horizon.

In Case D the frequency is so high that even those rays radiated horizontally from the transmitting antenna and striking the Kennelly-Heaviside layer at the smallest angle of incidence that is possible in view of the curvature of the earth's surface are not refracted sufficiently to be returned earthward. Long-distance communication at these frequencies is hence impossible.

The Critical Vertical Angle.—It is apparent from these considerations that all energy radiated at a vertical angle greater than a certain critical value passes on through the Kennelly-Heaviside layer into outer space and is lost. This critical angle becomes smaller as the frequency increases because of the low refracting power of the ionized region at high frequencies, and decreases with the electron density since the smaller the concentration of electrons the less will be the effect on the wave. Experience accumulated over a number of years indicates that under ordinary circumstances the highest frequency (shortest wave length) waves that can be refracted back to earth during the daytime have a length ranging from 8 to 11 meters. At night time the shortest waves that can be refracted by the Kennelly-Heaviside layer range from 15 to 25 meters in length.

The skip distance depends upon the critical vertical angle at which the sky wave is first refracted to earth, and is very nearly proportional to the height of the layer. The skip distance therefore increases as the frequency is raised, as the electron density in the Kennelly-Heaviside layer becomes less, and as the height of the layer is increased. The skip distance is ordinarily so sharply defined that during the sunset period when the height and electron density of the Kennelly-Heaviside layer are changing rapidly, the signals heard on a receiver located just at the edge of the skip distance may change from extreme loudness to complete inaudibility within a few minutes. Similarly when the conditions in the Kennelly-Heaviside layer are relatively constant and the receiver is at the edge of the skip distance a change in wave length of only 1 or 2 meters may cause the signal strength to vary from complete inaudibility to great loudness.[1]

Sky-wave Attenuation.—While the first requirement for short-wave radio transmission over great distances is that a sky wave be refracted earthward to the receiving point it is also necessary that the attenuation of the sky wave be low, for otherwise the energy reaching the receiving point will be too small to give a readable signal. Most of the attenuation of the sky wave takes place while the wave is traveling through the lower edge of the Kennelly-Heaviside layer and during the reflections at the surface of the earth. The attenuation in the space between the earth's surface and the lower edge of the ionized layer is negligible.

The factors governing the attenuation in the Kennelly-Heaviside layer are extremely complex. As the frequency is increased the attenuation tends to become less because the average velocity of the vibrating electrons is reduced, as explained in Sec. 130. At the same time the low refracting power of the Kennelly-Heaviside layer at high frequencies causes high-frequency waves to penetrate deeply into the layer and hence to travel considerable distances in the ionized region. The attenuation is affected by the angle of incidence of the ray, since the length of the ray path within the Kennelly-Heaviside layer and the average electron density encountered by the wave both vary with the angle of incidence. The loss also depends upon the height of the layer since at great elevations the gas pressure is low, making collisions between vibrating electrons and gas molecules less frequent.

When the sky waves strike the earth they are reflected upward because the earth is a relatively good conductor. Irregularities in the surface of the earth, however, cause a large fraction of the energy in the incident wave to be scattered and hence lost as far as radio communication is concerned. This scattering is particularly great in rough country but is appreciable even when the reflecting surface is the sea. The fraction of the energy lost in an earth reflection depends to a considerable extent upon the angle of incidence of the downward ray and becomes greater as the angle of incidence is reduced, until, with grazing waves, there is almost complete extinction.

Possible and Probable Wave Paths.—A consideration of the way in which the sky wave travels great distances shows that there are many paths which a wave might theoretically follow in traveling from a trans-

[1] Examples of such behavior at the edge of the skip distance are given by J. K. Clapp, Some Experiments in Short-distance Short-wave Radio Transmission, *Proc. I.R.E.*, vol. 17, p. 479, March, 1929.

mitter to a distant receiver. Some of the more important of these are illustrated in Fig. 322, which shows rays traveling by a single very long step path *a*, by a series of shorter long-step paths *b*, and by short-step paths *c* and *d*. The actual path which the main part of the energy arriving at the receiver follows is determined primarily by the relative attenuation along the different possible routes. Theoretical studies based on reasonable assumptions regarding the properties of the upper atmosphere and of the Kennelly-Heaviside layer indicate that the attenuation along the short-step path *d* is very great because of the numerous earth reflections and Kennelly-Heaviside layer refractions that are involved. The short-step path *c*, while suffering fewer reflections and refractions than *d*, is also severely attenuated in the Kennelly-Heaviside layer and is very nearly extinguished upon reflection at the earth's surface because of the grazing angle of incidence. The most probable paths appear to be the long-step paths such as *a* or *b*. At first sight it might

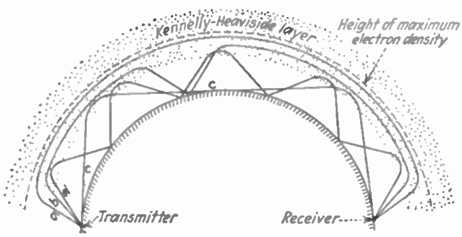


FIG. 322.—Possible routes by which short-wave signals can travel from transmitter to receiver. The height of the Kennelly-Heaviside layer has been greatly exaggerated to make the drawing clearer.

appear that since the long-step paths *a* and *b* travel nearly their entire length within the ionized region, their attenuation would be large. This is not necessarily the case, however, since these paths penetrate very deeply into the ionized region, and reach such an elevation above earth that the gas density is very low, with the result that the attenuation is correspondingly small.

Optimum Frequency.—The most satisfactory frequency for carrying on short-wave communication between two points depends upon the distance, the time of day, and the time of year. The best frequency ordinarily becomes higher as the distance is increased, is greater during the day than at night, and during the summer than in winter. Communication between two points is usually most satisfactory when the frequency employed approaches the highest frequency that will deliver a signal to the receiving point. Under these conditions the sky wave propagates to the receiver in the fewest steps and suffers a minimum of attenuation. While lower frequency signals will ordinarily be able to reach the receiver the signal strength will not be so great. The best frequency for carrying on short-wave communication varies from hour to hour, day to day, season to season, and even year to year, because of corresponding variations that take place in the Kennelly-Heaviside layer. In addition the best frequency depends upon the latitude of the transmitting and receiving stations, upon whether the transmission path lies in a north-south or east-west direction, and is affected by many special circumstances that may be either permanent or transient in character. In order to maintain relatively continuous communication over great distances by means of short waves it is always necessary to employ at least two frequencies. One of these is a high frequency for day service; the other a lower frequency for use at night. In addition, if continuous service is important, it is usually necessary to employ an intermediate or transition frequency that is used at those times during the day, usually near sunset and sunrise, when neither the day nor night frequency is satisfactory. Even then it is not always possible to maintain satisfactory contact between transmitter and receiver, and it is sometimes desirable to have available a fourth frequency to be used when the received signal is not satisfactory on either the normal day or night frequencies, or the normal transition frequency.

Experience accumulated as a result of a number of years' operation of various long-distance short-wave radio circuits indicates that the most satisfactory day frequency is in the order of 20,000 kc (15 meters), while the most useful night frequency is about 10,000 kc (30 meters). In addition a transition frequency approximately midway between these is desirable if fairly continuous contact is to be maintained. Thus the transatlantic short-wave radio telephone uses wave lengths between 15.1 and 16.3 meters for normal day communication, while the lengths for normal night service are between 30.4 and 32.7 meters for the various circuits, and the transition frequency ranges from 20.7 to 22.4 meters. In addition one of the short-wave circuits has available a wave length of 44.4 meters which finds its chief usefulness during winter nights. The short-wave telephone circuit from New York to Buenos Aires similarly makes use of day, transition, and night waves, having lengths of 14.0, 18.4, and 28.4 meters, respectively. These figures can be taken as representative of typical short-wave long-distance communication circuits.[1]

[1] Data on short-wave communication across the north Atlantic, showing the character of transmission that is obtained at different hours of the day, and at different seasons of the year, for frequencies lying in the range 6000 to 27,000 kc are given by C. R. Burrows, The Propagation of Short Radio Waves over the North Atlantic, *Proc. I.R.E.*, vol. 19, p. 1634, September, 1931.

An extensive discussion of the most satisfactory frequencies for carrying on short-wave communication at different hours of the day and at different seasons of the year over a wide variety of radio circuits radiating in different directions from Schenectady, N. Y., is given by M. L. Prescott, The Diurnal and Seasonal Performance of High-frequency Radio Transmission over Various Long-distance Circuits, *Proc. I.R.E.*, vol. 18, p. 1797, November, 1930.

Short-wave communication over distances up to several thousand miles is most satisfactorily carried on with frequencies somewhat lower than those preferred for greater distances.[1] This is because the lower frequency waves have a smaller skip distance and hence give a more reliable signal at moderate distances.

The above discussion of short-wave communication is intended to give a general picture of the most outstanding features involved and must not be taken too literally because the phenomena involved are not completely understood. Thus it is not known for certain whether the Kennelly-Heaviside region consists of a single or double layer, and the nature of the electron distribution within the layer is largely a matter of conjecture. In view of our limited knowledge the picture that has been drawn of the mechanism of short-wave communication is purely qualitative, although in general agreement with the observed facts.

Difference between East-west and North-south Propagation.—Short waves propagated over great distances in an east-west direction differ markedly in their behavior from waves traveling long distances in a north-south direction. This is because of the way in which the distribution of sunlight varies along the transmission path at any one time, and because of seasonal differences that exist between points on opposite sides of the equator. Thus when communication is carried on between points having the same longitude but on opposite sides of the equator the sunlight is more or less uniformly distributed along all parts of the path, but the seasons at the two terminals are opposite. On the other hand the distribution of sunlight along a great circle path lying between two points on the same latitude is non-uniform, so that one part of the path can be in sunlight while another part is in darkness. Under such a condition it is difficult to find a frequency that will propagate satisfactorily over the entire distance. Experience indicates that north-south transmission across the equator is more reliable and easier to maintain continuously than is communication over a like distance in an east-west direction. In particular when the great circle route between the terminals passes over the polar regions the transmission is often unsatisfactory.

Effect of the Earth's Magnetic Field.—The earth's magnetic field causes short waves which are refracted by the Kennelly-Heaviside layer to have their plane of polarization rotated, and to be split into several components which travel with different velocities through the ionized region. As a result of this, and because of the fact that there are several paths along which energy may travel to a given receiving point, it will ordinarily be found that at considerable distances from the transmitter short-wave signals will have both horizontally and vertically polarized

[1] Thus see N. H. Edes, Some Experiences with Short-wave Wireless Telegraphy, *Proc. I.R.E.*, vol. 18, p. 2011, December, 1930.

components. In fact the polarization of a short-wave signal observed at a distant point appears to have no relation whatsoever with the polarization of the transmitting antenna. Since the horizontally and vertically polarized components of the received signal are not necessarily in phase, the high-frequency radio waves will in general be elliptically polarized, *i.e.*, will consist of vertically and horizontally polarized waves of different time phase superimposed on each other.

Fading.—Short-wave signals observed at a distant receiving point nearly always fade severely. This is because sky waves of appreciable strength can ordinarily reach the receiver along several transmission paths of different lengths. The received signal is hence the vector sum of several components, and any slight change in the Kennelly-Heaviside layer will alter the relative phase of the components. Short-wave fading is normally much worse than at broadcast frequencies since when the wave length is small a given change in the condition of the Kennelly-Heaviside layer has a greater effect on the phase of the received signals.

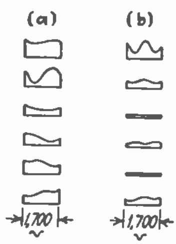


Fig. 323.—Typical fading patterns observed in short-wave transatlantic communication. The different figures in each column show transmission conditions over a 1700-cycle frequency band at successive moments.

The fading of short-wave signals is selective with respect to frequency, exactly as is the case with broadcast signals. Extensive investigations of selective fading on the transatlantic short-wave telephone circuit show that frequencies differing by 500 to 1000 cycles ordinarily fade quite independently of each other, and that at times frequencies differing by as little as 100 cycles fade independently.[1] Figure 323 shows the relative signal strength as a function of frequency over a band of frequencies 1700 cycles wide. It will be observed that at any instant the different frequencies within this range are transmitted with widely differing efficiencies, and that the signal received on any particular frequency varies greatly from moment to moment. Investigation of these fading patterns shows that under ordinary circumstances they can be accounted for on the assumption that there are not more than three paths by which energy is delivered to the receiving point.

The horizontally and vertically polarized components of short-wave signals fade independently of each other, and the fading patterns on antennas spaced 10 wave lengths or more apart also show no similarity. The fading on adjacent receiving antennas having different directivities in a vertical plane tends to be independent, while signals received on two antennas with the same vertical but differing horizontal directivity are usually in synchronism. It thus appears that the various paths

[1] See R. K. Potter, Transmission Characteristics of a Short-wave Telephone Circuit, *Proc. I.R.E.*, vol. 18, p. 581, April, 1930.

that the energy follows in arriving at the receiver differ in vertical direction, but all have the same bearing in a horizontal plane

Amount and Character of Directivity Desirable in Transmitting and Receiving Antennas.—The accumulated experience of many investigators indicates that in short-wave communication over great distances most satisfactory results are obtained when the transmitting antenna gives a high directivity in a horizontal plane and also concentrates most of the radiation at a low vertical angle. The reason for this appears to be that low-angle radiation reaches a distant receiver with the minimum number of steps and so is attenuated least. It is not desirable to have either excessive vertical or horizontal directivity, however, since the route followed by the sky wave in traveling to a distant receiving point is not always the same, and if extreme directivity is employed there will be found occasions when the received signal will be weaker than when a non-directional antenna is employed. The amount of directivity that can be usefully employed is usually much greater in a horizontal than a vertical plane, and in fact high-angle radiation is sometimes surprisingly effective for communication over great distances.[1]

In the reception of short-wave radio signals that have traveled great distances the receiving antenna is preferably one that gives considerable directivity in a horizontal plane and but very little vertical directivity. This is because the bearing of the waves seldom varies by more than 5° or 10° from the bearing of the great circle route, while the direction of arrival in a vertical plane will range from horizontal to vertical angles of 70° to 80°.

Echo and Multiple Signals.—The fact that short waves can travel to a receiver along paths of different lengths gives rise to echo signals of which a number of kinds have been observed. Thus if a short wave train lasting perhaps 10^{-4} sec. is sent out from a transmitter and recorded at the receiving point on an oscillogram the results will often be as in Fig. 324, in which the same impulse is shown as being received a number of times. When the receiver is within the range of the ground wave, the first received impulse travels along the ground, while the second impulse has reached the receiver by way of the Kennelly-Heaviside layer and so arrives several thousandths of a second later as a result of its longer path. The remaining echoes are apparently the result of multiple-step paths between the earth and the Kennelly-Heaviside layer.

Echo signals which arrive at the receiving point several hundredths of a second after the main signal are also observed at times. The source of such echo signals is difficult to establish, but in at least one instance

[1] Thus satisfactory communication has been maintained at 11 meters between Germany and South America when the radiation was directed at a vertical angle of 80°. See A. Meissner, Directional Radiation with Horizontal Antennas, *Proc. I.R.E.*, vol. 15, p. 928, November, 1927.

they have been found to represent scattered radiation reflected from points several thousand miles distant.[1]

Multiple signals are sometimes observed as a result of waves which have traveled around the earth. The possible paths for such around-the-world multiple signals are shown in Fig. 325, where it is seen that the signals may reach a distant receiving point along either the short or long great circle path, and that if the waves traveling in either of these directions make more than one complete circle of the earth they will be heard again. Multiple round-the-world signals occur regularly on certain short-wave circuits at definite times each year. There are cases on record where the same signal has been heard five times at a distant receiving

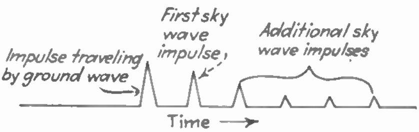


FIG. 324.—Typical oscillogram of received signal when transmitter within ground-wave range sends out a short impulse.

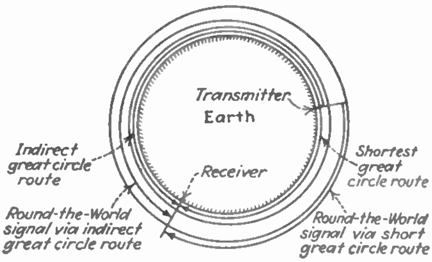


FIG. 325.—Paths followed by multiple and round-the-world signals.

point, while multiple signals repeated once or twice are relatively common.[2] In order that double signals may reach the receiver by propagation along both long and short great circle paths it is necessary that the conditions in the Kennelly-Heaviside layer be roughly the same along both routes. Otherwise a frequency that is satisfactorily propagated along one of the two paths will not be able to reach the receiver along the other path. The same requirement must also be fulfilled in the case of round-the-world signals. As a result multiple and round-the-world signals are observed only when the part of the great circle path that lies in darkness is experiencing summer, or when the great circle path coincides very closely with the twilight zone.[3] The time delay of round-the-world signals is in the order of $\frac{1}{7}$ sec. and is so great that when these signals occur it is necessary to reduce the speed of transmission to an extremely low value.

[1] See A. H. Taylor and L. C. Young, Studies of Echo Signals, *Proc. I.R.E.*, vol. 17, p. 1491, September, 1929; also see E. Quäck and H. Mögel, Short-range Echoes with Short Waves, *Proc. I.R.E.*, vol. 17, p. 824, May, 1929.

[2] See E. Quäck and H. Mögel, Double and Multiple Signals with Short Waves, *Proc. I.R.E.*, vol. 17, p. 791, May, 1929.

[3] See E. Quäck and H. Mögel, *loc. cit.*; A. H. Taylor and L. C. Young, Studies of High Frequency Radio-wave Propagation, *Proc. I.R.E.*, vol. 16, p. 561, May, 1928.

Echo signals having a time lag of several seconds have been reported a number of times, and several well-authenticated cases have been observed in which the time lag was several minutes. The cause of such echoes has not been definitely established, but theoretical work by P. O. Pedersen[1] indicates that retardations up to 10 sec. could be accounted for by low group-velocity propagation in the Kennelly-Heaviside layer, while signals with greater retardations can be accounted for only by waves which have traveled great distances in the empty space outside of the earth's atmosphere and which by a fortunate combination of circumstances have been finally reflected back to earth by ionized regions either within the influence of the earth's magnetic field or in the vicinity of the sun. These echo signals of long delay, while of extreme theoretical interest, are of little practical importance because they occur so rarely.

Communication between Aircraft and Ground.—Short-wave communication between airplane and ground possesses certain features not present in other types of short-wave communication. The important object in such communication is to maintain satisfactory contact between plane and ground over moderate distances. Since it is necessary to use a small antenna on the airplane, a high transmission frequency will be more efficiently radiated from the plane antenna, but at the same time low frequencies are more suitable for short-distance communication. The compromise most commonly employed uses approximately 6000 kc for daytime work, and approximately 3000 kc at night. Considerable attention must be paid to the directional characteristics in a vertical plane of the radiation from short-wave antennas employed in aircraft service in order to avoid skip-distance effects.[2]

134. Use of Radio Waves in Investigations of the Upper Atmosphere. Since the way in which radio signals propagate is dependent on the Kennelly-Heaviside layer, it is possible by working backward from observed propagation characteristics of radio signals to obtain information regarding the probable nature of the Kennelly-Heaviside layer and hence indirectly to obtain data on the composition of the upper atmosphere.

Pulse Experiments.—One method commonly used in making such investigations consists in transmitting a short-wave train lasting about 10^{-4} sec. and taking a record on an oscillogram of the signal as received at a point within range of the ground wave. Since the wave that reaches

[1] P. O. Pedersen, Wireless Echoes of Long Delay, *Proc. I.R.E.*, vol. 17, p. 1750, October, 1929.

[2] For additional discussion regarding the problem of obtaining satisfactory signal strength in air transport communication see R. L. Jones and F. H. Ryan, Air Transport Communication, *Trans. A.I.E.E.*, vol. 49, p. 187, January, 1930; F. H. Drake and Raymond M. Wilmotte, On the Daylight Transmission Characteristics of Horizontally and Vertically Polarized Waves from Airplanes, *Proc. I.R.E.*, vol. 17, p. 2242, December, 1929.

the receiver after refraction by the Kennelly-Heaviside layer must travel an appreciably longer path to reach the receiver than does the ground wave, there will be a time interval between the pulses arriving over the two routes. The length of this time interval is a measurement of the difference in path lengths and can be used to estimate the height of the layer.[1] In interpreting the results given by the pulse records it must be remembered that the pulse travels along a curved path while in the Kennelly-Heaviside layer, as shown in Fig. 326, and that the group velocity in the ionized region is less than the velocity of light. If it is assumed that the sky wave travels along the triangular path TAR of Fig. 326 with the velocity of light, the observed time delay of the sky-wave impulse corresponds to a "virtual height" in the Kennelly-Heaviside layer somewhat greater than the actual height. When this virtual height is known it is possible, by making reasonable assumptions about the electron distribution in the upper atmosphere, to estimate with fair accuracy the actual path followed by the wave.[2] This pulse method of investigating the height of the Kennelly-Heaviside layer can be used only at moderately high frequencies, since at low frequencies the pulses are too short compared with the time of the cycle to give a satisfactory wave train, while at extremely high frequencies the skip distance is much greater than the range of the ground wave.

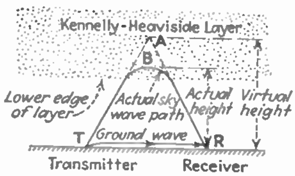


FIG. 326.—Diagram showing relation of actual and virtual height reached by a ray refracted by the Kennelly-Heaviside layer.

Miscellaneous Experimental Methods.—Appleton and Barnett[3] have investigated the Kennelly-Heaviside layer by varying the transmitted frequency and observing the variations that occur in the signal strength at a point within range of the ground wave. Because of the different path lengths of sky and ground waves the relative phase of the two component waves reaching the receiver will alternate between the same phase and phase opposition as the frequency is varied, and the increment in frequency that is required to change the relative phases by 180° can be used to estimate the layer height. Theoretical analysis indicates that the layer height obtained in this way is the virtual height given by the pulse method of Breit and Tuve. This method of determining the

[1] This method was first proposed by Breit and Tuve. See G. Breit and M. Tuve, A Test of the Existence of the Conducting Layer, *Phys. Rev.*, vol. 28, p. 554, September, 1926.

[2] An example of such a calculation is given by G. W. Kenrick and C. K. Jen, Measurements of the Height of the Kennelly-Heaviside Layer, *Proc. I.R.E.*, vol. 17, p. 711, April, 1929.

[3] See E. V. Appleton, Some Notes on Wireless Methods of Investigating the Electrical Structure of the Upper Atmosphere, *Proc. Phys. Soc.*, vol. 41, Part II, p. 43, December, 1928.

height of the Kennelly-Heaviside layer can be used at low radio frequencies but is not entirely satisfactory at high frequencies because of complications introduced by fading.

Appleton and Barnett have also made use of the angle of incidence of the down-coming sky wave to determine the height of the Kennelly-Heaviside layer. It is apparent from examination of Fig. 326 that if the angle of incidence, with which the sky wave strikes the earth, and the distance between receiver and transmitter are known, then the virtual height of the Kennelly-Heaviside layer can be found by simple triangulation.

Still another way of determining the height of the Kennelly-Heaviside layer consists of observing the variations in field intensity as the distance between transmitter and receiver is varied.[1] Thus Hollingworth has found that at moderate distances from long-wave transmitters the signal strength alternately decreases and increases as the distance between transmitter and receiver is varied. These variations in signal strength result from alternate reinforcement and cancellation between sky and ground waves. A somewhat similar method of investigating the Kennelly-Heaviside layer is due to Mirick and Hentschel and consists in observing at a fixed point the variations in strength of the signals received from an airplane in flight.[2]

A method of detecting slight changes in the position of the Kennelly-Heaviside layer (or the electron distribution within this layer) has been devised by Hafstad and Tuve. This consists in locating a receiver close to a crystal-controlled transmitter and supplying the receiver with a small amount of energy directly from the crystal. The power amplifier of the transmitter is then modulated so as to send out pulses of very short duration. By arranging the equipment so that the phase of the oscillation generated by the crystal is unchanged between impulses, slight variations in the phase of the pulses received after refraction by the Kennelly-Heaviside layer (and hence slight changes in path length) can be observed by noting the rate at which the received echo signal changes from in-phase to phase opposition with the oscillations generated by the crystal.[3]

135. Propagation of Ultra-high Frequency Waves.—Waves of frequencies above about 40,000 kc are not refracted sufficiently by the Kennelly-Heaviside layer to be returned to earth except under very

[1] See J. Hollingworth, Propagation of Radio Waves, *Jour.I.E.E.* (London), vol. 64, p. 579, May, 1926.

[2] C. B. Mirick and E. R. Hentschel, A New Method of Determining Height of the Kennelly-Heaviside Layer, *Proc. I.R.E.*, vol. 17, p. 1034, June, 1929.

[3] L. C. Hafstad and M. A. Tuve, An Echo Interference Method for the Study of Radio-wave Paths, *Proc. I.R.E.*, vol. 17, p. 1786, October, 1929; G. Breit, The Significance of Observations of the Phase of Radio Echoes, *Proc. I.R.E.*, vol. 17, p. 1815, October, 1929.

unusual and highly transient conditions. These waves tend to travel along straight-line paths and are quickly absorbed by the earth, or by buildings, trees, etc., but suffer negligible attenuation when traveling through the atmosphere. As a result satisfactory communication can be maintained with frequencies above about 40,000 kc only when the sending and receiving points are in sight of each other. The curvature of the earth hence makes it necessary to locate the transmitter and receiver at high points, such as on mountain peaks, if communication is to be carried on over appreciable distances. The relationship between the heights h_s and h_r of sending and receiving antennas, respectively, above the surrounding country (as measured in feet) to the maximum possible range is given by the equation

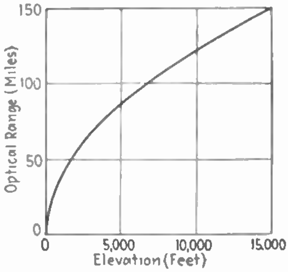


FIG. 327.—Maximum possible optical range between an elevated point and the surface of the earth. The maximum possible optical range between two elevated points is the sum of the ranges as obtained in the above diagram for the two heights involved.

$$\text{Maximum possible range in miles} = 1.225(\sqrt{h_s} + \sqrt{h_r}) \quad (173)$$

The results of this equation are shown graphically in Fig. 327. Experience indicates that there is enough bending of the rays around the curvature of the earth to make the practical range exceed the value given in Eq. (173) and Fig. 327 by perhaps 10 per cent.[1]

Propagation of ultra-high-frequency waves over distances so short that the curvature of the earth is unimportant is most satisfactory when the straight-line path between transmitting and receiving antennas is a reasonable distance above the earth and clears all objects, such as trees, buildings, hills, etc.

When a straight-line path is available between transmitter and receiver, ultra-high-frequency waves show no fading, no diurnal or seasonal variations in strength, and are substantially independent of weather conditions. Such waves also maintain their plane of polarization under ordinary conditions. Fog, rain, clouds, humidity, dust particles, etc., have little effect on the attenuation provided the wave length is greater than about 5 cm. Waves shorter than about 5 cm, however, suffer excessive attenuation from fog, etc., as a result of scattering and absorption.[2]

[1] Abraham Esau and Walter M. Hahnemann, Report on Experiments with Electric Waves of about Three Meters: Their Propagation and Use, *Proc. I.R.E.*, vol. 18, p. 471, March, 1930; also see H. H. Beverage, H. O. Peterson, and C. W. Hansell, Application of Frequencies above 30,000 kc to Communications Problems, *Proc. I.R.E.*, vol. 19, p. 1313, August, 1931.

[2] See J. A. Stratton, The Effect of Rain and Fog on the Propagation of Very Short Waves, *Proc. I.R.E.*, vol. 18, p. 1064, June, 1930; E. Karplus, Communication with Quasi-optical Waves, *Proc. I.R.E.*, vol. 19, p. 1715, October, 1931.

136. Relation of Solar Activity and Meteorological Conditions to the Propagation of Radio Waves.—The fact that the propagation of all except the very shortest radio waves depends to a marked extent upon the conditions in the Kennelly-Heaviside layer would lead one to expect some relation to exist between solar activity and meteorological conditions on the one hand, and wave propagation on the other, and this is the case.

The most striking relation of this type is the abnormal propagation characteristics of radio waves that always accompany magnetic storms. A magnetic storm is characterized by a rapid and excessive fluctuation of the earth's magnetic field that begins almost simultaneously over the entire earth with full intensity and then gradually subsides in 3 or 4 days.[1] Magnetic storms occur more or less irregularly, although showing a tendency to reoccur at intervals of 27 days, which is the period of rotation of the sun.

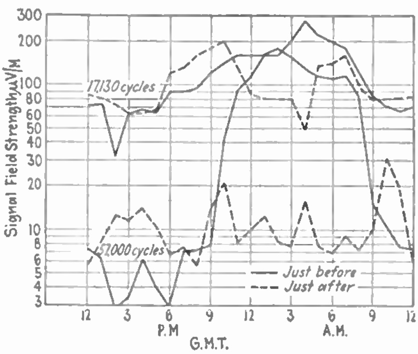


FIG. 328.—Effect of magnetic storm on low-frequency waves, showing how the day field strength is increased and the night field strength reduced by the magnetic storm.

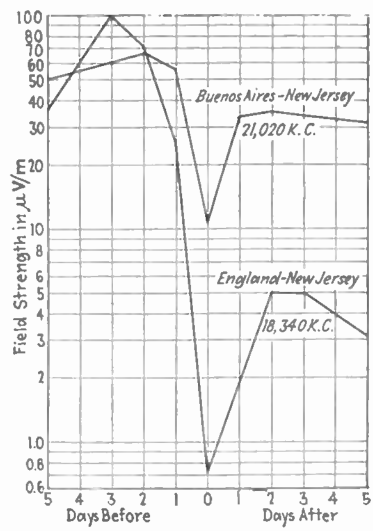


FIG. 329.—Effect of magnetic storm on the propagation of short waves. The drop in field strength is much greater on the east-west than on the north-south circuit.

During a magnetic storm the daytime field strength on long waves is increased above normal, the sunset drop in signal intensity disappears, and the night field is subnormal. These effects are illustrated in Fig. 328 and are more pronounced as the frequency of transmission is increased.

The propagation of high-frequency radio waves is very adversely affected by magnetic storms, particularly when the transmission path passes near the polar regions. Thus a severe magnetic storm ordinarily

[1] An excellent summary of the principal solar phenomena that are of importance in wave propagation is given by Clifford N. Anderson, Correlation of Long-wave Trans-atlantic Radio Transmission with Other Factors Affected by Solar Activity, *Proc. I.R.E.*, vol. 16, p. 297, March, 1928.

makes the short-wave circuits across the north Atlantic completely inoperative for a period of several days and causes subnormal field strengths for a week or more. The adverse effect is much less, however, when the entire transmission path is nowhere near the poles. This is strikingly brought out by Fig. 329, which shows the signal strength over the New York-London and the New York-Buenos Aires short-wave circuits during and after the same magnetic storm.

In addition to the abnormalities of wave propagation that are associated with magnetic storms there also appears to be some relation between sun spots and radio-wave propagation. Thus yearly averages of field

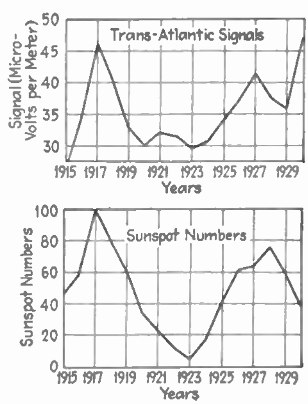


FIG. 330.—Diagram showing the close correlation between sun-spot numbers and yearly average of long-wave signal strength.

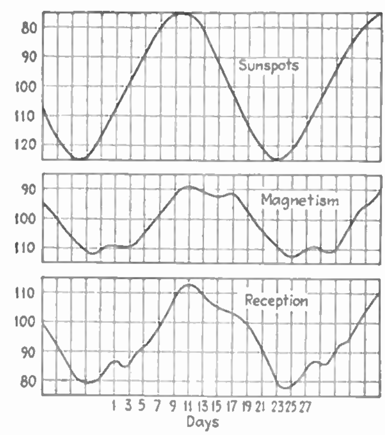


FIG. 331.—Average of sun-spot numbers, magnetic character of days, and radio reception on 1330 kc for eight solar rotations. These curves have been smoothed by the use of a 13-day moving mean.

strength of long-wave signals arriving from distant transmitters correlate surprisingly well with yearly averages of sun-spot numbers, as is brought out by Fig. 330.[1] In addition to the 11-year sun-spot cycle apparent in Fig. 330 there is also a 27.3-day sun-spot cycle corresponding to the period of rotation of the sun. While day-to-day signal strength appears to be independent of day-to-day variations in sun-spot activity, Pickard has found that when the signal strength over a number of successive 27.3-day cycles is averaged a very decided relation exists, as is illustrated in Fig. 331.[2]

[1] See L. W. Austin, Long-wave Receiving Measurements at the Bureau of Standards in 1930, *Proc. I.R.E.*, vol. 19, p. 1766, October, 1931.

[2] See Greenleaf W. Pickard, Correlation of Radio Reception with Solar Activity and Terrestrial Magnetism, part II, *Proc. I.R.E.*, vol. 15, p. 749, September, 1927.

Similar relations appear to exist between radio reception on the one hand and magnetic character of the days, the 6- and 15-month periods in solar activity, and meteoric showers.[1]

Some relation appears to exist between the strength of received signal, atmospheric temperature, barometric pressure, and weather conditions, although the correlation is not high and tends to be obscured by other influences.[2]

It is probable that the variations in wave propagation associated with magnetic storms, sun-spot numbers, magnetic character of the day, and meteorological conditions are not related in a cause-and-effect manner, but rather, that both wave propagation and these various meteorological and solar factors are the result of a common and as yet unknown force.

137. Noise and Static.—The output of a sensitive radio receiver nearly always contains a background of rumbles, crashes, rattles, etc., that disappears when the antenna is disconnected. This noise is the result of voltages induced in the antenna by either natural or man-made sources of interference and is often of sufficient magnitude to be the practical factor determining the minimum usable signal.

The principal sources of man-made noise are high-voltage power lines, ignition systems of airplane and automobile motors, brush motors, and electrical appliances. The noise from these sources is distributed throughout the entire frequency range used in radio communication and is always strong in cities and particularly in industrialized areas. There is very little that can be done to minimize the general level of man-made noise other than to suppress unusually bad localized sources of interference. Satisfactory reception in populated areas is hence obtained only from strong signals.

Radio waves generated by natural causes are referred to as static and produce the familiar clicks, rumblings, crashes, etc., sometimes heard in all radio receivers. Static normally has its origin in thunderstorms and similar natural electrical disturbances and is in the form of impulses, the energy of which is distributed throughout the range of useful radio frequencies. The energy level of static decreases as the frequency increases, ordinarily being very great at the lower radio frequencies and so small as to be unimportant at ultra-high frequencies. Since

[1] Greenleaf W. Pickard, Note on the Fifteen-month Period in Solar Activity, Terrestrial Magnetism, and Radio Reception, *Proc. I.R.E.*, vol. 19, p. 353, March, 1931; "A Note on the Relation of Meteor Showers and Radio Reception," *Proc. I.R.E.*, vol. 19, p. 1166, July, 1931.

[2] Thus see Greenleaf W. Pickard, Some Correlations of Radio Reception with Atmospheric Temperature and Pressure, *Proc. I.R.E.*, vol. 16, p. 765, June, 1928; L. W. Austin and I. J. Wymore, Radio Signal Strength and Temperature, *Proc. I.R. E.*, vol. 14, p. 781, December, 1926; R. C. Colwell, Weather Forecasting by Signal Radio Intensity, *Proc. I.R.E.*, vol. 18, p. 533, March, 1930.

static is fundamentally a radio signal containing frequency components distributed over a wide range of frequencies, the static within any frequency range is propagated over the earth in the same way as ordinary radio signals of the same frequency. Thus static impulses travel great distances under favorable conditions, arrive at a receiving point from a definite direction, and possess diurnal and seasonal variations in intensity as a result of corresponding variations in wave propagation.

It has been found that some of the static observed in the northern hemisphere is produced by local thunderstorms, but that a surprisingly large part of static interference originates in the tropics. Thus direc-

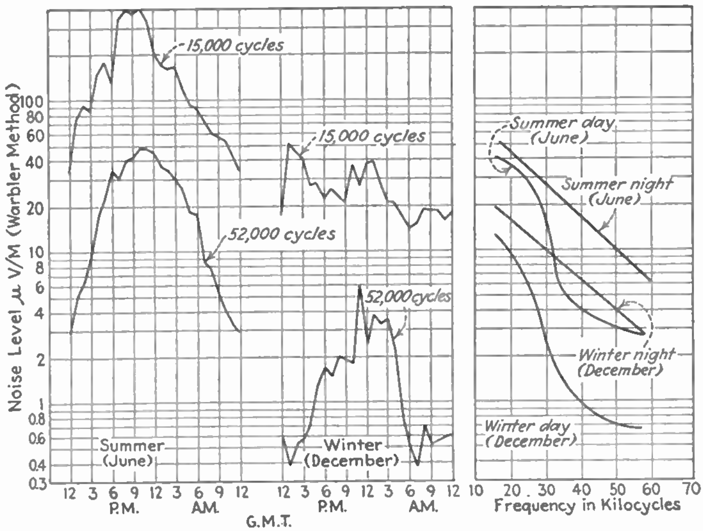


FIG. 332.—Curves summarizing behavior of low-frequency static received in Maine.

tional observations on long-wave static in Maine give a general southwesterly origin pointing toward the Gulf of Mexico and Texas, while similar observations in Europe indicate sources in Africa.[1] It has also been found that land areas, particularly mountains, are usually the most important sources of static, and that the static is worst in the summer season.

Low and Moderately Low Frequency Static.—At low radio frequencies the static intensity is high because most of the energy of a static impulse is concentrated on the lower radio frequencies, and because at low fre-

[1] The connection between storm areas and static is strikingly brought out by observations made in Maine by engineers of the American Telephone and Telegraph Company, in which it was found that storms within several thousand miles could be readily followed for days by making directional observations on static. See A. E. Harper, Some Measurements on the Directional Distribution of Static, *Proc. I.R.E.*, vol. 17, p. 1214, July, 1929; S. W. Dean, Correlation of Directional Observations of Atmospherics with Weather Phenomena, *Proc. I.R.E.*, vol. 17, p. 1185, July, 1929.

quencies radio waves propagate great distances with small attenuation.[1] The intensity of long-wave static becomes greater as the frequency is reduced, and in northern latitudes is greater at night and summer than in the daytime and winter, respectively. The curves of Fig. 332 summarize the more outstanding features of long-wave static as observed in the northern hemisphere.[2]

At moderate frequencies, such as those in the broadcast range, a large fraction of the static observed during the day is of local origin because of the poor propagation of such frequencies during daylight hours. At night however the lower attenuation causes static impulses of distant origin to be heard, with the result that the noise level is ordinarily greater at night than in the daytime.

Short-wave Static.—The static intensity at short waves, *i.e.*, frequencies from 6000 to 30,000 kc, is much less than at lower frequencies, and during a good part of the time is of the same order of magnitude as the noise level of a typical radio receiver. Investigations of high-frequency static show evidences of localized sources similar to those observed at lower frequencies, and it appears that a large fraction of high-frequency static represents waves that have traveled considerable distances. As a result, high-frequency static at any given wave length shows diurnal and seasonal variations in intensity which are identical with the corresponding variations in the strength of long-distance short-wave signals of the same frequency. This is well illustrated by the curves of Fig. 333, which show that during the day the static is strongest on the frequency best suited for daylight transmission (a high frequency),

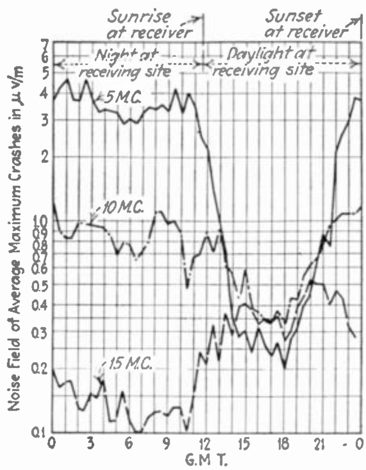


FIG. 333.—Diurnal variation of short-wave static received at Doria, Fla., in December. Note how the noise level on each frequency is highest at the part of the day most favorable for long-distance propagation of waves of the same frequency.

[1] Thus the same static impulse has been observed at Berlin, Germany, and the Hawaiian Islands. See M. Baumler, Simultaneous Atmospheric Disturbances in Radio Telegraphy, *Proc. I.R.E.*, vol. 14, p. 765, December, 1926; also see S. W. Dean, Long-distance Transmission of Static Impulses, *Proc. I.R.E.*, vol. 19, p. 1660, September, 1931.

[2] Excellent discussions of long-wave static are given by L. W. Austin, The Present Status of Radio Atmospheric Disturbances, *Proc. I.R.E.*, vol. 14, p. 133, February, 1926; Lloyd Espenschied, C. N. Anderson, and Austin Bailey, Trans-atlantic Radio Telephone Transmission, *Proc. I.R.E.*, vol. 14, p. 7, February 1926; and R. A. Watson Watt, Present State of Knowledge of Atmospherics, Abstract in *Exp. Wireless and Wireless Eng.*, vol. 5, p. 629, November, 1928.

while at night a reversal of the situation takes place because a low frequency is best for short-wave night transmission.[1]

The intensity of static interference at frequencies above 40,000 kc is very small, and it appears that nature, like man, has found difficulty in generating waves of such high frequencies. Furthermore, whatever ultra-high frequency static is produced has a very short range.

Means of Overcoming Static.—The only successful means that have been found for minimizing static interference are to make the frequency band to which the receiver responds as narrow as possible and to use directional receiving antennas. Since static energy is more or less uniformly distributed through the frequency spectrum it is obvious that the amount of static interference picked up by a receiver is almost directly proportional to the frequency range to which the receiver responds. This range should therefore be no wider than is necessary to accommodate the side bands of the desired signal.[2]

Directional receiving systems are of advantage in eliminating static interference when the interference and the desired signal arrive from different directions. It has been found that in east-west transmission at high latitudes very great gains in signal-to-noise ratio are nearly always realized by using directive receiving antennas because the major sources of static heard on almost any frequency are to the south and so are not in the direction of the transmitting station. After the full benefits of narrow-band reception and directional receiving antennas have been realized, the only remaining possibility for improvement in the signal-to-static ratio is to increase the transmitter power in order to override the interference, or to move the receiver to a location where the static intensity is less.

Static Eliminators.—Much effort has been expended in attempting to devise "static eliminators," but all of these devices have been failures. The reason is that static is a radio wave similar in all respects to the signal that is to be received, and any balancing scheme that balances out static will also balance out the received signal. In cases where apparent improvement has been obtained it can be shown that this is the result of improved selectivity (*i.e.*, a narrowing of the width of the response band of the receiver) rather than because of any balancing action that is present. In code reception the width of the response band of ordinary receivers is much greater than the minimum necessary to accommodate

These statements based on a

[1] ~~The result of a~~ very thorough investigation of high-frequency noise is reported by R. K. Potter, High-frequency Atmospheric Noise, *Proc. I.R.E.*, vol. 19, p. 1731, October, 1931.

[2] See John R. Carson, Selective Circuits and Static Interference, *Trans. A.I.E.E.*, vol. 43, p. 789, 1924. This is a classical paper in which it is shown that if static is a random series of impulses, then the amount of static energy absorbed by a receiver is directly proportional to the frequency range to which the receiver responds.

the side bands represented by the dots and dashes, so that considerable improvement in signal-to-static ratio of code signals can ordinarily be obtained by increased receiver selectivity.[1]

[1] See John R. Carson, The Reduction of Atmospheric Disturbances, *Proc. I.R.E.*, vol. 16, p. 966, July, 1928. In this paper Carson shows the fallacy behind a number of proposed methods of balancing out static interference while still preserving the signal.

CHAPTER XVI

RADIO AIDS TO NAVIGATION

138. Fundamental Principles of Radio Direction Finding.[1]—The fact that radio waves propagate away from the transmitter along a great circle route can be utilized in direction-finding work. Thus a ship or airplane can obtain its location by determining the direction of the radio waves sent out by transmitters at known locations. Similarly it is possible to determine the location of a radio transmitter by taking bearings on the radio waves at two receiving locations.

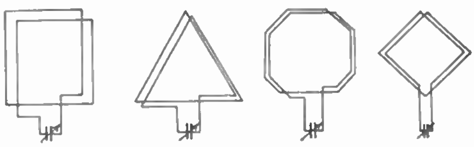


FIG. 334.—Examples of typical loop antennas.

Analysis of Loop Characteristics.—All practical direction-finding systems make use of a loop antenna, which is essentially a large coil of any conveniently-shaped section (see Fig. 334, for examples). Such an antenna has the directional characteristic shown in Fig. 335 and abstracts energy from passing waves as a result of phase differences between the voltages induced in the opposite legs. Thus consider the case of a rectangular loop in the path of a vertically polarized radio wave. When the plane of the loop is perpendicular to the direction of travel of the waves the voltages induced in the two legs are of equal magnitude and the same phase, and being directed around the loop in opposite sense, cancel each other and result in zero response. As the plane of the loop is brought nearer to parallel with the direction of wave travel the wave front reaches the two legs at slightly different times, causing a phase difference between the voltages induced in the two legs and giving rise to a resultant voltage that acts around the loop and is maximum when the plane of the loop is parallel to the direction along which the waves travel. Vector diagrams

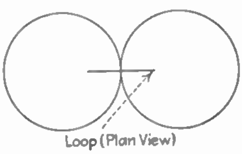


FIG. 335.—Directional characteristic of loop antenna. This applies to loops of all shapes.

[1] For further information on direction finding and for references to the extensive literature on the subject one should read R. L. Smith-Rose, Radio Direction Finding by Transmission and Reception, *Proc. I.R.E.*, vol. 17, p. 425, March, 1929.

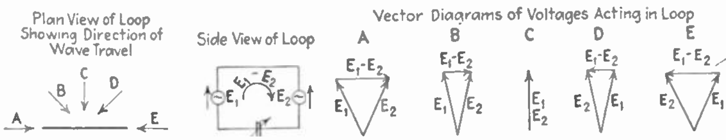
illustrating the situation for several loop orientations are shown in Fig. 336.

The resultant voltage acting around a rectangular loop is given by the following equation:[1]

$$\text{Resultant voltage acting around loop} = 2\epsilon lN \sin\left(\frac{\pi s}{\lambda}\cos\theta\right) \qquad (174)$$

where

ϵ = strength of radio wave in volts per meter
l = height of loop in meters
s = width of loop in meters
N = number of turns in loop
λ = wave length of radio wave in meters
θ = direction of travel of wave with respect to plane of loop.



reverse this arrow

FIG. 336.—Vector diagrams showing how the voltages induced in the two sides of a loop by a passing radio wave combine to give a resultant voltage acting around the loop.

In practical loops the size is small compared with a wave length so that $\sin\left(\frac{\pi s}{\lambda}\cos\theta\right)$ may be written as $\frac{\pi s}{\lambda}\cos\theta$ without appreciable error, giving

$$\text{Resultant voltage acting around loop} = 2\pi\epsilon N\frac{ls}{\lambda}\cos\theta \qquad (174a)$$

$$\text{Resultant voltage acting around loop} = 2\pi\epsilon N\frac{(\text{loop area})}{\lambda}\cos\theta \qquad (174b)$$

Equation (174*b*) applies to loops of all shapes provided only that the loop is small compared with a wave length.

The effectiveness of the loop antenna as a means of abstracting energy from radio waves is measured in terms of the effective height, which is the resultant loop voltage as given by Eq. (174*b*) divided by the field strength ϵ. Hence

$$\text{Effective height of loop} = 2\pi N\frac{(\text{loop area})}{\lambda}\cos\theta \qquad (175)$$

The effective height of a loop of reasonable size is low because of the tendency for the two vertical legs to cancel each other's effects. The

[1] This formula can be readily derived as follows: The voltage induced in each vertical leg is ϵNl, while the phase difference between the voltages is $\frac{2\pi s}{\lambda}\cos\theta$ radians, since the wave front must travel a distance $s\cos\theta$ to pass from one leg to the other. Subtracting the voltages in the two legs while taking into account this phase difference gives Eq. (174).

value of the loop antenna arises from its convenient form and its directional characteristic.

When a loop antenna is used in transmission the strength of the vertical component of the radiated field in a horizontal plane can be found by substituting the effective height as given in Eq. (176) for the length δl of the elementary antenna, giving:

$$\left.\begin{array}{l}\text{Vertically polarized component of}\\ \text{radiated field in volts per meter}\end{array}\right\} = 1183\, I \frac{N(\text{loop area})}{d\lambda^2} \cos\theta \qquad (176)$$

where d is the distance in meters, I is the loop current in amperes, and the remainder of the notation is the same as in Eq. (174).

Direction Finding with Loop.—The direction of travel of a radio wave is determined by rotating the loop until zero signal indicates that the plane of the loop is perpendicular to the direction in which the waves are traveling. The position of minimum rather than maximum response is used because this permits much more accurate settings. In order to determine from which side of the loop the waves are arriving a vertical antenna is used in conjunction with the loop. The vector diagrams of Fig. 336 show that when the loop is adjusted for maximum response the resultant voltage acting around the loop is very nearly 90° out of phase with the wave at the center of the loop and may either lead or lag according to the direction from which the waves arrive. A vertical antenna located at the loop center and coupled inductively to the loop circuit will induce a voltage in the loop that is 90° out of phase with the current in the vertical antenna and hence with the wave at the center of the loop and so will add to, or subtract from, the resultant voltage acting around the loop depending on the direction of arrival of the waves. The procedure for obtaining the true bearing is hence first to adjust the loop for zero response with the vertical antenna disconnected. The loop is then rotated 90° in a specified direction in order to give maximum response, after which the vertical antenna is tuned to resonance and coupled to the loop circuit. The 180° uncertainty in the loop bearing is then removed by noting whether the vertical antenna causes the loop response to increase or decrease.

Errors from Down-coming Waves and Their Elimination.—The bearings obtained by the use of a loop antenna are accurate only when down-coming horizontally polarized sky waves are absent. Horizontally polarized down-coming waves induce voltages in the top and bottom parts of the loop which give rise to a resultant voltage acting around the loop circuit even when the loop is set for minimum response to vertically polarized waves traveling parallel with the surface of the earth. The result is either that there is no loop position giving zero response, or that the position of zero response represents a false bearing. As a consequence the usefulness of the loop as a means of direction finding is limited to the lower radio frequencies, where the waves when observed in the vicinity of the earth are vertically polarized, or when the sky wave is of negligible strength in comparison with the ground wave. Since the sky wave is always strongest at night the errors that result from down-coming horizontally polarized waves are frequently referred to as "night effects" although they are always present to some extent in daytime.

The errors in bearing caused by down-coming horizontally polarized sky waves can be eliminated by replacing the loop antenna with the Adcock aerial system, which in its simplest form consists of two spaced vertical antennas connected as shown in Fig. 337. The action of such an antenna as far as vertically polarized waves are concerned is identical with the loop since the resultant current in the output coil of the Adcock antenna is proportional to the vector difference of the voltages induced in the two vertical members, exactly as is the case with the loop. Horizontally polarized down-coming waves do not affect the Adcock antenna, however, since the connections are arranged so that all voltages induced in the horizontal members balance out each other's effects. The usefulness of the Adcock aerial system is greatly limited by the fact that it is equivalent to a single-turn loop and hence has an extremely low effective height in proportion to size.

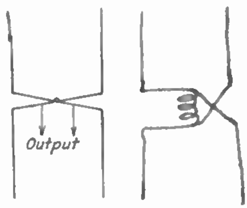


FIG. 337.—Simple form of Adcock aerial system.

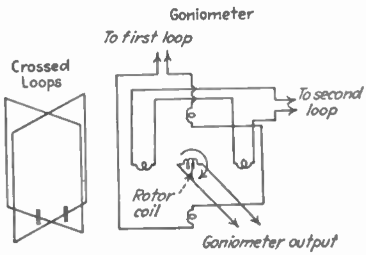


FIG. 338.—Goniometer arrangement for shifting directional characteristic of loop antenna without moving the loop.

Goniometer Arrangements.—When it is desirable to avoid rotating a loop antenna, as for example when very large loops are employed, it is possible to obtain the effect of rotation by using two loop antennas at right angles to each other and combining the outputs in a goniometer. The goniometer consists of two pairs of primary coils (one pair for each loop) arranged at right angles to each other and coupled to a secondary coil as shown in Fig. 338. If the goniometer is built so that the mutual inductance between each pair of primary coils and the secondary is proportional to the cosine of the angle which the axis of the secondary coil makes with the axis of the primary coils, then for any position of the secondary the directional characteristic of the antenna system will be identical in form with the directional characteristic of a single loop antenna. The orientation of this characteristic depends upon the position of the rotating coil, however, and can be changed at will by rotating the goniometer secondary just as the orientation of the directional characteristic of a loop is changed by rotating the loop.

Effect of Capacity Unbalances.—In order to obtain the theoretical directional characteristic of the loop antenna as given in Eq. (174) it is necessary to balance the loop carefully with respect to ground. If this is not done currents flowing through the stray capacities between loop and ground will either obscure the point of minimum response or will cause a false minimum. Thus if one side of the loop tuning condenser is grounded as shown in Fig. 339, the current through the ground capacity C_g' will produce a voltage across the tuning condenser and hence cause a current to circulate around the loop even when the loop is orientated in the position that should give zero response. Errors of this type can be eliminated by using balanced circuits, such as those shown in Fig. 339.

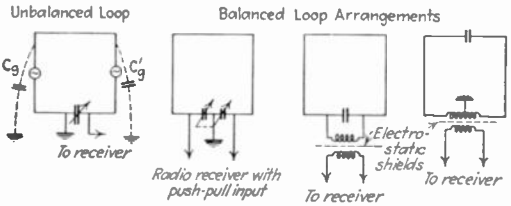


FIG. 339.—Unbalanced and balanced loop arrangements.

139. Practical Systems of Direction Finding.—The principal use of direction-finding systems is to enable ships and airplanes to obtain their position. The methods commonly used to accomplish this are:

1. Loop antenna and associated receiver on ship for the purpose of obtaining bearings on radio waves from any radio transmitter of known location that may be within range.
2. Loop antennas and associated receivers located at strategic positions and used to obtain bearings on the radio waves sent out by the ship (or plane) that desires to know its position.
3. Special beacon transmitters located at important points.

The first two of these systems ordinarily employ a balanced loop that can be orientated to any desired position. A sensitive radio receiver is used, and provision is made to obtain the sense of the bearing by means of a vertical antenna that can be coupled into the loop output. The special-beacon method consists of a radio transmitter which employs as its radiating element a loop antenna that is slowly rotated. When the loop orientation is such that the minimum signal is in a stated direction (ordinarily north-south) a distinctive or marking signal is radiated on a non-directional antenna. Any radio receiver within range of the beacon transmitter can then obtain its bearing with respect to the beacon by using a stop watch to measure the time that elapses between the sending of this distinctive signal and the instant when the received beacon signal goes through zero.

The choice between these direction-finding systems is a matter of convenience, flexibility, cost, etc., since the accuracy is the same in all cases. In particular the Rayleigh-Carson reciprocal theorem (see Sec. 119) shows that under the conditions where accurate bearings are possible (no effect from down-coming sky waves) the observed bearings are unchanged by interchanging the transmitter and the direction-finding receiver.

It is always necessary to calibrate a loop after installation in order that the distorting effect which metal objects, buried pipes, local topography, etc., have on the bearings may be corrected for. It is also necessary of course that the loop be installed where these disturbing factors are of a fixed character and do not change from time to time.

A properly compensated and calibrated loop antenna will give bearings on near-by radio stations that are accurate to within 1° or 2°. The accuracy diminishes, however, as the distance to the transmitter increases because of the errors introduced by down-coming sky waves. The range over which satisfactory bearings can be obtained is greater during the day than at night, is reduced as the frequency is raised, and is somewhat more when the transmission path is over water than when over land. This is because all of these factors cause the ground wave to be stronger in proportion to the sky wave. The maximum distance is in the order of 50 to 200 miles at frequencies in the order of 500 kc and may be as great as several thousand miles at frequencies in the order of 20 kc. When down-coming sky waves are not present, or when their effects are eliminated by means of an Adcock aerial system, it is possible to obtain relatively accurate bearings on distant transmitting stations since radio waves seldom deviate laterally by more than a few degrees from the great circle route even in transmission over great distances. The Adcock aerial system, while superior to the loop in not having a night-error effect, is seldom used in commercial direction-finding equipment because of its very low sensitivity in comparison with a multiturn loop of similar dimensions.

Direction Finding at High Frequencies.—At frequencies in the order of 3000 kc and higher loop antennas are utterly useless in direction-finding work because of the strong down-coming sky waves that are always present. An Adcock aerial system must be employed at such frequencies and, in addition to being balanced to ground, must also be thoroughly shielded electrostatically to prevent horizontally polarized waves traveling parallel with the earth's surface from producing stray currents which affect the indicated bearings. With a carefully balanced and electrostatically shielded Adcock system it has been found possible to obtain reasonably accurate bearings on short waves up to distances in the order of several hundred kilometers.[1] It is found, however, that while these bearings give the direction in which the waves are traveling they do not necessarily represent the bearing of the transmitter. Thus when an

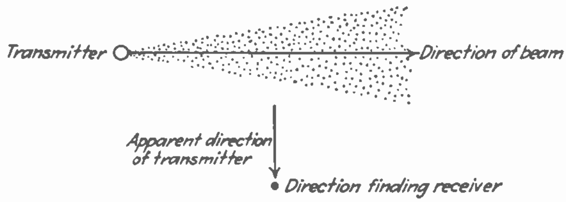


FIG. 340.—Schematic diagram showing how short-wave signals radiated from a highly directional antenna may have their direction changed as a result of scattering by the Kennelly-Heaviside layer.

attempt is made to determine the bearing of signals radiated from a highly directional antenna system using a receiver somewhat to the side of the main beam, as shown in Fig. 340, the indicated bearing will ordinarily be toward the nearest point of the main beam. This is apparently because of scattering in the Kennelly-Heaviside layer, which causes most of the energy arriving at such a direction-finding receiver to be scattered radiation coming from the nearest point on the main beam rather than directly from the transmitter. Effects such as this appear to place an absolute limit on the ultimate possibilities of short-wave direction finding.

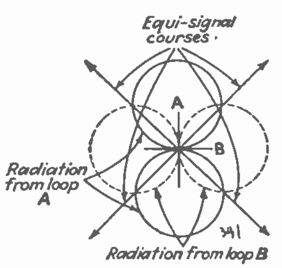


FIG. 341.—Directional characteristics of signals radiated from an aural-type radio range showing equisignal courses that are produced.

140. The Radio Range.[2]—The radio range is a type of radio beacon which lays down a course in a predetermined direction. The radio range may be of either the aural or visual type. The aural type employs two crossed loops which are alternately excited from a common source of radio-frequency power. The directional characteristics of the crossed loops are shown in Fig. 341, where it is seen that in certain directions radiating away from the transmitter the signals from the two loops are of equal strength. By interlocking the two loops so that one of them is always radiating energy, and then sending out complementary signals

[1] For a discussion of short-wave direction finding with the Adcock aerial see R. H. Barfield, Recent Developments in Direction Finding Apparatus, *Exp. Wireless and Wireless Eng.*, vol. 7, p. 262, May, 1930.

[2] A summary (with bibliography) of the work that has been done on the radio range is given by J. H. Dellinger, H. Diamond, and F. W. Dunmore, Development of the Visual-type Airway Radio Beacon, *Proc. I.R.E.*, vol. 18, p. 796, May, 1930.

such as N(—— -) and A(- ——), the signal heard along the equisignal line is a continuous dash, while at points to the side of this equisignal course either one or the other of the code characters dominates, depending on which side of the course one is located.

The visual type of radio range is made in a number of forms. The simplest arrangement employs the same antenna system as the aural type, but instead of alternately exciting the separate antennas, it radiates a carrier wave modulated at a convenient low frequency (usually 65 cycles) from one loop and the same carrier frequency modulated at a different frequency (usually 86.7 cycles) from the second loop. The result of these two radiations is to produce equisignal zones in which the 65- and 86.7-cycle side bands are of the same amplitude. An airplane making use of signals from the visual-type radio range is supplied with a radio receiver which delivers its output to two reeds tuned to 65 and 86.7 cycles, respectively, which vibrate with an amplitude corresponding to the strength of the corresponding side-band components. These reeds are mounted on the pilot's instrument board with the tips visible. The tips are painted white and when in vibration give white lines proportional to the side bands involved. When on the course laid out by the visual radio range the two reeds vibrate with equal amplitude while when off the course one or the other of the reeds will vibrate with greater amplitude than the other and will indicate in which direction the true course lies.

The Courses and Their Alignment.—The courses produced by the visual and aural types of radio range are not necessarily the same, because in the visual type the carrier is radiated continuously from both loops. The situation that exists when the carrier currents in the two antennas are in the same time phase is illustrated at Fig. 342*a*. The side-band energy radiated by each loop possesses the usual figure-of-eight characteristic just as though the other loop were not present, but the carrier energy radiated by the separate loops combines to form a carrier having the directional characteristics shown in the figure. The result is a beacon providing only two instead of four courses. In order to obtain four courses from the visual type of range it is necessary that the carrier currents in the two antennas be of different time phase. The situation that exists when the two carriers are 90° out of phase is illustrated at Fig. 342*b*.

In the practical application of the radio range it is necessary to align the beacon course to actual routes followed by air traffic. This can be done in a number of ways. One method consists in supplying the two antennas with equal carrier waves modulated to different degrees, which produces the effect shown at Fig. 342*c*. Another arrangement consists in using a vertical antenna coaxial with the beacon tower and excited by coupling to the output of either one or both of the amplifiers supplying energy to the crossed loops. This arrangement has effects such as illustrated at Fig. 342*d*. By using combinations of these methods it is possible to align the four courses of a visual beacon in four arbitrary directions. It is also possible for a pilot to select a special course to either side of the normal beacon course by simply shunting the appropriate reed in the visual indicator in order to reduce its sensitivity.

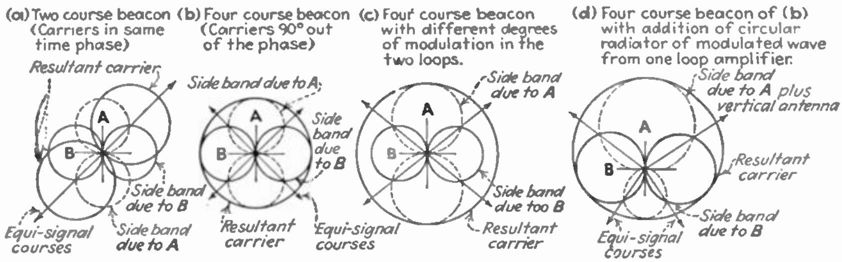


FIG. 342.—Directional characteristics of visual-type radio range, showing how the orientation of the equisingal courses can be changed.

Miscellaneous Features of the Radio Range.—The above discussion has been confined to the two- and four-course types of visual radio range. It is also possible to obtain a single-course beacon by combining the loop radiation with circular radiation of suitable intensity from a vertical antenna. More than four courses can be obtained by the addition of a third modulation frequency, such as 108.3 cycles at the transmitter. Such an arrangement leads to a beacon similar in its properties to the two- and four-course beacons but possessing 12 courses which can be arbitrarily aligned along 12 air lanes.

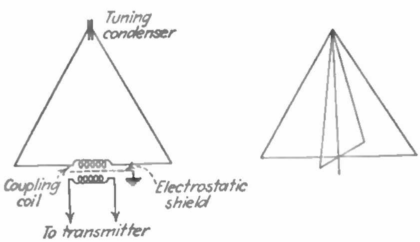


FIG. 343.—Details of loop antennas used in radio-range transmitter.

The radio-range transmitters operate in the frequency range 285 to 315 kc, and make use of two large triangular single-turn crossed-loop antennas mounted on a common pole, as shown in Fig. 343. A tuning condenser is placed at the apex of the loop while the center point of the coupling inductance is grounded. This arrangement balances the loop and reduces the potential difference between the loop and ground. The large size of the loops makes it inconvenient to orient the directional characteristics by rotation, and goniometer arrangement is used instead.

The loops are ordinarily fed from separate power amplifiers which are excited from a common oscillator, and the phase relations are controlled by slightly detuning the circuits.

In receiving the radio-range signals it is necessary that the receiving antenna respond only to vertically polarized radio waves. If down-coming horizontally polarized sky waves affect the receiving antenna there will be errors similar to those present in radio direction finding. These errors are greatest at night but can be appreciable even in the daytime. In receiving the radio-range signals it is therefore necessary to use either a simple vertical antenna or a vertical antenna with symmetrically arranged flat top. Trailing wire antennas are worse than useless in radio-range work.

Radio-range transmitters of the aural type have been in operation at some of the important airports in the United States for a number of years. As time goes on it is planned to replace these with the visual-type range and to make additional installations at strategic points along all important air routes. The usual radio transmitter supplies approximately 1 kw of radio power to each of the crossed antennas and produces signals that have a useful daylight range of approximately 100 miles. The width of the equisignal zone produced by a transmitter depends upon a number of factors but is normally approximately 10 miles at a distance of 100 miles and narrows down as the transmitter is approached.

The radio range and the radio compass represent the only widely used radio aids to airplane navigation. Many other radio devices for aiding air navigation have been proposed, however, and it is probable that with time some of these will be generally adopted. Among these are radio methods of determining the altitude of a plane above ground, means of making blind landings, homing devices by which a plane may follow a course directly to an airport provided with a radio transmitter, etc.

CHAPTER XVII

RADIO MEASUREMENTS[1]

141. Resistance.—The effective resistance offered by a circuit to alternating current of power and audio frequencies is most conveniently obtained with the aid of an alternating-current bridge. Many forms of such bridges have been devised, and the more standard arrangements are too well known to need discussion here.[2]

When the resistance being measured is high, such as 50,000 ohms or more, trouble is usually experienced with stray capacity currents flowing from bridge to ground and between different parts of the bridge. Errors arising from this source can be avoided by the use of a Wagner earth connection, as illustrated at Fig. 344*b* and *c*. A moderate- or

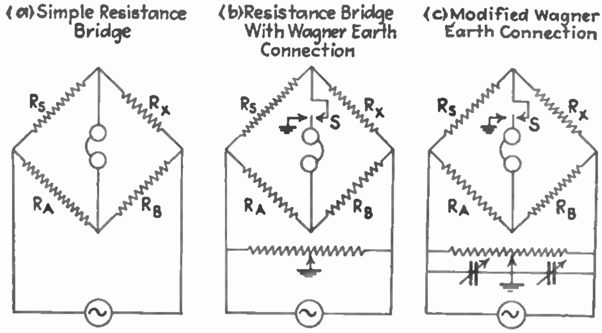


FIG. 344.—Simple alternating-current bridges with and without Wagner earth connection.

low-resistance potentiometer is placed across the power supply and the adjustable contact is grounded, while the switch S is arranged to connect the receivers either across the bridge diagonal, or from one corner of the bridge to ground. The procedure for adjusting the bridge is as follows: The bridge is first balanced as well as possible in the usual way, after which the switch S is thrown to the left, and the earth connection on the potentiometer is adjusted until no sound is heard in the phones.

[1] This chapter gives a brief summary of the measurements and measuring methods most widely used in radio work. More complete treatments are to be found in books devoted exclusively to this one subject, such as: H. A. Brown, "Radio Frequency Measurements," McGraw-Hill Book Company, Inc., New York; Bur. Standards *Circ.* 74, Radio Instruments and Measurements; E. B. Moulin, "Radio Frequency Measurements," Charles Griffin & Co., Ltd., London.

[2] For an extensive discussion of impedance bridges see Frank A. Laws, "Electrical Measurements," McGraw-Hill Book Company, Inc., New York, 1917.

The switch S is thrown to the right, and the balance of the bridge is completed.

In some cases, particularly at the higher frequencies, the effect of inductances and stray capacities in the bridge requires provision for adjusting the phase angle of the impedances in the Wagner earth connection between ground and each side of the power source. This is usually accomplished by the use of adjustable shunt condensers, as indicated at Fig. 344c, but with some bridge arrangements it requires the use of a variable inductance in series with one side of the potentiometer.

Radio-frequency Resistance—Neutralization Method.—Radio-frequency resistance is most conveniently measured by making the unknown resistance a part of a resonant circuit tuned to the frequency for which the results are desired, and then measuring the total circuit resistance by the resistance-neutralization, the resistance-variation, or the reactance-variation method.

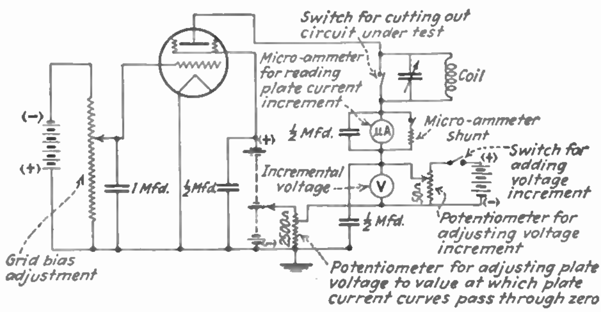


FIG. 345.—Circuit connections for measuring the parallel resonant impedance of a tuned circuit by the resistance-neutralization method.

In the resistance-neutralization method, the tuned circuit is shunted by an adjustable negative resistance produced by a dynatron, as shown in Fig. 345 (also see Sec. 79).[1] This negative resistance is varied by changing the control grid bias until oscillations just barely fail to be produced, in which condition the resistance developed across the tuned circuit by parallel resonance is exactly equal to the negative plate-cathode resistance of the dynatron. The series resistance of the tuned circuit is then given by the formula

$$\text{Series resistance of circuit} = \frac{(\omega L)^2}{r} \tag{177}$$

where

ωL = the inductive reactance across which the dynatron is connected

r = negative plate-cathode resistance of dynatron at the adjustment at which oscillations just cease.

[1] See Hajime Iinuma, A Method of Measuring the Radio-frequency Resistance of an Oscillatory Circuit, *Proc. I.R.E.*, vol. 18, p. 537, March, 1930.

The resonant frequency of the tuned circuit can be readily determined by adjusting the dynatron so that oscillations are produced and then measuring the resulting frequency by heterodyning with a calibrated oscillator. The inductance L that appears in Eq. (177) is the true inductance of the coil, *i.e.*, the inductance disregarding the effects of distributed capacity, and can be obtained with sufficient accuracy by measurements at audio frequencies. The negative plate-cathode dynatron resistance r can be measured either by an alternating-current bridge, or by adding a small increment, such as 1 to 5 volts, to the plate potential and measuring the resulting current increment with the aid of a microammeter. The ratio

$$\frac{\text{Added plate voltage}}{\text{Resulting current increment}}$$

represents the negative plate resistance of the tube. In order to avoid errors the voltage increment must be so small that the range of tube characteristic over which it extends can be considered as a straight line. The circuit of Fig. 345 includes provision for measuring the negative plate resistance in this way. When the negative plate-cathode resistance of the dynatron is obtained by the incremental method it is desirable to operate the tube at the plate voltage for which the plate-current curves all pass through zero (see Fig. 159) in order that the initial plate current will be zero before the addition of the voltage increment. The resistance-neutralization method of measuring radio-frequency resistance does not require correction for distributed coil capacity, involves no special precautions or delicate manipulation, and with reasonable care gives results with an accuracy satisfactory for all practical purposes.

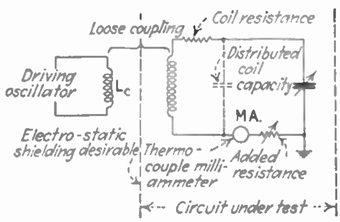


FIG. 346.—Circuit arrangement for measuring radio-frequency resistance by the resistance-variation method.

Radio-frequency Resistance—Resistance-variation Method.—The resistance-variation method of determining radio-frequency resistance takes advantage of the fact that when a circuit is tuned to resonance the current that flows is the applied voltage divided by the circuit resistance. If the applied voltage is kept constant and the circuit resistance is varied by known increments it is possible to deduce the actual circuit resistance from the resulting current changes. The circuit arrangements for carrying out the necessary measuring operations are illustrated in Fig. 346, which shows the circuit under test loosely coupled to a driving oscillator and in series with a thermo-milliammeter and an adjustable resistance. The circuit is first tuned to resonance with the driver, and the circuit current observed when the added resistance is zero. Some resistance is then added, the circuit is retuned to resonance without changing the coupling to the driver, and the resulting current noted. The apparent resistance is then given by the formula:[1]

$$\text{Apparent series resistance of tuned circuit} = R_1\left(\frac{I_1}{I_0 - I_1}\right) \quad (178)$$

where

R_1 = the added resistance
I_0 = the current with no added resistance
I_1 = the current with added resistance R_1.

In order to obtain accurate results with the resistance-variation method it is necessary that the current through the coupling coil L_c be kept constant, and that the only coupling between the oscillator and the circuit under test be inductive. These requirements can be most satisfactorily met by loose coupling between the two circuits and in some cases by providing an electrostatic shield. As a check it is always desirable to repeat the measurements with several values of added resistance. In order to avoid errors from stray ground capacities it is desirable to ground one side of the condenser and to place the milliammeter and added resistance on the grounded side of the circuit, as shown in Fig. 346. The added resistance should of course be constructed with negligible skin effect, so that its value can be measured on direct currents.

The chief drawback of the resistance-variation method is that the results must be corrected for the effects of distributed coil capacity. This is made apparent by reference to Fig. 346 where it is seen that the added resistance is not in series with the complete circuit but only with the portion of the tuning capacity included in the tuning condenser. The circuit resistance as obtained from Eq. (178) is therefore the apparent coil resistance as defined in Sec. 14, plus the resistance of the tuning condenser, and is too high. If it can be assumed that the losses in the coil distributed capacity are a negligible part of the entire circuit losses, which is true if the coil is properly constructed, it is possible to use Eq. (35) to correct the results given by Eq. (178). The correction that must be applied to the observed results is in the order of 12 per cent at a frequency that is one-fourth of the frequency at which the distributed

[1] This is derived as follows: With no added resistance the induced voltage E from the driver operates against the circuit resistance R_0 to give

$$I_0 = \frac{E}{R_0}$$

The addition of a resistance R_1 to the circuit does not change E, so the current now becomes

$$I_1 = \frac{E}{R_0 + R_1}$$

Simultaneous solution of these two equations to eliminate E gives Eq. (178).

capacity is in resonance with the coil inductance, and increases rapidly as the natural frequency of the coil is approached more closely.

Radio-frequency Resistance—Reactance-variation Method.—In the reactance-variation method of measuring radio-frequency resistance the circuit is tuned to resonance with a driving oscillator of the desired frequency, and the current I_0 is observed. The tuning capacity is then increased in known amounts to some value C_2 at which the current has dropped to some convenient value I', after which the capacity is reduced to a value C_1 such that the current is again reduced to the value I'. Then

$$\text{Apparent series resistance of circuit} = \frac{C_2 - C_1}{2\omega C_1 C_2}\sqrt{\frac{I_1^2}{I_0^2 - I_1^2}} \qquad (179)$$

The circuit for the reactance-variation method is similar in all essential respects to the circuit for resistance variation, the only difference being that instead of an added resistance it is necessary to provide a calibrated tuning condenser. An accurate evaluation of $C_2 - C_1$ can be obtained by using a small calibrated vernier condenser in shunt with the main tuning condenser. The results given by Eq. (179) must be corrected for distributed coil capacity by the use of Eq. (35), exactly as in the resistance-variation method.

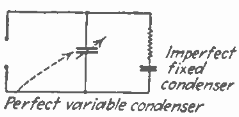


FIG. 347.—Equivalent electrical circuit of a variable condenser constructed so that the energy losses do not vary with the capacity setting.

Radio-frequency Resistance of Condensers.—After the total series resistance of the circuit has been obtained by one of the methods outlined above, it still remains to determine how this is divided between coil and condenser. The fraction of the total circuit resistance that is contributed by the condenser is usually relatively small, particularly if the condenser is well constructed, so that for approximate work it is permissible to assume that all of the measured resistance is coil resistance. The condenser series resistance can be ascertained with the aid of an auxiliary high-quality variable condenser constructed so that its losses are independent of the capacity setting. Such a condenser is equivalent to a fixed condenser having a loss, shunted by a perfect variable condenser, as shown in Fig. 347. The series resistance of the condenser under test is determined by first measuring the resistance of a circuit which is tuned to resonance with the auxiliary condenser. The condenser under test is then connected in shunt with the auxiliary condenser, the capacity of which is reduced until the circuit is again resonant at the original frequency. The difference between the circuit resistance in the two cases then gives the desired series resistance by means of the formula:[1]

$$\text{Equivalent series resistance of condenser} = R\frac{C_o^2}{C_x^2} \qquad (180)$$

[1] C. T. Burke, Substitution Method for the Determination of Resistance of Inductors and Capacitors at Radio Frequencies, *Trans. A.I.E.E.*, vol. 46, p. 482, 1927.

where R is the increase in circuit resistance caused by adding the condenser being tested, while C_x is the capacity of the condenser under test, and C_o is the capacity of the auxiliary condenser before adding the capacity C_x. The capacity C_x is conveniently obtained as the difference between the capacities corresponding to the two settings of the auxiliary condenser.

142. Measurement of Capacity.—Condensers having capacities of 0.01 μf or more can be measured on any one of a number of types of alternating-current impedance bridges without any special precautions. When the capacity is less than about 0.01μf, however, the reactance is so great at the usual 1000-cycle bridge frequency that a Wagner earth connection must be employed if satisfactory results are to be obtained, while if the capacity is much less than 0.001μf specially shielded bridges must be employed if accuracy is important.

An example of a shielded bridge intended for the measurement of capacities of 1$\mu\mu$f upward is shown in Fig. 348. This consists of an ordinary unity-ratio impedance bridge in which the unknown capacity is balanced by an adjustable standard condenser, with the power factors of the unknown and standard equalized by a variable resistance R. The entire bridge is shielded electrostatically from external influences by means of a grounded copper case. Transformers with grounded electrostatic shields between primaries and secondaries are provided for supplying the bridge with audio power and for detecting balance. Errors arising from stray capacities are avoided by using identical resistances symmetrically arranged for the ratio arms and by enclosing the variable resistance R in a separate shield.[1]

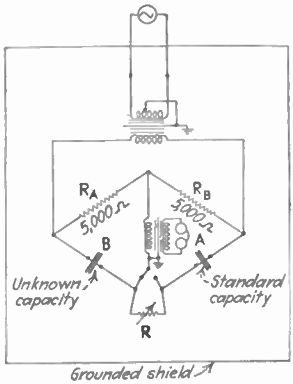


Fig. 348.—Shielded capacity bridge suitable for accurately measuring small capacities.

In addition to these bridge methods it is also possible to measure capacity at radio frequencies by substitution. In this method a coil is first tuned to resonance with a driving oscillator by means of an auxiliary calibrated condenser. The unknown condenser is then placed in shunt with the calibrated condenser, and the circuit is brought back to resonance by readjusting the calibrated condenser. The unknown capacity is then the change in capacity of the calibrated auxiliary condenser.

[1] Excellent discussions of the problems involved in shielding bridges for both audio- and radio-frequency measurements are given by: W. J. Shackelton, A Shielded Bridge for Inductive Impedance Measurements, *Trans. A.I.E.E.*, vol. 45, p. 1266, 1926; J. G. Ferguson, Shielding in High-frequency Measurements, *Trans. A.I.E.E.*, vol. 48, p. 1286, October, 1929.

The accuracy is limited by the accuracy with which the condensers can be set to give resonance and is not as great as might be desired because of the flap top of the resonance curve.

Distributed Capacity of Coils.—The distributed capacity associated with an inductance coil can be measured by determining the capacities that must be shunted across the coil terminals to give resonance at several different frequencies. A curve, such as Fig. 349, is then plotted between capacity shunted across the coil and the reciprocal of the resonant frequency squared. This curve will be a straight line because the tuning capacity is inversely proportional to the square of the resonant frequency, and will intercept the capacity axis at a negative capacity which is equal to the distributed capacity of the coil. The slope of the line is a measure of the radio-frequency inductance, which is given by the equation

$$\text{Coil inductance in henrys} = 0.0253m \tag{181}$$

where m is the slope of curve $1/f^2$ plotted against capacity when the frequency is in megacycles, and the capacity is in micromicrofarads.

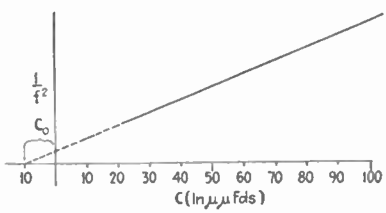


FIG. 349.—Capacity C shunted across coil plotted as a function of $\frac{1}{(\text{resonant frequency})^2}$. The value of negative capacity C_o at which the extrapolated curve intercepts the capacity axis is the distributed capacity of the coil.

The necessity of plotting a complete curve, such as that of Fig. 349, can be avoided by using a calibrated condenser to tune the coil to resonance with an oscillator and then reducing the condenser capacity until the coil is brought into resonance with the second harmonic of the oscillator. If C_1 is the tuning capacity required for the fundamental frequency, and C_2 the capacity at the second harmonic, the distributed capacity is then[1]

$$\text{Distributed coil capacity} = \frac{C_1 - 4C_2}{3} \tag{182}$$

The distributed capacity of a coil is sometimes determined by measuring the resonant frequency of the coil when tuned by its own capacity. This

[1] See Ralph R. Batcher, Rapid Determination of Distributed Capacity of Coils *Proc. I.R.E.*, vol. 9, p. 300, August, 1921.

Equation (182) results from the fact that the total capacity (*i.e.*, added capacity plus distributed capacity) required to tune the coil to the fundamental frequency is exactly four times the total capacity required to tune to the second harmonic.

can be accomplished by loosely coupling the coil to an oscillator and observing the reaction of the coil on the oscillator as the frequency is varied. A knowledge of this resonant frequency and the coil inductance will permit a determination of the apparent distributed capacity. The results obtained in this way do not agree with the previous two methods because the voltage and current distribution in the coil with no external capacity present is quite different from the distribution in the presence of a normal tuning capacity. For this reason the determination of distributed capacity by means of the natural resonant frequency of the coil is not recommended.

Direct-capacity Measurements—Electrode Capacity of Tubes.—Circumstances sometimes arise where it is desired to determine the capacity that exists between two terminals which also possess capacities to other terminals. An excellent illustration of this is supplied by an ordinary three-electrode vacuum tube, where one may desire to determine the grid-plate capacity but finds that the measurement is complicated by the presence of the grid-filament and plate-filament capacities. Such a direct capacity between two electrodes can be measured by means of the Campbell-Colpitts bridge arrangement[1] which is illustrated in Fig. 350. The terminals between which the unknown capacity exists are connected in one arm of the bridge, while all other terminals are brought together and connected to the grounded corner of the bridge. In this way some of the auxiliary capacities are shunted across the output diagonal of the bridge and so have no effect, while the remainder are shunted across the arm AB and are balanced by the condenser C_2. The unknown capacity is measured directly in terms of the setting of condenser C_s required to give balance.

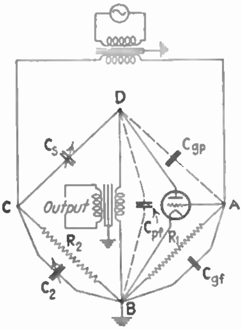


FIG. 350.—Campbell-Colpitts capacity bridge for measuring direct capacity between two electrodes, which in the case shown are the grid and plate of a vacuum tube.

Another method of measuring interelectrode tube capacities consists in applying a known radio-frequency voltage to the desired capacity shunted by an adjustable standard capacity, and measuring the current on a thermo-milliammeter. The unknown capacity is then removed and the standard capacity readjusted until the condenser current is the same as before. The difference between the standard capacities in the two cases represents the tube capacity. This method is particularly well adapted to the measurement of extremely small capacities, such as those

[1] Lincoln Walsh, Direct-capacity Bridge for Vacuum-tube Measurements, *Proc. I.R.E.*, vol. 16, p. 482, April, 1928; E. T. Hock, A Bridge Method for the Measurement of Interelectrode Admittance in Vacuum Tubes, *Proc. I.R.E.*, vol. 16, p. 487, April, 1928.

existing between the control grid and plate electrodes of screen-grid tubes.[1]

143. Inductance and Mutual Inductance.—Inductance measurements are most conveniently carried out by an alternating-current bridge at frequencies in the order of 1000 cycles. A Wagner earth connection is required when the reactance of the inductance under test is high, such as 100,000 ohms, at the frequency supplied to the bridge.

The measurement of inductance at radio frequencies is very difficult because of the effect which distributed capacity has at these frequencies, and it is usually more accurate to assume that the inductance at radio frequencies is the same as at audio frequencies, where the distributed capacity has negligible effect. If the inductance must be evaluated at radio frequencies then Eq. (181) can be used.

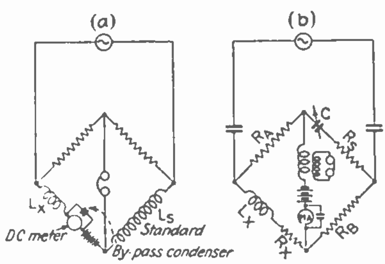


Fig. 351.—Bridge circuits for measuring the inductance which a coil offers to an alternating-current voltage superimposed upon a direct-current magnetization.

In coils which carry a direct current, such as filter inductances and the primaries of audio-frequency transformers, it is often necessary to determine the inductance which is offered to a small alternating-current voltage superimposed on this direct current. There then arises the problem of how to pass the direct current through the inductance without interfering with the bridge measurement. One way of accomplishing this is to use a bridge in which the resistance arms have sufficient current-carrying capacity to permit the introduction of the direct-current voltage directly in the bridge, as shown at Fig. 351*a*. Another method is shown in Fig. 351*b*, and involves balancing the unknown inductance against an adjustable standard condenser. In this way the necessity of a large low loss standard inductance is avoided, and furthermore the direct current is not affected by the balancing operation. When the bridge is balanced $R_x = R_A R_B R_s/(R_s^2 + (1/\omega C)^2)$, and $L_x = R_A R_B (1/\omega^2 C)/(R_s^2 + (1/\omega C)^2)$. Normally $R_s \ll (1/\omega C)$, so that to a good opproximation $L_x = R_A R_B C$, and $R_x = R_A R_B R_s (\omega C)^2$.

In measuring the inductance of filter coils and audio-frequency transformer primaries it is usually necessary to use a very low frequency, such as 60 to 400 cycles, because of the high distributed capacity which such inductances possess. Thus in the case of audio-frequency transformers this capacity will usually cause the coil to have a natural resonant frequency of about 1000 cycles. Measurements at frequencies of this order of magnitude will hence give the parallel-resonant impedance

[1] See A. V. Loughren and H. W. Parker, The Measurement of Direct Interelectrode Capacitance of Vacuum Tubes, *Proc. I.R.E.*, vol. 17, p. 957, June, 1929.

rather than the inductance of the coil. If the departure of the apparent inductance from the true inductance is to be kept within 5 per cent it is necessary that the frequency at which the measurements are made be not more than 22.4 per cent of the resonant frequency.

The mutual inductance between two coils can be most conveniently measured by the method outlined in Sec. 7. The procedure is briefly to connect the two coils in series and measure the inductance of the combination. The connection to the terminals of one of the coils is reversed and the inductance is again measured. The difference between the two measured inductances is then four times the mutual inductance.

144. The Measurement of Current.—The most widely used method of measuring currents at radio and audio frequencies consists in passing the current to be measured through a small high-resistance wire which is associated with a thermocouple so that the heat generated by the passage of the unknown current in the wire causes the thermocouple to generate a small voltage which is then measured on a direct-current millivoltmeter.

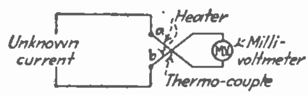


Fig. 352.—Circuit diagram of thermocouple instrument.

The thermocouple ammeters made commercially are of two general types. The most commonly used arrangement consists of a portable d'Arsonval type of galvanometer containing a crossed-wire thermocouple, such as illustrated in Fig. 352. Such thermocouple meters can be made to give full-scale deflections on currents as low as 100 milliamp., and their resistance to the current being measured ranges from about 5 ohms for the most sensitive instruments to a small fraction of 1 ohm where large currents are to be handled. One of the chief drawbacks of these thermocouple meters is that the deflection of the galvanometer is proportional to the square of the current passed through the thermocouple heater because the temperature that the heater reaches is proportional to the energy dissipated in the resistance, and this varies as the square of the current. This disadvantage can be overcome by having galvanometer pole pieces shaped to compensate for the square-law action of the thermocouple.[1] When currents smaller than 100 milliamp. are to be measured it is necessary to place the thermocouple and its heater in a vacuum in order to reduce the rate at which heat is lost and to permit the use of high heater temperatures. Such vacuum thermocouples are used extensively in telephone work and can be made to give a full-scale deflection on a sensitive galvanometer with alternating currents as low as 1 milliamp.

Thermocouple instruments give an indication dependent on the effective value of the current and can be calibrated with direct currents. In making such calibrations with instruments in which the couple is in contact with the heater it is necessary to reverse the direct current for

[1] See U. S. Patent 1,782,588 to F. E. Terman.

each point and average the readings for the two polarities in order to prevent errors from small spurious direct currents in the galvanometer.

Shunts and Current Transformers for Radio-frequency Measurements.—The measurement of extremely large currents at radio frequencies presents a number of difficulties. Ordinary resistance shunts cannot be employed because at radio frequencies the inductance of these shunts, while small, will be sufficient to have considerable influence on the current distribution between the shunt and meter. An ingenious method of overcoming this limitation is by means of the condenser shunt shown in Fig. 353, in which a thermocouple instrument is placed in series with a small condenser, and the whole combination is shunted by a much larger condenser.[1] The sizes of these condensers are such that their reactance is large compared with the meter resistance and the inductive reactance of the connections, so that the current distribution is practically in proportion to the condenser capacities. At low radio frequencies shunts consisting of a number of wires arranged on the circumference of a cylinder with one of these wires used as the thermocouple heater have been successful. Uniform division of current is insured by the fact that the shunts and the heater are symmetrically arranged.

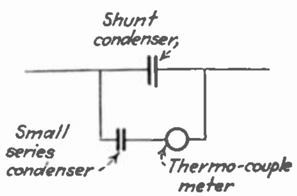


FIG. 353.—Circuit diagram of thermocouple meter provided with condenser multiplier for the measurement of large currents.

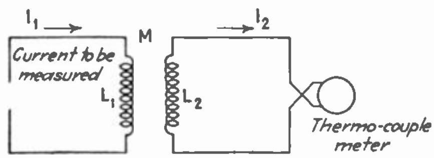


FIG. 354.—Circuit diagram of radio-frequency current transformer used for measuring large currents.

Large radio-frequency currents are sometimes measured by current transformers. These consist of a small primary inductance coupled to a secondary inductance, the terminals of which go directly to a thermocouple meter as shown in Fig. 354. If the ratio of inductive reactance to resistance for the secondary is not less than about 3, and if the capacity between the primary and secondary of the transformer has a negligible effect, then the ratio of primary to secondary current is given by the equation[2]

$$\frac{\text{Secondary current}}{\text{Primary current}} = \frac{M}{L_2} \tag{183}$$

[1] Alexander Nyman, Condenser Shunt for Measurement of High-frequency Currents of Large Magnitude, *Proc. I.R.E.*, vol. 16, p. 208, February, 1928.

[2] This is derived as follows: The primary current I_1 induces a voltage $\omega M I_1$ in the secondary which has an impedance very nearly ωL_2. Hence the secondary current I_2 is

$$I_2 = \frac{\omega M I_1}{\omega L_2} = \frac{I_1 M}{L_2}$$

where

M = mutual inductance.
L_2 = secondary inductance

By suitably proportioning the inductances and coupling of the primary and secondary coils any desired current ratio can be obtained. Electrostatic coupling between the primary and secondary windings of the transformer has the effect of altering the ratio of secondary to primary current. Exact analysis shows that the variation of the ratio with frequency is given by an equation of the form

$$\frac{\text{Secondary current}}{\text{Primary current}} = \frac{M}{L_2}(1 + \omega^2 K) \tag{183a}$$

where K is a constant that depends on the construction. The correction for frequency as given by Eq. (183a) can be made small by placing an electrostatic shield between the two windings.

Radio-frequency currents are sometimes measured by passing them through a known inductance or capacity and measuring the resulting voltage drop. In this way it is possible to obtain a measurable voltage drop with relatively small currents and hence to obtain high sensitivity.

Rectifier Instruments.—Currents of audio frequencies can be very satisfactorily measured by passing the current to be evaluated, or a portion of it, through a full-wave rectifier using elements of the copper oxide type and then measuring the rectified output on an ordinary direct-current instrument. Copper oxide meters of this type are relatively rugged, are permanent in their calibration, and can be made with a sensitivity that will give full-scale deflection on currents as low as 1 milliamp. Instruments making use of copper oxide rectifiers have practically no frequency correction up to about 1000 cycles, but at higher frequencies the capacity of the rectifier elements by-passes sufficient current to cause the instrument to read low. A typical calibration curve showing the percentage of error as a function of frequency is shown in Fig. 355.[1] Instruments of the copper oxide type tend to give an indication proportional to the average value of the wave and so have a wave-form error when harmonics are present.

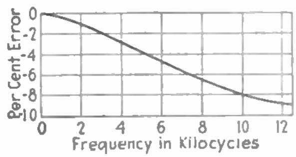


FIG. 355.—Calibration curve of a typical copper oxide rectifier type of alternating-current instrument, showing how the observed indication drops off at high frequencies as a result of the electrostatic capacity of the rectifier unit.

145. Audio- and Radio-frequency Voltages.—The methods most commonly employed to evaluate audio-frequency voltages are the copper

[1] An excellent discussion of the characteristics of instruments using copper oxide rectifiers is given by Joseph Sahagen, The Use of the Copper Oxide Rectifier for Instrument Purposes, *Proc. I.R.E.*, vol. 19, p. 233, February, 1931.

oxide rectifier voltmeter, the thermocouple voltmeter, and the vacuum tube voltmeter. With radio-frequency voltages the vacuum-tube voltmeter is almost standard except for frequencies below about 1000 kc, where thermocouple voltmeters can also be employed.

In the copper oxide rectifier type of voltmeter the unknown voltage is applied across the terminals of a standard resistance and the resulting current is measured by a copper oxide rectifier meter. Such meters have a high value of ohms per volt and are very satisfactory for audio frequencies. They cannot be used at radio frequencies, however, because the capacity of the copper oxide elements by-passes high-frequency currents around the rectifier. Voltmeters of the copper oxide rectifier type tend to give an indication which is proportional to the average value of the wave and as a result have wave-form errors when harmonics are present.

Thermocouple voltmeters consist essentially of a standard resistance to which the unknown voltage is applied, and a thermocouple for measuring the current that flows. Such instruments have a lower value of ohms per volt than the copper oxide type of voltmeter, but can be used up to relatively high frequencies, the limiting factor being the difficulty of obtaining a standard resistance free from inductance and capacity. With careful shielding, thermocouple voltmeters have been successfully used up to frequencies in excess of 1000 kc.[1] Thermocouple voltmeters give an indication which is proportional to the effective wave being measured irrespective of the wave form.

The Vacuum-tube Voltmeter.—Vacuum-tube voltmeters are essentially vacuum-tube rectifiers (detectors) in which the rectified direct current in the detector output is used as a measure of the voltage that is applied to the detector input. The voltages that can be measured in this way extend from about 0.01 volt upward, the power consumption is negligible, and there is virtually no frequency error up to the highest frequencies employed in radio communication. One of the greatest advantages of vacuum-tube voltmeters is that a calibration made at audio or power frequencies holds with negligible correction up to the highest radio frequencies. Vacuum-tube voltmeters can be of either the grid- or plate-rectification type. The former are more sensitive but less desirable from a number of points of view and so are employed only when the potential being measured is so small that the efficiency of plate rectification gives insufficient sensitivity.

Circuits of Vacuum-tube Voltmeters.—Vacuum-tube voltmeters of the plate-rectification type can employ a number of circuit arrangements. A simple example consisting of an ordinary vacuum-tube amplifier biased to approximately cut-off and with a milliammeter in the plate circuit is shown in Fig. 356. A condenser of large capacity is connected between

[1] Leon T. Wilson, A Novel Alternating-current Voltmeter, *Trans. A.I.E.E.*, vol. 43, p. 220, 1924.

plate and filament to by-pass the alternating component of the plate current around the milliammeter. When an alternating-current voltage is applied to the grid of such a tube the positive half-waves of the applied voltage produce impulses of plate current, while the negative half-waves are suppressed, with the result that the average plate current is increased by the presence of the applied alternating-current voltage, and the amount of increase as indicated on the milliammeter is a measure of the magnitude of the unknown voltage. The calibration curve of such an instrument is shown in Fig. 357, and upon careful examination it will be found that the increase in plate current caused by the presence of the applied voltage is proportional to the square of the potential being measured until large amplitudes are reached. In the range over which the square-law characteristic holds the plate-current increment produced by the voltage being measured is proportional to the square of the effective value of this voltage.

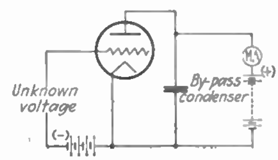


FIG. 356.—Circuit diagram of simple vacuum-tube voltmeter of the plate-rectification type.

While it is usually desirable to operate the vacuum-tube voltmeter at the cut-off grid bias, substantially the same characteristics including the same sensitivity are obtained over an appreciable range of grid-bias

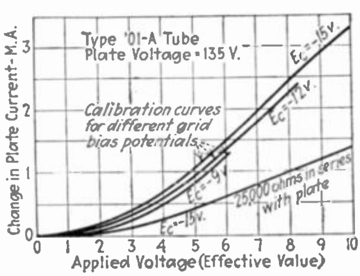


FIG. 357.—Calibration curves of typical vacuum-tube voltmeters of the plate-rectification type.

Fil battery can be used for balancing

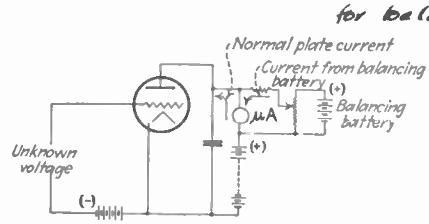


FIG. 358.—Circuit diagram of vacuum-tube voltmeter provided with balancing battery for balancing normal plate current out of the plate meter.

voltages, as is apparent from Fig. 357. If the bias is too low, however, the sensitivity is reduced, while if the bias is much greater than cut-off the instrument is capable of reading only very high voltages and becomes a crest voltmeter because plate current flows only during the peaks of the positive half-cycles.

While all vacuum-tube voltmeters of the plate-rectification type are based on the circuit of Fig. 356 it is possible to introduce a number of modifications that increase the usefulness of the instrument. A commonly used variation consists in balancing out the steady direct current that flows through the milliammeter in the circuit of Fig. 356

when no signal voltage is applied. An arrangement for doing this is illustrated in Fig. 358 and consists of a local battery, a resistance, and an adjustable potentiometer for passing through the milliammeter a current equal in magnitude to the residual direct-current plate current but in opposite direction. When this is done the deflections of the plate meter represent the increase in plate current produced by the unknown voltage applied to the grid. Another modification consists in placing a resistance in series with the plate milliammeter as shown in Fig. 359. The effect of this resistance is to decrease the sensitivity of the voltmeter and to cause the calibration curve to become more nearly linear, as is illustrated in Fig. 357. The instrument also tends to become a peak voltmeter because if the grid bias is at cut-off with no applied voltage, then the application of a large signal to the grid increases the current flowing through the plate resistance, reducing the plate voltage and causing the bias to exceed cut-off. Another form of the plate-rectifier type of voltmeter is also shown in Fig. 359, and is known as the slide-back voltmeter.

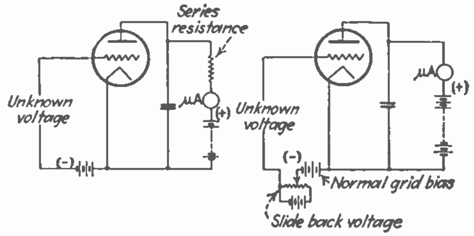


FIG. 359.—Commonly used types of vacuum-tube voltmeters. Both of these arrangements tend to give an indication proportional to the crest value of the unknown potential.

In this instrument the unknown voltage is measured as follows: The voltage to be evaluated is applied to the grid of the tube, causing an increase in the plate current. The negative grid bias is then increased until the plate current is reduced to the value existing before the application of the unknown voltage. The amount by which the grid bias must be increased to accomplish this result is then a measure of the peak value of the unknown voltage.

A very convenient form of vacuum-tube voltmeter which is made commercially and is capable of maintaining its accuracy of calibration over long periods of time is shown in Fig. 360.[1] This is essentially a direct-current bridge in which the direct-current plate-filament resistance of the voltmeter tube forms one arm, and the plate meter is the bridge detector. A single battery is employed to supply grid, filament, and plate potentials and to balance out the residual direct current in the plate meter. The bridge is prepared for operation by short-circuiting

[1] S. C. Hoare, A New Thermionic Instrument, *Trans. A.I.E.E.*, vol. 46, p. 541, 1927.

the input terminals and adjusting the resistance in series with the battery until the plate meter reads zero. The unknown voltage is then applied to the voltmeter, causing the direct-current plate-filament resistance of the tube to change, unbalancing the bridge and giving an indication on the milliammeter that is a measure of the effective value of the unknown voltage. Other forms of vacuum-tube voltmeters employing a single battery have been devised by Moulin, and by Jansky and Feldman.[1]

Several circuits for vacuum-tube voltmeters of the grid-rectification type are shown in Fig. 361. These differ from each other in the location of the grid leak and whether or not a balance battery is used for eliminat-

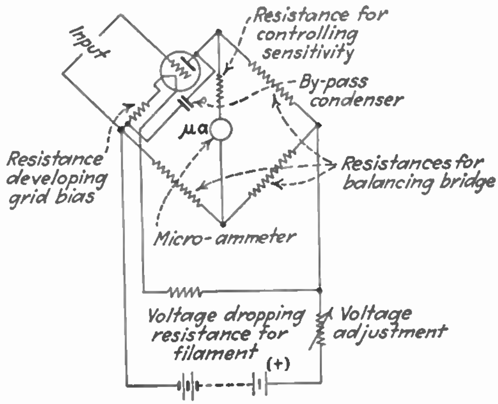


FIG. 360.—Circuit of convenient form of vacuum-tube voltmeter requiring only a single battery and capable of maintaining its calibration over long periods of time.

ing the steady plate current. The grid-rectification type of vacuum-tube voltmeter is much more sensitive to small alternating-current voltages than is the plate-rectification type of instrument, but consumes energy and has a large residual plate current and so is employed only when high sensitivity is essential. With small applied alternating-current voltages the grid-rectification type of vacuum-tube voltmeter gives a change of plate current proportional to the square of the effective value of the applied voltage, while with large applied voltages the instrument tends to become a peak voltmeter.

"*Turnover.*"—Reversing the input terminals of a vacuum-tube voltmeter will sometimes change the reading. This is known as "turnover" and is caused by even harmonics in the voltage wave that make the positive and negative half-cycles have a different shape and, in particular, different crest amplitudes. When the voltmeter gives an indication

[1] E. B. Moulin, "Radio-frequency Measurements"; C. M. Jansky and C. B. Feldman, A Two-range Vacuum-tube Voltmeter, *Trans. A.I.E.E.*, vol. 47, p. 307, January, 1928.

that tends to be proportional to the peak of the wave, and when this wave contains even harmonics, one can always expect to find the turnover effect present. In order to obtain accurate readings in such cases it is necessary either to modify the voltmeter so that the indications will depend upon the effective value of the voltage or to measure the apparent voltage for both polarities, and average the result.

Miscellaneous Features of Vacuum-tube Voltmeters.—The vacuum-tube voltmeter is the most important measuring instrument employed in radio work. When the plate circuit is properly by-passed with condensers so as to offer negligible impedance to the alternating currents being measured, the calibration is virtually independent of frequency, permitting calibrations made at low frequencies to be used at radio frequencies. The accuracy can be made better than 1 per cent and is sufficient for all ordinary purposes. The calibration is reasonably permanent provided the tube is always operated at the same electrode voltages, but frequent recalibration is advisable unless special precautions, such as those employed in the meter of Fig. 360, are provided for insuring the same electrode voltages at all times.

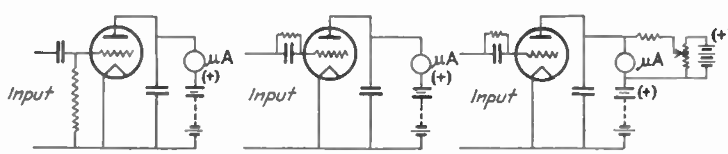


FIG. 361.—Circuit diagrams of typical vacuum-tube voltmeters of the grid-rectification type.

When the voltage to be measured is larger than can be conveniently applied to the grid of a vacuum tube, as for example when the voltage is in the order of several hundred volts, the most satisfactory measuring instrument is an inverted vacuum-tube voltmeter.[1] This instrument is ideally suited for such conditions and has proved very successful in practice.

A method used extensively in measuring small audio-frequency voltages and currents by means of a vacuum-tube voltmeter is shown schematically in Fig. 362. The measurement is made by applying the unknown voltage, or a voltage derived by passing the unknown current through a known resistance, to the amplifier, the gain of which is adjusted by suitable volume controls until the meter in the plate circuit of the vacuum-tube voltmeter gives a convenient deflection. The amplifier input is then connected to the output of a calibrated voltage divider, which is adjusted until the vacuum-tube voltmeter in the amplifier output gives the same indication as with the unknown voltage. The

[1] See F. E. Terman, The Inverted Vacuum Tube, *Proc. I.R.E.*, vol. 16, p. 447 April, 1928.

output of the voltage divider is then equal to the unknown voltage and can be evaluated in terms of the divider input and setting. Instruments of this sort are able to measure very small voltages and currents because of the amplification used and give good accuracy because the amplifier and vacuum-tube voltmeter are in effect calibrated for each reading.

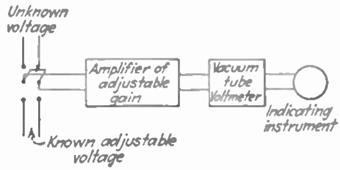


FIG. 362.—Schematic arrangement for measuring small audio-frequency voltages by comparing the unknown potential with a known adjustable alternating-current voltage of the same frequency.

Screen-grid tubes are sometimes used in vacuum-tube voltmeters but have little advantage for this purpose because the electrostatic shielding provided by the screen electrode is of no benefit when a large condenser by-passes the plate circuit of the tube to alternating currents. The screen-grid tube voltmeter has the disadvantage of requiring a screen-grid potential, which adds another factor that can affect the calibration.

146. Determination of Frequency.—The usual method of measuring an audio frequency is by means of an alternating-current bridge arranged so that the balance depends upon frequency. Examples of such bridges are shown in Fig. 363.

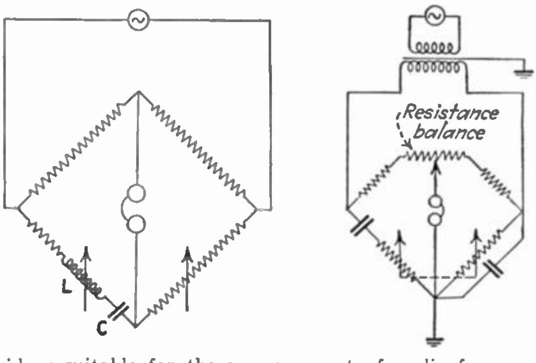


FIG. 363.—Bridges suitable for the measurement of audio frequencies. In both of these bridges the balance is dependent on both frequency and the circuit elements, so that if the latter are known the frequency can be calculated.

Radio frequencies are usually measured either by means of an adjustable calibrated tuned circuit or by comparison with a known frequency. In the case of very high frequency waves (very short wave lengths) the wave length can be directly measured by means of standing waves set up on a transmission line (Letcher wire method).

The tuned-circuit method of measuring frequencies makes use of a calibrated resonant circuit which is adjusted to resonance with the frequency to be measured. Such a device is commonly known as a wave meter. Resonance between the wave meter and the frequency to be measured can be detected by a low-resistance thermocouple milliammeter in series with the tuned circuit or coupled to it by a small coil, or by a neon light placed across the wave-meter condenser. These arrangements are shown in Fig. 364. Resonance between a wave meter and a small oscillator can also be detected by observing the effect of the wave meter on the grid or plate currents of the oscillator. When the meter is tuned to resonance with the oscillations, the energy which the wave meter absorbs causes the oscillator plate current to increase, and if the coupling between the wave meter and oscillator is close enough and the oscillator is small, the oscillations will stop at resonance. This phenomenon is often employed in rough measurements, but is not very accurate because of the close coupling required, and the consequent reaction on the oscillator frequency.

In place of the simple circuit of Fig. 364 wave meters sometimes employ a small variable condenser shunted by a much larger fixed capacity, as shown at Fig. 365*a*. In this way the frequency range that is covered by the condenser dial is reduced, thus spreading out the calibration curve.

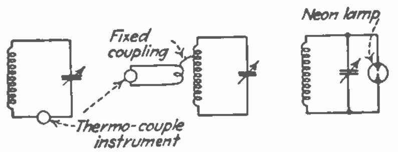


Fig. 364.—Methods of indicating when a wave meter is tuned to resonance with the frequency being measured.

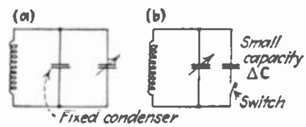


Fig. 365.—Special types of wave meters.

Another modification is shown at Fig. 365*b* and consists of an ordinary wave meter with the addition of a very small condenser ΔC which can be connected into the circuit by means of a push button. The adjustment is made so that the addition of ΔC detunes the wave meter as much on one side of resonance as the removal of ΔC detunes on the other side of resonance, as shown by no change in the resonance indicator. This arrangement increases the accuracy with which the wave meter can be adjusted to read an unknown frequency because it makes use of the steep sides of the resonance curve instead of the flat top.

The tuned circuit of the wave meter is sometimes used as the frequency-determining element of an oscillator, giving a "heterodyne wave meter." The most satisfactory type of heterodyne wave meter is the dynatron oscillator circuit of Fig. 178, in which a tuned circuit is shunted by the negative plate-cathode resistance of the dynatron. The chief advantages of the dynatron oscillator are the extreme simplicity and the fact that the tube adjustment has little effect on the frequency calibration.

General Radio says tuned circuit wavemeters are accurate to 0.5% with special precautions and 0.1% with extreme care. Accurate dials can usually be read closer than this.

The precision with which frequency may be measured by means of a wave meter depends upon the accuracy of the calibration and upon the exactness with which the wave meter can be adjusted to resonance with the frequency to be measured. The calibration depends in the first place upon the accuracy of the standard frequency against which the calibration was made and also upon the permanence with which the calibration is maintained. The exactness with which the wave meter can be adjusted to resonance with the unknown frequency depends upon the Q of the wave-meter tuned circuit. Under ordinary conditions a properly constructed and calibrated wave meter can be depended upon to measure frequency to within less than 1 per cent, and if extreme care is used in the construction, calibration, and maintenance, accuracies as great as 0.1 per cent can be obtained with reasonable certainty.

Comparison Method of Measuring Frequency.—The comparison method of frequency measurement makes use of a standard frequency which is known to a high degree of precision, and with which the unknown frequency is compared. By using harmonics (or if necessary, subharmonics) of the standard, it is possible to determine the unknown frequency with great accuracy. Thus if the standard frequency is 50 kc it is possible by harmonic generation to obtain known frequencies ranging from 50 to about 10,000 kc in 50-kc steps, and the unknown frequency can then be compared with the nearest of these harmonics. Interpolation between harmonics can be carried out in the following ways:[1]

1. Direct heterodyne methods in which the beat frequency between the unknown frequency and the nearest harmonic of the standard is measured directly by means of a wave meter, frequency bridge, calibrated oscillator, or other frequency-measuring device.
2. Direct ~~heterodyne~~ interpolation methods in which the fundamental frequency of an auxiliary oscillator is adjusted to zero beat first with the unknown frequency and then with the adjacent harmonics of the standard. The unknown frequency is then obtained by interpolation between the oscillator settings for the two known frequencies.
3. Successive heterodyning with harmonics and subharmonics of a standard frequency in such a way as to subtract known amounts from the unknown frequency until the residual is only a small fraction of the unknown frequency and need be measured only approximately.

The standard frequency required in the comparison method is always a carefully designed crystal-controlled oscillator operated with extreme care. The frequency of the standard can be accurately measured by an electric clock driven by a subharmonic of the standard. A typical circuit arrangement used to carry out the necessary operations is shown in Fig.

[1] For a thorough discussion of the first two general methods of interpolation see J. K. Clapp, Interpolation Methods for Use with Harmonic Frequency Standards, *Proc. I.R.E.*, vol. 18, p. 1575, September, 1930. The third method of interpolation is described at length by F. A. Polkinghorn and A. A. Roetken, A Device for the Precise Measurement of High Frequencies, *Proc. I.R.E.*, vol. 19, p. 937, June, 1931.

366, in which the standard frequency is generated by a 50-kc crystal oscillator provided with an oven and thermostat for keeping the crystal temperature constant. The output of this crystal oscillator is injected into the circuits of a multivibrator having a fundamental frequency approximating 10 kc and thus causing the multivibrator to synchronize at exactly one-fifth the crystal frequency. The 10 kc thus obtained is amplified and injected into a second multivibrator having a fundamental frequency of approximately 1 kc. The output of the 1-kc multivibrator, which has a fundamental frequency exactly one-fiftieth of the frequency of the crystal oscillator, is applied to a synchronous clock that keeps correct time when the frequency is exactly 1000 cycles and hence when the crystal plate vibrates at exactly 50 kc. Comparison of the clock time with observatory time makes it possible to evaluate the crystal frequency to a very high order of precision.[1]

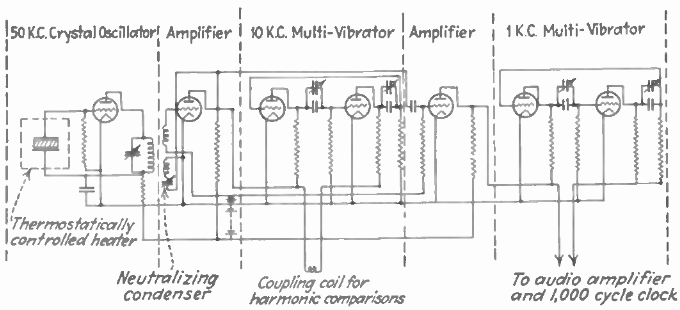


FIG. 366.—Circuit diagram of standard frequency outfit consisting of a 50-kc crystal and a two-stage multivibrator for deriving a frequency one-fiftieth of the crystal frequency for operating a 1000-cycle synchronous clock.

Letcher Wire Method of Measuring Frequency.—When the frequency is extremely high it is possible to measure the wave length directly by establishing standing waves along a transmission line consisting of a pair of parallel wires (often called Letcher wires). If one end of such a transmission line is short-circuited, and a voltage is applied at the other end, the current distribution shows decided resonances, as illustrated in Fig. 33, and the distance between the adjacent current minima or adjacent current maxima is almost exactly one-half wave length. This phenomenon can be used to measure the wave length of high-frequency waves by an arrangement such as is shown in Fig. 367. The transmission line

[1] This particular standard frequency equipment is described by L. M. Hull and J. K. Clapp, A Convenient Method for Referring Secondary Frequency Standards to a Standard Time Interval, *Proc. I.R.E.*, vol. 17, p. 252, February, 1929. The Bell Telephone Laboratories have also developed a somewhat similar standard frequency outfit differing only in detail and in refinements which make the frequency precision somewhat greater. See W. A. Marrison, A High Precision Standard of Frequency, *Proc. I.R.E.*, vol. 17, p. 1103, July, 1929.

is excited from an oscillator generating the frequency that is to be measured. A thermo-milliammeter is then shunted across the transmission line and moved along until a maximum current reading is obtained. This point is marked and the meter is again moved until the current is again maximum. The distance between these two points is then one-half wave length. The Letcher wire method of measuring frequency is extremely convenient in the case of very short waves, and with proper precautions will give a precision in excess of 1 part in 1000.[1]

147. Tube Characteristics.—The most important characteristics of the three-electrode tube are the amplification factor, plate resistance, and mutual conductance. In the case of screen-grid tubes the mutual conductance is by far the most important characteristic, but the plate amplification factor u_p, the screen-grid amplification factor u_{sg}, and the plate resistance may also be of interest.

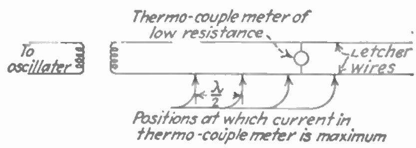


FIG. 367.—Letcher wire arrangement for directly measuring wave length.

Amplification Factor of Triodes.—There are three principal methods for determining the amplification factor of tubes, which for convenience will be referred to as the static-curve, incremental, and bridge methods. The approximate value of the amplification factor can be determined directly from the static curves of the tube such as those of Fig. 53 or 54. Since the amplification factor can be defined in terms of the increase in plate voltage which will counteract the effect which an increase in the negative grid bias has on the plate current, it is merely necessary to determine from the characteristic curves the plate-voltage increment that will maintain the plate current unchanged when an increment is added to the negative grid bias. The ratio of these two increments is the amplification factor. This method of evaluating the amplification factor has the disadvantage that small increments cannot be accurately read on the characteristic curves.

The increment method overcomes this disadvantage by accurately measuring the grid- and plate-voltage increments on meters instead of determining them from the characteristic curves. Typical circuit arrangements for carrying out the necessary operations are shown in Fig. 368, in which the voltage increment added to the grid is measured by

[1] See Eijiro Takagishi, On a Double Hump Phenomenon of Current through a Bridge across Parallel Lines, *Proc. I.R.E.* vol. 18, p. 513, March, 1930; Francis N. Dunmore and Francis H. Engels, A Method of Measuring very Short Wave Radio Lengths and Their Use in Frequency Standardization, *Proc. I.R.E.*, vol. 11, p. 467, October, 1923; August Hund, Correction Factor for the Parallel Wire System Used in Absolute Radio-frequency Standardization, *Proc. I.R.E.*, vol. 12, p. 817, December, 1924.

voltmeter V_g while the voltage increment that must be added to the plate potential in order to maintain the plate current constant is V_p. The plate current is indicated by meter *MA* which can be arranged to read the total plate current or may be provided with means to balance out the initial direct current. This latter arrangement permits greater accuracy since it allows the use of a very sensitive meter to detect any change in the plate current.

The bridge method of measuring amplification factor can be used in a number of forms, the most common being the Miller bridge shown in Fig. 369. When the ratio of resistances, R_1/R_2, is adjusted to give no sound in the telephone receivers the amplification factor is

$$\text{Amplification factor} = \frac{R_2}{R_1} \tag{184}$$

This equation results directly from the fact that when the current from the audio source is I_0, the alternating-current voltage I_0R_1 that is applied to the grid has its effect on the plate current balanced by the voltage I_0R_2 developed by I_0 flowing through R_2, so that by definition the amplification factor is I_0R_2/I_0R_1. In order to obtain a sharp balance it is

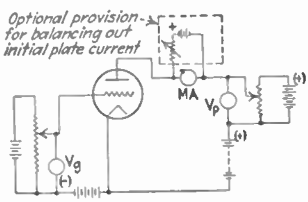


FIG. 368.—Circuit for measuring the amplification factor, plate resistance, and mutual conductance of a triode by the incremental method.

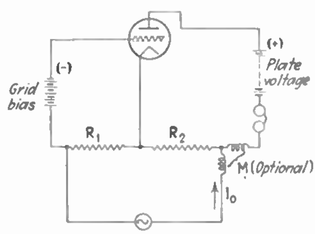


FIG. 369.—Miller bridge for measuring the amplification factor of triodes.

desirable to use a small mutual inductance *M* as shown to balance out quadrature currents resulting from electrode capacities. As ordinarily employed, the resistance R_1 is a fixed resistance of 10 ohms, while R_2 is variable and gives the amplification factor directly by shifting the decimal point. In making measurements it is desirable to use the smallest feasible inputs from the audio-frequency source in order to avoid overloading, with consequent production of harmonics.

Plate Resistance of Triodes.—The plate resistance of triodes can be determined by the characteristic-curve, voltage-increment, and bridge methods, as in the case of the amplification factor. Plate resistance can be determined directly from plate-current plate-voltage static curves, such as those of Fig. 54, by noting that the reciprocal of the slope of these curves is the plate resistance. A more accurate way of determining the slope of the characteristic is to add a voltage increment ΔE_p to the plate potential and read the resulting plate-current increment ΔI_p on a meter. The plate resistance is then $\Delta E_p/\Delta I_p$. The necessary operations can be carried out using the circuit of Fig. 368.

The bridge method of measuring plate resistance consists in placing the plate-cathode circuit of the vacuum tube in the "unknown" arm of a resistance bridge excited from an audio-frequency source. When the bridge is balanced in the usual way the resulting apparent "unknown" resistance is the resistance offered to the small superimposed alternating-current voltage and hence is the dynamic plate resistance. A typical circuit arrangement for the bridge measurement of plate resistance is shown in Fig. 370 and includes a condenser C for the purpose of balancing out the tube capacities. When the plate resistance is high it is desirable to add a Wagner earth connection. In using the bridge method the alternating-current input to the bridge should be kept low, and the

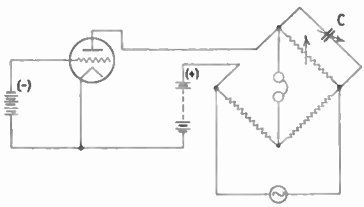


FIG. 370.—Circuit for measuring the dynamic plate resistance of a triode with an alternating-current resistance bridge.

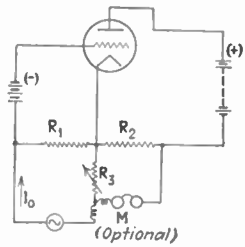


FIG. 371.—Bridge circuit for measuring mutual conductance.

resistance in the bridge arms should not be great enough to develop large direct-current voltage drops as a result of the direct-current plate current which must flow through the bridge.

Mutual Conductance.—The mutual conductance can be determined from the grid-voltage plate-current characteristic curves by noting the increment ΔI_p of plate current that results when a voltage ΔE_g is added to the grid potential. The ratio $\Delta I_p/\Delta E_g$ is then, by definition, mutual conductance. As in the case of amplification factor and plate resistance, the accuracy can be increased by measuring the increments ΔE_g and ΔI_p directly on indicating instruments as shown in Fig. 368, rather than by obtaining them from the characteristic curves of the tube.

The bridge method of measuring mutual conductance employs the circuit shown in Fig. 371. When the resistance R_3 is adjusted to give minimum sound in the telephone receivers the mutual conductance in mhos is given by the equation

$$\text{Mutual conductance in mhos} = \frac{R_3}{R_1R_2} \qquad (185)$$

This is derived as follows: The current I_0 from the audio-frequency source produces in R a voltage I_0R_1 that is applied to the grid. This develops an alternating-current plate current $I_{ac} = g_mI_0R_1$, which in flowing through R_2 produces a voltage drop $g_mI_0R_1R_2$. When no sound is heard in the phones $g_mI_0R_1R_2 = I_0R_3$, and Eq. (185) follows at once. In order to improve the sharpness of balance it is usually desirable to employ a small mutual inductance as indicated in the figure in order to compensate for stray capacities. In practice R_1 is usually 1000 ohms, while R_2, which must be kept low in order to minimize direct-current voltage drops, is ordinarily 100 ohms. The resistance R_3 is variable, and its value at balance multiplied by 10 will then give the mutual conductance directly in micromhos. In order to avoid errors it is necessary that the output of the audio-frequency oscillator be kept low.

Characteristics of Screen-grid Tubes.—The amplification factor, plate resistance, and mutual conductance of the screen-grid amplifier can be

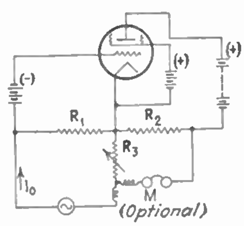


FIG. 372.—Bridge circuit of Fig. 271 modified to measure the mutual conductance of screen-grid tubes.

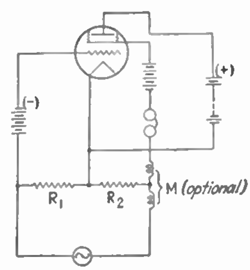


FIG. 373.—Modified Miller bridge arranged to measure the amplification factor μ_{sg} of the screen grid.

determined by the same methods used with triodes. Bridge methods for the determination of the amplification factor and plate resistance are not recommended, however, since the very high values which these constants have in screen-grid tubes make it difficult to avoid errors. The most satisfactory way of measuring the characteristics of screen-grid tubes is to use the bridge method illustrated in Fig. 372 for determining the mutual conductance and then to evaluate the plate resistance by the increment method analogous to that illustrated in Fig. 368. If the amplification factor is desired this can be most conveniently found by the formula

$$u_p = g_mR_p \tag{186}$$

The screen-grid amplification factor μ_{sg}, which represents the relative ability of screen- and control-grid potentials to produce electrostatic fields at the surface of the cathode is of the same order of magnitude as the amplification factor of triodes and so can be determined by modifying the Miller bridge, as shown in Fig. 373, or can be evaluated by the increment or static-curve method.

Detection Coefficients.—The method of evaluating the efficiency of rectification and the effective plate resistance of the anode power detector was outlined in Sec. 60 and need not be repeated here. The detection constants v for weak-signal grid rectification, and v_m or v_p for weak-signal plate detection can be determined as described in Sec. 62 and 63. The detection characteristics of a vacuum tube can also be measured by applying to the detector input a carrier wave of known amplitude and degree of modulation and then measuring the modulation-frequency voltage that appears in the detector output.[1]

148. Voltage Amplification.—Audio-frequency voltage amplification is usually determined by applying a known potential to the amplifier input and measuring the output voltage with a vacuum-tube voltmeter. A typical circuit arrangement for carrying out these steps is illustrated

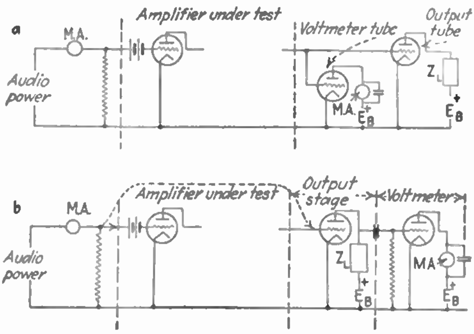


FIG. 374.—Methods of measuring the voltage amplification of audio-frequency amplifiers.

at Fig. 374*a*. In making such measurements it is desirable to control the input by passing a known current through a resistance that is varied until a predetermined output is obtained. In this way the danger of overloading the amplifier is avoided, and it is necessary to determine only one point on the vacuum-tube voltmeter calibration.

At Fig. 374*a* the input capacity of the voltmeter tube, while small (usually 15 to 20$\mu\mu$f), is still sufficiently large to affect the amplification at the higher audio frequencies. This disadvantage can be overcome by arrangements of the type illustrated at Fig. 374*b*, in which the vacuum-tube voltmeter is placed across an impedance in the plate circuit of the output tube. The voltage applied to the amplifier input is adjusted until a convenient scale reading is obtained on the output meter. The grid of the output tube is then disconnected from the amplifier and brought directly to the input voltage which is adjusted to give the same

[1] A description of apparatus suitable for making such measurements is given by E. L. Chaffee and G. H. Browning, A Theoretical and Experimental Investigation of Detection for Small Signals, *Proc. I.R.E.*, vol. 15, p. 113, February, 1927.

meter reading as before. The voltage amplification is then the ratio of the input voltages for the two conditions.

When the phase shift as well as the voltage amplification is desired, the balance method illustrated in Fig. 375 should be employed. Here the output voltage of the amplifier is balanced by means of an equal and opposite voltage derived with the aid of a resistance and a mutual inductance. Balance is most conveniently determined on telephone receivers operated from an amplifier, and in order to avoid errors introduced by capacities to ground the telephone receiver is preferably connected to the secondary of a shielded transformer. The amplifier used for detecting balance should be built in a compact portable unit isolated as far as

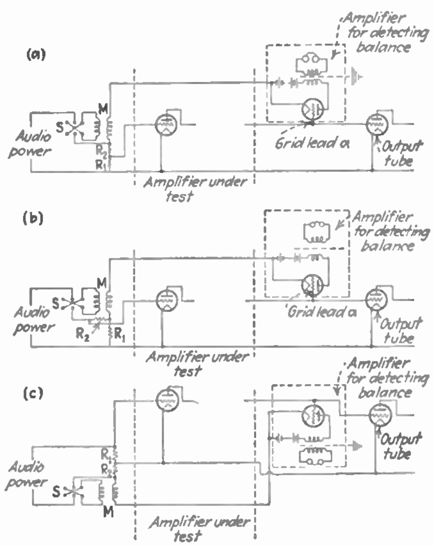


FIG. 375.—Various forms of balance circuit for determining vector ratio of output to input voltages of an audio-frequency amplifier.

possible from ground and with the shortest possible grid lead a, while the input and balance voltages should be obtained by passing a relatively large input current through small mutual inductances and resistances. The exact circuit arrangement depends upon the relative phase and magnitudes of the input and output voltages, and it may be necessary to change circuits for different frequencies. When a condition of balance is obtained the vector amplification for the different circuits of Fig. 375 is given by the following formulas[1]

[1] Other types of balance circuits for testing audio-frequency amplification are given by H. Diamond and J. S. Webb, The Testing of Audio-frequency Transformer-coupled Amplifiers, *Proc. I.R.E.*, vol. 15, p. 769, September, 1927. The arrangements described in this paper all have the disadvantage of shunting the amplifier output with the input capacity of an auxiliary tube, however, and so are not so satisfactory as the circuits of Fig. 375.

For circuit a:

$$\text{Vector amplification} = \frac{(R_1 + R_2) \pm j\omega M}{R_1}$$

For circuit b:

$$\text{Vector amplification} = \frac{R_1 \pm j\omega M}{R_1 + R_2} \tag{187}$$

For circuit c:

$$\text{Vector amplification} = \frac{R_2 \pm j\omega M}{R_1}$$

The choice of sign is determined by the position of the switch S.

Load Limit of Amplifiers.—The load limit of an audio-frequency amplifier, *i.e.*, the input at which amplitude distortion becomes excessive, cannot be determined objectively since "excessive distortion" is a qualitative term. The most convenient ways of detecting overloads in an amplifier consist in testing for grid current with a microammeter, and in observing the change in direct-current plate current drawn by the amplifier as the signal voltage is applied and removed. The presence of grid current means that the input voltage is large enough to make the grids positive during part of each cycle, while a change in plate current indicates that distortion is being produced in the plate circuit.

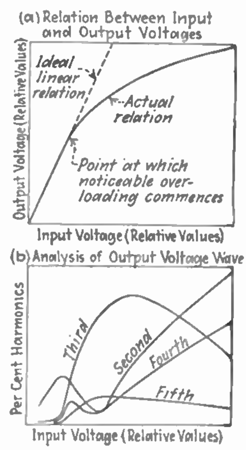


FIG. 376.—Variation of magnitude and distortion of output voltage as the input voltage of an amplifier is varied.

The characteristics of a vacuum tube are such that when there is amplitude distortion the output voltage is no longer proportional to the input voltage, and this fact can be used to detect the point at which distortion becomes appreciable. The procedure is simply to plot a curve showing output voltage or current as a function of input voltage, maintaining the frequency constant. The overload point is indicated by a departure from linearity as shown at Fig. 376a.[1] The exact character of amplitude distortion in amplifiers can be determined only by analyzing the wave shape of the output, but this requires equipment not usually available. An analysis of the output wave of a typical amplifier is shown in Fig. 376b, where it is observed that at low loads the distortion consists largely of a second-harmonic component but that as the overload point is reached the third and other odd harmonics become prominent. In a push-pull amplifier these even harmonics are

[1] An ingenious method for making use of this principle to test amplifiers on a production basis is given by Arthur E. Thiessen, The Accurate Testing of Audio Amplifiers in Production, *Proc. I.R.E.*, vol. 18, p. 231, February, 1930.

eliminated entirely, thus giving substantially no distortion up to the overload point.

Measurement of Radio-frequency Amplification.—The measurement of radio-frequency amplification presents difficulties not encountered in audio-frequency amplifiers since at high frequencies the amplification is much more sensitive to capacities introduced by the measuring equipment. Vacuum-tube voltmeters can be used when the amplification is small, but with large amplifications, as in the case of sensitive radio receivers, a standard signal generator must be employed. The proper circuit arrangements and test procedure will depend upon the special circumstances involved, and a general procedure applicable to all cases cannot be laid down.

149. Receiver Characteristics.—The technique for measuring the characteristics of radio receivers was outlined briefly in Sec. 108, and is described in detail in the *Reports* of the Standardization Committee of the Institute of Radio Engineers. In brief the measurements make use of a standard signal generator capable of delivering adjustable known carrier voltages covering the range 1 to 500,000μv, with means for modulating to a known degree. The receiver sensitivity is defined as the input that must be supplied from the signal generator to obtain an arbitrary audio-frequency output to the loud-speaker. The selectivity is obtained by determining the input voltage required to develop the standard output as the signal generator is detuned from the frequency the receiver is adjusted to receive. The fidelity, *i.e.*, frequency distortion, is evaluated by setting the signal generator to give a standard output when the modulation frequency is 400 cycles and the receiver is tuned to the carrier wave of the signal generator. The modulation frequency is then altered without changing the degree of modulation or the carrier amplitude, and the variation in the output voltage of the receiver noted.

Standard signal generators are carefully shielded oscillators provided with means of obtaining a known degree of modulation. The output of the oscillator is applied to an attenuation network consisting of a network of resistances, a calibrated mutual inductance, or a small inductance of known value, and the current flowing into this attenuator is measured by means of a thermocouple milliammeter. When proper shielding is employed, and when the attenuator is arranged to avoid errors from stray capacitive and inductive couplings, it is possible to obtain known voltages as small as 1μv with reasonable accuracy.

The actual construction of a standard signal generator represents a formidable task since the shielding problems involved are very severe. For this reason such equipment is usually developed in large laboratories after extensive preliminary developmental work, the cost of which can be spread over a large number of units.

150. Field-strength Measurements.[1]—The strength of radio waves is most satisfactorily measured by comparing the voltage that the radio waves induce in a receiving antenna of calculable effective height with an adjustable known voltage inserted in the antenna. The measurement of field strength is schematically illustrated in Fig. 377. The received signals are delivered to a sensitive receiver (usually of the superheterodyne type), which is provided with a milliammeter in the plate circuit of the detector for indicating relative amplitudes of received carrier voltages. A loop antenna is employed, since the effective height of a loop can be accurately computed. The measuring procedure is to turn off the standard signal generator, adjust the loop for maximum reception from the distant station, and then set the receiver gain to give a convenient scale reading on the output milliammeter. The loop is then turned at right

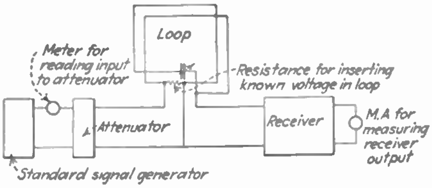


FIG. 377.—Schematic diagram illustrating the measurement of field strength by the substitution method using a loop antenna, a sensitive radio receiver, and a standard signal generator.

angles so that no energy will be picked up from the distant transmitter, and the signal generator is used to insert in series with the loop a voltage that gives the same receiver output as that obtained from the passing wave. The field strength is then the voltage output of the standard signal generator divided by the effective height of the loop.

The difficulty of shielding the standard signal generator is such that at frequencies in excess of 1500 kc it is impractical to produce extremely small voltages of known amplitude. This limitation can be overcome by taking advantage of the fact that when the first detector of a superheterodyne receiver is operated on the square-law part of its characteristic, the amplitude of the intermediate frequency output of the first detector is proportional to the signal amplitude. This makes it possible to introduce a multiplying factor in the first detector by reducing the intermediate-frequency amplification in known amounts with an attenuation network. The voltage which a weak wave induces in the loop can thus be compared with a standard signal 100 or 1000 times as great by reducing the intermediate-frequency amplification by a factor of $1/100$ or $1/1000$, respectively, when making the comparison.

[1] An excellent survey of field-strength measuring equipment is given by C. R. Englund and H. T. Friis, Methods for the Measurement of Radio Field Strength, *Trans. A.I.E.E.*, vol. 46, p. 492, 1927.

151. Degree of Modulation.—The degree of modulation possessed by a wave depends upon the average amplitude of the envelope and the amount by which the envelope amplitude varies above and below the average. In terms of the notation illustrated in Fig. 378 the degree of modulation for the peaks and troughs is E_1/E_0 and E_2/E_0, respectively. When the modulation envelope does not contain even harmonics the degree of modulation is the same for the peaks and troughs, but when even harmonics are present in a modulation envelope there will usually, although not always, be a difference.

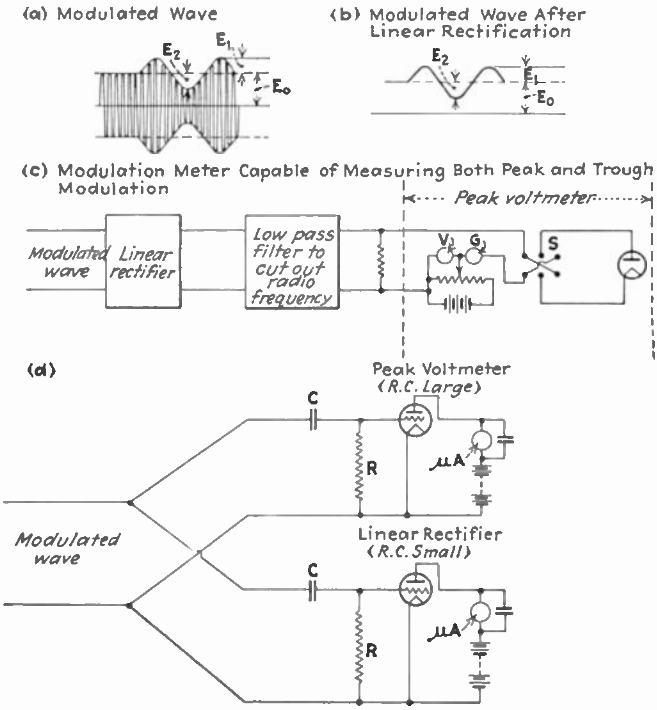


FIG. 378.—Circuit diagrams for measuring degree of modulation in two ways, together with wave forms involved.

The degree of modulation possessed by a wave can be determined by rectifying with a linear detector, which gives a current that varies in accordance with the modulation envelope. The direct-current component of the rectified wave is proportional to the average envelope amplitude and so is a measure of E_0, while the maximum and minimum amplitudes determine the peak and trough modulations and can be measured by means of the crest voltmeter illustrated in Fig. 378. This consists of a rectifier tube (which can be a triode with the grid and plate connected together) in series with a bias battery and a galvanometer.

The bias voltage is adjusted until the galvanometer indicates the current flowing through the rectifier is just approaching zero. The bias voltage as read by the voltmeter V then represents either the maximum or minimum of the rectified wave according to the position of the switch S.

When the modulation envelope contains no even harmonics the degree of modulation can be measured by utilizing a linear detector to determine the average amplitude of the modulated wave, and a peak voltmeter operating directly on the radio-frequency wave to give the peak amplitude. These vacuum-tube voltmeters can be two power detectors of the grid-rectification type as shown in Fig. 378, one of which employs a grid-leak and grid-condenser combination having a low ratio of R/X (small RC) in order to give distortionless linear rectification, while the other has a large ratio R/X (large RC) and so measures the peak value of the modulated wave.

Distortion in the production or subsequent amplification of a modulated wave can ordinarily be detected by modulation measurements. The procedure is to employ a sinusoidal modulating voltage and then either to measure the peak and trough modulation, or to observe whether the carrier amplitude changes when the modulation is removed. Distortion is present if the carrier amplitude depends upon the modulation present, or if the peak and trough modulations are not the same.

152. Wave Form.—Practically all determinations of wave form that are made in radio work are concerned with audio-frequency waves and have as their object the measurement of distortion frequencies, cross-talk components, etc. Wave-form investigations of radio-frequency waves are seldom necessary since practically all circuits handling radio-frequency power are resonant circuits and so discriminate very strongly against harmonics that result from amplitude distortion. When it is desired to investigate the character of a wave consisting of a number of components of slightly different frequencies, such as a modulated wave, the customary procedure is to reduce the frequency of the wave to be investigated to an audio frequency by heterodyning with a local oscillator.

The most obvious method of determining the wave form of an audio-frequency voltage or current is by the use of an oscillograph. This has serious disadvantages, however, since the oscillograph does not perform satisfactorily at the higher audio frequencies and is not capable of indicating small percentages of distortion components. These limitations are overcome in two types of wave analyzers developed by the Bell System.[1] In one of these analyzers the wave that is to be measured

[1] R. G. McCurdy and P. W. Blye, Electrical Wave Analyzers for Power and Telephone Systems, *Trans. A.I.E.E.*, vol. 48, p. 1167, October, 1929; C. R. Moore and A. S. Curtis, An Analyzer for the Voice Frequency Range, *Bell System Tech. Jour.*, vol. 6, p. 217, April, 1927; A. G. Landeen, Analyzer for Complex Electric Waves, *Bell System Tech. Jour.*, vol. 6, p. 230, April, 1927.

is amplified, after which the frequency component to be determined is selected by means of suitable tuned circuits and measured on a vacuum-tube voltmeter. In the other the unknown voltage is heterodyned with a carrier wave of adjustable frequency and the resultant heterodyne signal applied to a square-law detector. The output of this detector is passed through a sharply resonant circuit tuned to 11,000 cycles (a mechanical resonator is used) and is then measured by means of a vacuum-tube voltmeter. The amplitude of any frequency component of an unknown wave can be determined by simply adjusting the local oscillator to a frequency such that the difference between the frequency of the component to be measured and the local oscillator is 11,000 cycles.

Instruments such as those just mentioned, while very satisfactory, are relatively expensive. A method of analyzing wave forms that does not involve any special equipment is illustrated in Fig. 379.[1] The unknown voltage is combined with a much larger voltage from an audio-frequency oscillator and applied to the grid of a vacuum-tube voltmeter that gives an output proportional to the square of the input voltage. The local oscillator is adjusted until its frequency differs by less than one cycle from the frequency of the voltage component to be measured. When this situation exists the needle of the milliammeter in the plate circuit of the vacuum-tube voltmeter will swing slowly at the beat frequency with an amplitude that is proportional to the amplitude of the frequency component responsible for the beat. By calibrating the arrangement with voltages of known amplitudes it is possible to determine the amplitude of small components of a wave with unusually good accuracy.

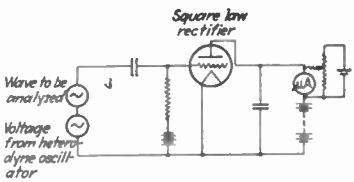


FIG. 379.—Method of analyzing wave form which consists in superimposing a voltage upon the unknown wave, applying the resultant heterodyne signal to the grid of a square-law vacuum-tube voltmeter, and observing the amplitude of the swings in the plate milliammeter as the frequency of the superimposed voltage is adjusted to approach zero beat with the frequency component to be measured.

153. Beat-frequency Oscillators.—A beat-frequency oscillator is a generator of audio-frequency energy which operates by heterodyning two radio-frequency voltages of slightly different frequencies. The usefulness of such an oscillator lies in the fact that by varying one of the radio frequencies through a range of about 10 kc, which can be readily accomplished by the turn of a single dial associated with a variable condenser, while keeping the frequency of the other oscillator fixed, the resulting audio frequency can be varied continuously through the entire range of audio frequencies.

[1] See Chauncey Guy Suits, A Thermionic Voltmeter Method for the Harmonic Analysis of Electrical Waves, *Proc. I.R.E.* vol. 18, p. 178, January, 1930.

In designing a beat-frequency oscillator it is necessary to give attention to several details if satisfactory results are to be obtained. In the first place there is a tendency for the two radio-frequency oscillators to synchronize automatically when their frequencies are very nearly the same, *i.e.*, when low audio frequencies are being produced. In order to avoid this tendency it is necessary to avoid coupling between the two circuits either by placing buffer amplifiers between the rectifier and oscillator tubes or by the use of a balanced input circuit as shown in Fig. 380. It is also necessary to place a low pass filter in the output of the heterodyne detector to prevent radio-frequency voltages from reaching the audio-frequency amplifier. The omission of this filter causes the audio-frequency amplifier to become overloaded and results in the production of unwanted audio frequencies. In order to obtain a sine-wave audio-frequency voltage having an amplitude that is independent of frequency, the coupling from the radio-frequency oscillators to the

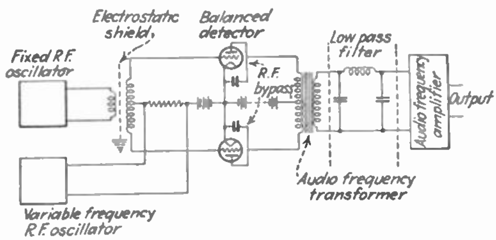


FIG. 380.—Schematic circuit diagram of beat-frequency oscillator, showing balanced detector which avoids coupling between the two radio-frequency oscillators.

heterodyne detector should be such that the fixed-frequency oscillator supplies a small voltage free of harmonics, while the variable-frequency oscillator supplies a large voltage. If the detector has a linear characteristic the beat-frequency output will then have an amplitude independent of the variable-frequency oscillator (see Sec. 65), while the wave shape of the beat-frequency output will be sinusoidal.

The frequency stability of the output of a beat-frequency oscillator is relatively poor because a small percentage change in the frequency of one radio oscillator will have a large influence on the comparatively small difference frequency. The effect of such things as temperature, plate voltages, etc., can be minimized by making the two radio-frequency oscillators as nearly alike as possible, so that they will behave in the same way under varying conditions. Beat-frequency oscillators are also always provided with a trimming condenser that can be adjusted so that a specified point on the frequency calibration is correct. This point is obtained either by comparing the output frequency with a tuning fork or by using the zero beat point.

154. Cathode-ray Oscillograph.—The cathode-ray oscillograph is the only satisfactory method available for the observation of wave shapes at radio frequencies. A typical cathode-ray oscillograph is shown in Fig. 381.[1] It consists of an electron-emitting cathode surrounded by a cylinder that is negative with respect to the cathode. This cylinder focuses the emitted electrons into a beam that is attracted to the anode

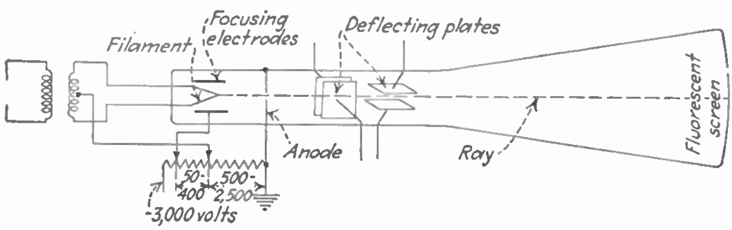


FIG. 381.—Schematic illustration of a typical cathode-ray oscillograph.

electrode. Some of these electrons pass through the small hole in the anode and form a concentrated ray that passes on to the fluorescent screen, where the impact produces a spot of light. By subjecting the electrons in the concentrated ray to transverse deflecting forces produced either by voltages applied to the deflecting plates shown in the figure, or by magnetic fields, the luminous spot on the screen will trace out a pattern

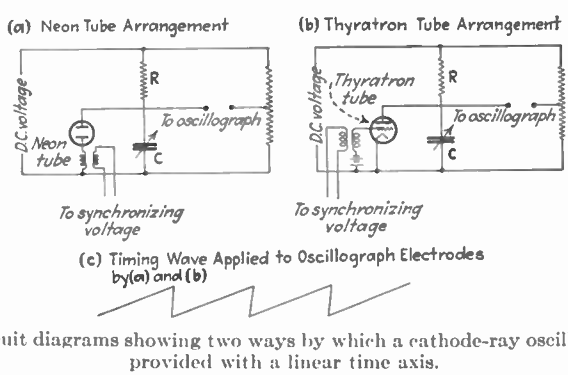


FIG. 382.—Circuit diagrams showing two ways by which a cathode-ray oscillograph may be provided with a linear time axis.

determined by the wave shapes of the deflecting forces even when these are high radio frequencies.

The possible uses of the cathode-ray oscillograph are too numerous to enumerate. It can be used to synchronize frequencies in harmonic relation, to measure degree of modulation, to detect distortion, to measure phase relations, etc.

[1] See Manfred von Ardenne, A Braun Tube for Direct Photographic Recording, *Exp. Wireless and Wireless Eng.*, vol. 7, p. 66, February, 1930. Another type of cathode-ray tube is described by J. B. Johnson, A Low-voltage Cathode-ray Oscillograph, *Bell System Tech. Jour.*, vol. 1, p. 142, Oct., 1922.

vol. 63 Jl of I.E.E. for best summary of C-R tubes

When observing a cyclically repeating phenomenon on a cathode-ray oscillograph it is desirable to obtain a linear time axis by applying a saw-toothed voltage wave, such as shown at Fig. 382*c*, to one pair of deflecting plates, and applying the phenomenon to be observed on the other pair of deflecting plates. If the period of the saw-toothed wave is made an exact submultiple of the period of the phenomenon being investigated then the spot on the fluorescent screen will retrace the same path each cycle and will give a stationary pattern on a linear time base similar to the records made with an ordinary Duddel type of oscillograph.

A saw-tooth wave such as shown at Fig. 382*c* can be generated with the aid of gas tubes as shown at Fig. 382*b*.[1] The frequency of the saw-tooth wave can be brought to approximately the required value by adjusting the capacity C and may then be brought into exact synchronism by injecting a small voltage from the circuits being investigated, as indicated in the circuit diagrams.

[1] A. L. Samuel, A Method of Obtaining a Linear Time Axis for a Cathode-ray Oscillograph, *Rev. Sci. Instruments*, vol. 2, p. 532, September, 1931; F. Bedell and H. J. Reich, The Oscilloscope, *Trans. A.I.E.E.*, vol. 46, p. 546, 1927.

CHAPTER XVIII

SOUND AND SOUND EQUIPMENT

155. Characteristics of Audible Sounds.—Sound is a mechanical vibration lying within the frequency range to which the ear responds, and is characterized by pitch, loudness, and phase. The pitch represents the frequency and is expressed in cycles per second. The loudness depends upon the amplitude of the wave and is measured either in terms of the pressure in bars (dynes per square centimeter) that is produced by the sound waves, or as the ratio of energy contained in the actual sound to the energy contained in the weakest audible wave of the same frequency. The phase of the wave gives the instant at which the vibration passes through zero. Sound travels in air at a velocity of approximately 1130 ft. per second.

Most actual sounds are complex waves containing several components having frequencies which are in harmonic relation to each other. The fundamental frequency corresponding to these harmonics (or overtones) is called the "pitch" of the complex sound, while the presence of the harmonics determines what is commonly termed the "quality" of the sound. Changing the relative amplitudes of the different components without changing their frequencies hence merely changes the quality without affecting the pitch.

Ordinary sound can be classified as speech, music, or noise.

Speech.[1]—Speech represents those sounds employed in conversation, and can be subdivided into vowels, diphthongs, transitionals, semi-vowels, voiced and unvoiced fricative consonants, and voiced and unvoiced stop consonants. Table XIII gives such a classification of the speech sounds involved in the English language.

The vowels are relatively powerful sustained sounds produced by using the lungs as bellows and setting the resulting air stream in vibration by means of the vocal cords. The resulting sound is rich in harmonics, and the various vowel sounds are produced by positioning the lips, tongue, etc., so that the throat, mouth, and nasal cavities reinforce harmonics lying within certain frequency ranges as a result of resonance, while tending to suppress other harmonics. The frequency spectrums of two vowels each pronounced by a male and a female voice are shown in

[1] For a more detailed discussion of the nature of speech sounds the reader should consult Harvey Fletcher, "Speech and Hearing," D. Van Nostrand Company, New York.

Fig. 383. It will be observed that each vowel is characterized by a reinforcement of the harmonics within certain frequency ranges, and that the distinguishing difference between the vowels is in the location of these regions of reinforcement, rather than the pitch of the complex sound. The fundamental pitch of the vocal cords varies considerably

TABLE XIII.—CLASSIFICATION OF SPEECH SOUNDS

Pure vowels
- Long: ū(tool), ō(tone), ó(talk), a(far), ā(tape), ē(team)
- Short: u(took), o(ton), á(tap), e(ten), i(tip)

Diphthongs: i, ou, oi, ew

Transitionals: w, y, h

Semi-vowels: l, r, m, n, ng

Fricative consonants

Voiced	Unvoiced
v	f
z	s
th(then)	th(thin)
z̄h(azure)	sh

Stop consonants

Voiced	Unvoiced
b	p
d	t
j	ch
g	k

with different individuals and also changes somewhat during speech, but is about 125 cycles for a normal male voice and about twice as high for a normal female voice. The pitch tends to increase as the voice intensity rises, and in shouting may reach twice the normal value. This

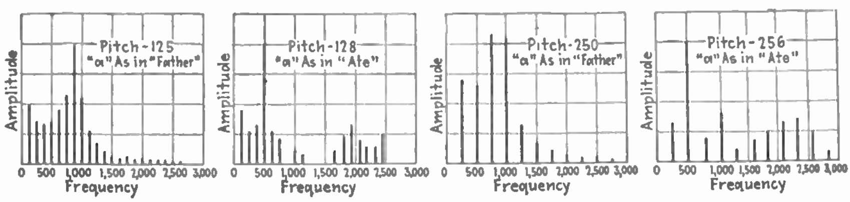


FIG. 383.—Frequency spectrum of typical vowel sounds when spoken at different pitches. Note that the distinguishing feature of the vowels is the frequency regions that are reinforced and not the pitch.

fact makes it possible to estimate the original level of speech independently of the loudness of the sound reaching an observer. The frequency ranges in which harmonics are reinforced in the different vowel sounds lie between 375 and 2400 cycles.

The diphthongs are a combination of two vowels in which the change from one vowel sound to another is accomplished without pause by changing the position of the mouth. The transitionals represent particular ways of beginning the vowel sound. The semivowels differ from the true vowels in that the passage from the vocal cords to the outside air is partially blocked either by forcing the sound wave to flow around the tongue or through the nasal cavity.

Fricative consonants are characterized by the rushing sound of air forced through a very small outlet and may be either voiced or unvoiced, depending on whether or not the vocal cords are allowed to vibrate. The fricative consonants, particularly when unvoiced, contain important components having frequencies as high as 5000 or 6000 cycles per second.

The stop consonants represent short wave trains formed by the lips and tongue and may be either voiced or unvoiced depending upon whether or not the vocal cords participate in making the sounds.

Power of Speech.—The power of average conversational speech varies widely with different persons, but under average conditions for American speech averages approximately 10 microwatts. If the silent intervals during conversation are excluded the average is increased to approxi-

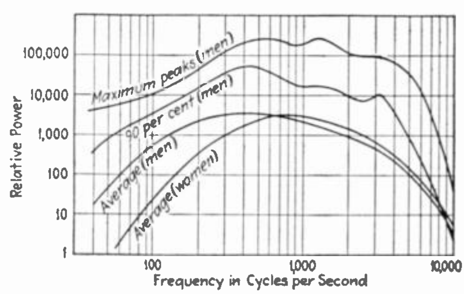


FIG. 384.—Distribution of average and peak speech energy over the frequency spectrum. When the speech is divided into time intervals of ⅛-sec. duration, the peak intensity in 90 per cent of these intervals does not exceed the value given by the curve marked "90 per Cent."

mately 15 microwatts. This represents an extremely small amount of energy, as is apparent when it is realized that the total average power that would be produced by all inhabitants of the earth talking simultaneously is less than the power radiated from a large broadcasting station. When one talks as loudly as possible the average speech power rises to about one hundred times the normal average, while in speaking with as weak a voice as possible without whispering the average power is reduced to approximately one-hundredth. The peak power of the loudest sound encountered in conversation is in the order of 5000 microwatts, while the power of a faint whisper is in the order of 0.01 microwatt. This represents an intensity range of 500,000 to 1. The pure vowels and the diphthongs are the most powerful sounds, while the *th* as in *thin* is the weakest.

The distribution of speech power over the audible frequency range is approximately as shown in Fig. 384. In the case of male voices the most powerful sounds are more or less evenly distributed throughout the frequency range 500 to 1500 cycles, while the average power is greatest at approximately 500 cycles. The higher frequency peaks, while about as intense as the lower frequency peaks, are less numerous. This is apparent from the 90 per cent curve of Fig. 384, which indicates that when the speech is divided into intervals of ⅛ sec., the peak intensity in 90 per cent of the intervals is less than the value given by the curve. In the case of women's voices the higher fundamental pitch tends to shift the energy distribution toward higher frequencies.

A general picture of the frequencies and intensities involved in the different speech sounds as used in normal conversation is shown in Fig. 385. Sounds which have several important frequency ranges appear more than once in the diagram. Powerful sounds are at the top of the figure, weak sounds at the bottom, while positions to the right or left represent important high- or low-frequency components, respectively.

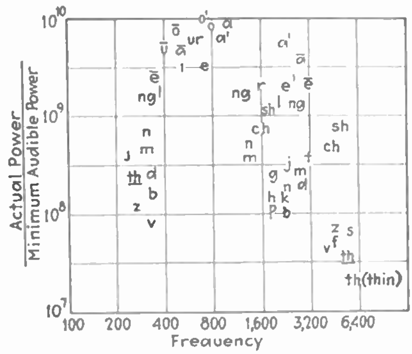


FIG. 385.—Chart showing intensities and region of most important frequency components of fundamental speech sounds. (*After Fletcher.*) When a sound has several principal components the intensity and frequency of each are indicated.

Music.—Musical sounds are ordinarily sustained tones having a distinct pitch which normally changes in definite steps called musical intervals. Musical sounds are practically always rich in harmonics, the presence of which causes notes having the same fundamental pitch to sound differently when produced by different instruments.

The frequency range of musical sounds is much greater than with speech. Thus, the bass tuba, bass viol, piano, organ, and drums have fundamental frequencies in the order of 60 cycles or less, while many musical instruments can produce notes which have harmonics extending up to 15,000 cycles. Musical sounds are often accompanied by a noise caused by key clicks, hissing of air, etc. This noise represents energy more or less uniformly distributed over a range of frequencies at the high-frequency end of the audible spectrum. The results of an extensive investigation of the audible frequency range of musical sounds produced by typical instruments is shown in Fig. 386.[1] These were obtained from listening tests in which the frequencies at one end of the spectrum were progressively cut out until trained observers were able to detect a difference from the original sound. The results therefore represent

[1] See W. B. Snow, Audible Frequency Ranges of Music, Speech, and Noise, *Bell System Tech. Jour.*, vol. 10, p. 616, October, 1931.

the useful frequency range but do not necessarily mean that there is absolutely no energy outside of the ranges indicated.

Power of Musical Sounds.—The power involved in musical sounds is often very large and in the case of large orchestras may reach peak values approaching 100 watts. The peak power depends of course upon the number and character of the instruments involved and upon the nature of the musical composition being rendered. Typical values of peak sound power obtained from musical instruments when played loudly are given in Table XIV.[1] The instruments producing the largest power are

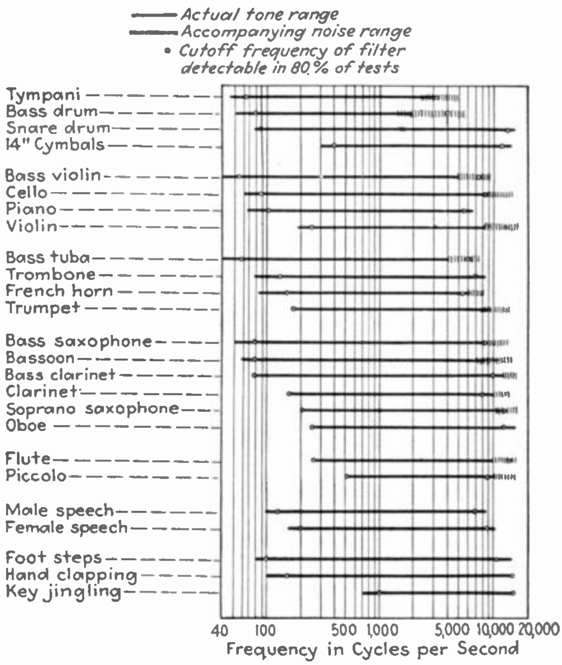


FIG. 386.—Frequency range of representative musical instruments as determined by listening tests (as given by W. B. Snow).

the bass drums, which generate peaks reaching 25 watts (approximately 15,000 times the peak power of the average voice), the pipe organ, the snare drum, the cymbals, and the trombone.

The way in which the sound power produced by musical instruments is distributed over the frequency spectrum depends upon the instrument involved and upon the fundamental frequency being played. Most, although not all, musical instruments are so constructed as to produce more sound power on the lower notes (*i.e.*, those below 500 to 1000 cycles)

[1] See L. J. Sivian, H. K. Dunn, and S. D. White, Absolute Amplitudes and Spectra of Certain Musical Instruments and Orchestras, *Jour. Acous. Soc. Amer.*, vol. 2, p. 330, January, 1931.

TABLE XIV

Instrument	Peak power, watts	Per cent of ⅛-sec. intervals in which power is at least one-fourth of peak	Band containing maximum peaks, C.P.S.
36 by 15-in. bass drum—A	24.6	6	250 to 500
36 by 15-in. bass drum—B	1.2	1½	20 to 62.5; 250 to 500
30 by 12-in. bass drum—C	13.4	1	125 to 250
34 by 19-in. bass drum—D	4.9	3	20 to 62.5
Snare drum	11.9	2½	250 to 500
15-inch cymbals	9.5	7½	8,000 to 11,300
Triangle	0.050	1	5,600 to 8,000
	0.012	37	
Bass viol	0.156	2	62.5 to 125
			125 to 250
Bass saxophone	0.288	25	250 to 500
BB♭ tuba	0.206	17	250 to 500
Trombone	6.4	5	500 to 700; 2,000 to 2,800
Trumpet	0.314	18	250 to 500; 500 to 700
French horn	0.053	6	250 to 500
Clarinet	0.050	5½	250 to 500
Flute	0.055	1	700 to 1,000
	0.014	1½	1,400 to 2,000
	0.0035	38	
Piccolo	0.084	½	2,000 to 2,800
	0.021	10	
Piano-A First method	0.166	16	250 to 500
Second method	0.437	16	250 to 500
Third method	0.198	16	250 to 500
Average	0.267	16	250 to 500

TABLE XIV.—(*Continued*)

Instrument	Peak power, watts	Per cent of $\frac{1}{8}$-sec. intervals in which power is at least one-fourth of peak	Band containing maximum peaks, C.P.S.
Piano-B—average	0.248	16	250 to 500
15-piece orchestra	9.0	$1\frac{1}{2}$	250 to 500
Average of two methods	2.2	16	2,000 to 2,800
18-piece orchestra	2.5	8	250 to 500
Average of two methods			2,000 to 2,800
75-piece orchestra—A	8.2	6	125 to 250
			250 to 500
			2,000 to 2,800
75-piece orchestra—B	13.4	9	250 to 500
	66.5	1	8,000 to ∞
75-piece orchestra—C	13.9	$1\frac{1}{2}$	250 to 500
			2,000 to 2,800
75-piece orchestra—D	13.8	6	125 to 250
			250 to 500
			2,000 to 2,800
Pipe organ—A	3.5	$1\frac{1}{2}$	250 to 500
Pipe organ—B	12.6	36	20 to 62.5

than on the higher frequencies, as is evident in Table XIV. This table also gives the percentage of time intervals of $\frac{1}{8}$-sec. duration in which the peak power is at least one-fourth of the maximum power observed, and it is apparent that different instruments vary greatly in this respect. The range of sound power encountered in the rendition of musical pieces depends upon the selection involved and upon the number of instruments participating. In the case of a symphony orchestra the sound power during the loudest passages may reach peaks which are 10,000,000 times as great as the sound power during the softest passages.

Noise.—Noise represents sounds in which the energy is more or less uniformly distributed over a considerable frequency range without a definite pitch being present. In a crude way noise can be considered as consisting of a mixture of sound of all frequencies. Noises differ in the way in which their energy is distributed with frequency, as is apparent from Fig. 387. Many noises, such as the jingling of keys, clapping of hands, and footsteps, contain important components having frequencies above 8500 to 10,000 cycles and do not sound natural when these are suppressed (see Fig. 386).

156. Elements of Acoustics.[1]—The sound reaching an observer will generally differ from the sound as generated because of reflections

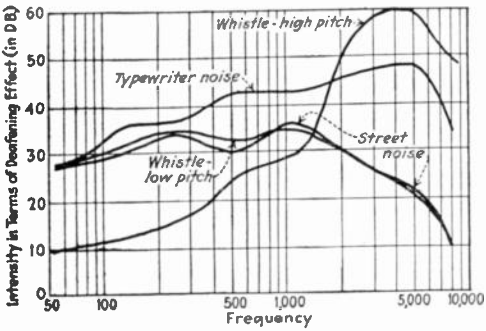


FIG. 387.—Energy distribution of typical noises expressed in terms of the deafening (or masking) effect. The distinguishing feature of noises is the more or less continuous energy distribution over a wide frequency range.

from near-by objects. Consider for example the situation illustrated in Fig. 388, which shows only a few of the paths by which sound produced in a room may travel from source to observer. The direct route involving no reflections is most important, but unless the observer is very close to the source, or unless the walls are lined with sound-absorbing material, large amounts of sound energy will reach the observer by way of the longer indirect paths involving reflections from the bounding surfaces.

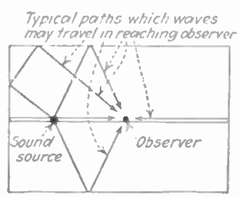


FIG. 388.—Diagram illustrating a few of the many routes which sound produced in a room may travel in reaching a listener.

The principal effects which these reflections have on the sound as observed are as follows:

1. The average intensity of the observed sound is raised because sound originally sent out in other directions is reflected back to the observer.

2. The relative amplitudes of the different frequency components of the sound may be altered as a result of selective absorption of the reflecting surfaces, which usually tend to reflect low frequencies more efficiently than high frequencies.

3. The relative amplitude of the different frequency components of the sound will always be altered as a result of interference effects resulting from the fact that the phase with which the energy traveling along the different possible paths combines, depends upon the position of the observer and upon the frequency.

[1] For further information on the subject of acoustics the reader should consult: W. C. Sabine, "Collected Papers on Acoustics," Harvard University Press, Cambridge, Mass.; F. R. Watson, "Acoustics of Buildings."

4. The observed sound persists for some time after the original sound has ceased as a result of the greater time it takes the sound traveling along the indirect routes to reach the observer. This effect is known as reverberation.

The magnitudes of the first three of these effects depend primarily upon the fraction of the total energy reaching the observing point that has traveled an indirect path, and this in turn is determined by the relative lengths of the direct and the indirect paths and the fraction of the sound-wave energy that is absorbed upon reflection. When the direct path is short, or when the bounding surfaces are of such a character as to absorb a large fraction of the energy of sound waves striking them, most of the energy reaching the observer travels along the direct path, and interference, selective absorption, etc., are not important.

When the observer is at reasonable distance from the sound source, interference effects will produce considerable quality distortion. The

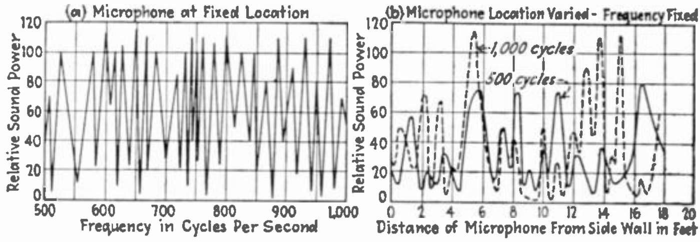


Fig. 389.—Variations in observed sound power that result from changes in position and changes in frequency because of interference effects between standing wave trains existing in small classroom.

extent of the quality distortion that can be expected is indicated in Fig. 389, which shows sound powers actually observed in a small-sized classroom.[1] It will be noted that the relative sound intensity varies widely with frequency and microphone position. The result is that the actual sounds observed depend to a large extent upon the position of the observer. Fortunately, however, the ear is not very critical with regard to the exact character of the sound waves provided that there is no systematic distortion, such as selective absorption, that discriminates more or less uniformly against a wide band of frequencies.

Reverberation Time.—Reverberation in a room depends upon the ratio of the enclosed volume to the area of the bounding surfaces, and upon the average coefficient of sound absorption of the walls. It is measured in terms of the time after the sound source has been silenced that it takes

[1] The writer is indebted to P. G. Caldwell, formerly graduate student in Electrical Engineering at Stanford University, for these curves.

a sound uniformly distributed throughout the room and having 1,000,000 times the minimum audible energy to die down to inaudibility. This time, called the reverberation time, is given by the formula:[1]

$$\text{Reverberation time in seconds} = \frac{0.05V}{-S \log_e (1 - \alpha)} \qquad (188)$$

where V is the volume of the room in cubic feet, S the total area of the bounding surfaces plus the equivalent area of sound absorbing objects such as chairs, people, etc., and α is the average coefficient of sound absorption of the surfaces.

The reverberation time of theaters, auditoriums, etc., is often of the order of several seconds because of the large volume in proportion to the wall area; while in rooms of the size encountered in ordinary homes the reverberation time is too small to be of any consequence. Upon first consideration it might appear that the smaller the reverberation time the better, but this is not always true because certain oratorical and musical effects are enhanced by reverberation. Furthermore the power of the human voice is so small that a speaker is unable to be heard by a large audience unless sound reflections with low energy absorption are depended upon to increase the average sound intensity throughout the room. There is always a particular reverberation time best suited for a particular set of conditions. The reverberation time can be controlled by means of drapes, rugs, presence or absence of an audience and other means. When the reverberation time is small the room is said to be "dead," while a large reverberation time results in a "live" condition.

Sound Absorbents.—When a sound wave strikes a wall it is reflected, but the reflected wave will have a smaller amplitude than the incident wave because of energy lost in reflection. The ratio of this energy lost to the energy of the incident wave is called the coefficient of absorption of the reflecting surface and may vary from nearly zero to unity. Typical values for a number of substances are given in Table XV.

The way in which the reflecting surface absorbs energy from the sound wave depends upon the material involved. In the case of hard solid walls composed of such material as brick a certain amount of vibrational energy is actually transferred to the wall, but as this is small the absorption coefficient is likewise low. An open window, on the other hand, allows all of the sound striking it to pass out of the enclosed space and hence has an absorption coefficient of 1.0. With porous material, such as acoustic tiles, part of the incident wave enters the pores and is then extinguished as a result of the friction from the sides of the pores. Felt, rugs, celotex, drapes, and upholstery are all effective absorbers of sound because the sound wave passes into them readily as a result of their porous nature and then suffers a high rate of attenuation as a

[1] See C. F. Eyring, Reverberation Time in Dead Rooms, *Jour. Acous. Soc. Amer.*, vol. 1, p. 217, January, 1930.

result of friction between the vibrating air particles and the solid parts of the absorbent.

TABLE XV.—ABSORPTION COEFFICIENT OF COMMON MATERIALS

Material	Coefficient of absorption
Open window	1.000
Concrete and stone	0.015
Hard plaster on wood lath	0.034
Wood varnished	0.030
Acoustic plasters	0.10 to 0.30
Acoustic tiles	0.20 to 0.70
Heavy drapes	0.50
Carpet, unlined	0.15
Carpet, lined	0.25
Individual objects	**Equivalent square feet of 100 per cent absorption**
Adult person	3.0 to 4.7
Plain wood seats	0.15
Upholstered seats, leather	0.75 to 2.0
Upholstered seats, velour	2.0 to 3.5

The acoustic properties of buildings being constructed are most satisfactorily controlled by the use of porous tiles and plasters, while drapes, rugs, and felt are used primarily for corrective purposes in structures already built, or to vary the acoustic properties from time to time. In applying acoustic treatment it is necessary to keep in mind that the relative absorption for different frequencies depends upon the nature of the material and the way in which it is used. For example when a layer of felt is backed by a hard wall the absorption coefficient is maximum for a particular frequency which depends upon the exact composition and thickness of the absorbent. Furthermore in the case of drapes, rugs, and thin porous absorbents in general, the effective absorption coefficient at a particular frequency will be a maximum when the absorbing material is placed away from the hard wall by a distance that is a quarter of a wave length for the sound frequency in question.

157. Characteristics of the Human Ear.[1]—The properties of the ear are of fundamental importance in sound work since it is through the medium of the ear that sound waves are observed. The frequency and amplitude range over which a normal ear receives auditory sensations is illustrated in Fig. 390. Frequencies below about 20 cycles are perceived by feeling rather than hearing, while frequencies above about 20,000 cycles are not heard by most ears. The limit marked "Threshold of Feeling" in Fig. 390 represents the point at which the sound intensity

[1] Much of the material in this section is summarized from Fletcher, *loc. cit.*

becomes great enough to produce a sensation of pain, while the limit marked "Threshold of Audibility" represents the minimum sound audible to the normal ear. It is apparent that the sensitivity of the ear depends upon the frequency and is maximum in the range 1000 to 3000 cycles.

The smallest variation in sound amplitude which the ear is able to perceive is roughly a constant percentage of the original intensity, which is equivalent to stating that the loudness of a sound as measured by the ear is proportional to the logarithm of the sound intensity. The minimum percentage change in sound energy that is detectable is not strictly constant but, if extremes of frequency and intensity are avoided, is in the order of 25 per cent (which represents a 12.5 per cent change in pressure).

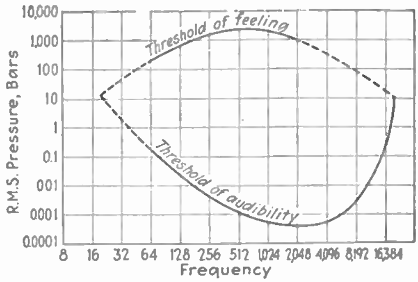


FIG. 390.—Average auditory sensation area of normal ears.

The wide range of intensities to which the ear responds and the fact that the sensitivity of the ear varies logarithmically make it convenient to use a logarithmic scale for measuring sound intensities. The unit commonly employed for this purpose is the decibel (abbreviated db), which is described in Appendix B, and is ten times the common logarithm of a power ratio. *The intensity level* of a sound wave is given by the relation:

$$\text{Intensity level in decibels} = 10 \log_{10} \left[\begin{array}{l}\text{sound power in micro-}\\ \text{watts per square centimeter}\end{array}\right] \qquad (189)$$

The intensity level is the ratio (measured in decibels) of the actual sound energy compared with an energy of 1 microwatt per square centimeter, which is approximately the average speech intensity close to the mouth. The *sensation level* represents the ratio of actual sound power to the minimum audible power when the ratio is measured in decibels, and is hence

$$\text{Sensation level in decibels} = 10 \log_{10} \frac{\text{Actual sound power}}{\text{Minimum audible power}} \qquad (190)$$

The minimum perceptible difference in sensation level is approximately 1 db.

The minimum perceptible change in frequency is roughly a constant percentage of the initial frequency. The exact ratio varies with pitch, but over the frequency range 500 to 8000 cycles is very close to 0.03 per cent.

With tones of very short duration the ear acts very much as a ballistic instrument as far as intensities are involved. Thus a sound having a given amplitude, and lasting for 1⁄10 sec., appears to have approximately the same loudness as a sound containing twice as much energy but persisting for only 1⁄20 sec.

Non-linearity of the Ear.—The ear has a non-linear response to sound waves of large amplitude. The result is that with powerful sound waves the ear produces harmonics, as well as sum and difference tones, which are not present in the original sound, and yet which are actually present in the hearing organs and are perceived by the brain. Such frequencies produced in the ear are called "subjective tones" and explain a number of sound phenomena. Thus the pitch of a sound is not changed by removing the fundamental frequency since the harmonics combine in the ear to produce a difference frequency which recreates the fundamental component in the form of a subjective tone. This non-linear character of the ear also makes many radio receivers and loud-speakers at least passably acceptable by regenerating in the form of subjective tones the low frequencies which the equipment itself fails to reproduce.

Another important consequence of the non-linear character of the ear is the phenomenon known as masking, which appears as a deafening to high-frequency sounds caused by the presence of a lower pitched sound. Masking arises from the fact that when the ear produces harmonics of the low frequencies, these harmonics interfere with the perception of the higher pitched sounds, which are then said to be masked. Masking is particularly important in noisy locations, since it is equivalent to a deafening, and is the reason that it is necessary to raise the voice when carrying on a conversation in a noisy location.

Articulation.—The extent to which a transmission system is capable of reproducing the original speech is measured in terms of the *articulation.*[1] Articulation is ordinarily based upon the accuracy with which the fundamental voice sounds are perceived, and can be tested using random combinations of vowel and consonant sounds in vowel-consonant, consonant-vowel, and consonant-vowel-consonant combinations. An articulation of 90 per cent on such a test would mean that 90 per cent of the individual voice sounds were correctly received.

The articulation is affected by the intensity of the reproduced sound and the amount of frequency distortion involved. With low sound intensities the weaker sounds, such as th (in thin), s, and f, become inaudible. Noise, in addition to its annoyance value, has an effect equivalent to reducing the sound intensity because of masking. The effect on articulation of eliminating the higher and lower frequency components of speech is particularly important because of the practical

[1] See Fletcher, *loc. cit.*, for more detailed information on articulation and articulation testing.

limitations to the frequency range of audio amplifiers and loud-speakers. The reduction in articulation that results from restricting the frequency range varies with the different speech sounds, but a general picture of the situation can be gained from Fig. 391. It will be observed from this figure that while most of the energy of speech is in the lower frequency range, the higher frequencies are essential to satisfactory articulation. Thus when all frequencies above 1550 cycles are suppressed the articulation is reduced to 65 per cent although less than 12 per cent of the sound energy is eliminated, while suppressing all frequencies below 1550 cycles gives an articulation of 65 per cent even though only about 12 per cent of the original sound energy remains.

It must be kept in mind that naturalness and articulation are not synonymous. Thus Fig. 391 shows that the articulation is not appreciably reduced by suppressing all frequencies below 500 cycles, although doing so destroys the naturalness of the voice. In general, satisfactory naturalness requires the preservation of a much greater frequency range than is needed from the standpoint of mere articulation. It is impossible to measure the degree of naturalness objectively, but listening tests indicate that speech does not sound entirely natural unless frequencies from 100 to 9000 are preserved, while a detectable distortion is observed in music unless frequencies from about 20 to 15,000 are reproduced (see Fig. 386). High-quality reproduction of orchestral music and speech requires a frequency range of at least 80 to 8000 cycles, while many noises require frequencies above 10,000 cycles in order to sound natural.

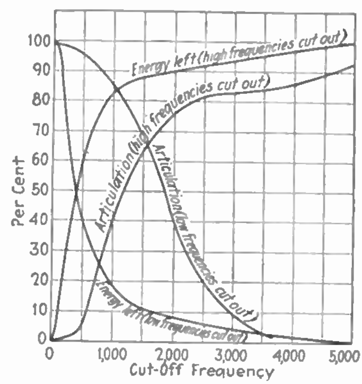


FIG. 391.—Variation of articulation for voice sounds as one end of the frequency spectrum is cut out, together with the fraction of the original sound power remaining after the restriction of the frequency band.

158. The Telephone Receiver.—The term telephone receiver is used here to denote those devices which convert electrical energy into sound waves and which are held against the ear when used. This is in contrast with loud-speakers, which are arranged to produce sound waves that spread over a large volume.

All types of telephone receivers make use of a diaphragm which is effectively sealed to the ear by means of a vented cap, so that, as the diaphragm vibrates, the pressure of the small quantity of air trapped between the diaphragm and the ear drum varies in accordance with the displacement of the diaphragm. Investigations have shown that in order to obtain distortionless reproduction the amplitude of the diaphragm vibrations should be proportional to the current supplied to the telephone receiver and independent of the frequency of this current.

Magnetic-diaphragm Telephone Receiver.—The type of telephone receiver most widely employed makes use of a permanent magnet, upon the pole tips of which are coils that carry the voice currents. The magnetic circuit is closed by means of a magnetic diaphragm, which is set in vibration by the audio-frequency currents passed through the receiver windings. The cross section of a typical telephone receiver of the type worn with a head band is shown in Fig. 392. The receivers used in the telephone system are similar except for the fact that the permanent magnet is located inside the hand piece.

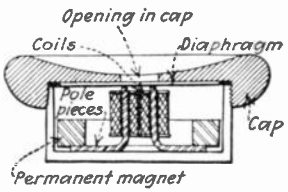


FIG. 392.—Cross section of typical watch-case type of telephone receiver.

The permanent magnet increases the sensitivity and prevents distortion. This results from the fact that the pull on each unit area of the diaphragm is proportional to the square of the flux density in the air gap. Thus if B_o is the flux density produced by the permanent magnet, and $B_s \sin \omega t$ the flux density produced by the current in the receiver windings, then

$$\text{Pull on diaphragm} \propto (B_o + B_s \sin \omega t)^2 = B_o^2 + 2B_oB_s \sin \omega t + \frac{B_s^2}{2}(1 - \cos 2\omega t) \qquad (191)$$

The first term in the right-hand side of Eq. (191) represents the constant pull produced by the permanent magnet. The second term is a force which varies in accordance with the current passing through the receiver winding and is proportional to the strength of the permanent magnet. This is the force which produces the desired diaphragm vibrations. The final term consists of a constant pull and a double frequency distortion alternating-current force, both of which are relatively small if the flux from the permanent magnet is large, but which in the absence of a permanent magnet represent the only forces exerted on the diaphragm.

The vibration of a diaphragm under the action of a mechanical force is determined by the equivalent mass, elasticity, and frictional resistance of the moving system, and has a velocity and amplitude given by the equations

$$\text{Velocity in centimeters per second} = \frac{\text{force in dynes}}{r + j\left(\omega m - \dfrac{s}{\omega}\right)} \qquad (192)$$

$$\text{Amplitude of vibration in cms} = \frac{\text{velocity}}{\omega} = -\frac{\text{force in dynes}}{\omega\left[r + j\left(\omega m - \dfrac{s}{\omega}\right)\right]} \qquad (193)$$

where

r = coefficient of friction in dynes per centimeter per second
m = equivalent mass of diaphragm in grams
s = equivalent elastic stiffness in dynes per centimeter
ω = 2π times frequency.

It will be observed that these equations are identical in form with equations giving the relationship between current and voltage in a tuned electrical circuit, and that the mechanical quantities of mass, stiffness, friction, force, velocity, and amplitude correspond in the equivalent electrical circuit to inductance, elastance, resistance, potential, current, and charge, respectively. Examination of Eqs. (192) and (193) shows that when the driving force is constant the velocity of vibration will be independent of frequency when the friction is large compared with the mass and elastance, while in order that the amplitude of vibration may be constant with change in frequency the elastance force must dominate (*i.e.*, the friction must not be too high, and the resonant frequency of the diaphragm must be greater than the frequency of the force). If the friction is small the amplitude and velocity of vibration will be very great at the resonant frequency of the diaphragm.

The vibrating diaphragm of a telephone receiver varies the reluctance of the magnetic circuit of the permanent magnet, and the resulting flux changes induce a back voltage in the receiver windings. This back voltage is proportional to the current in the receiver coil and so acts as a series impedance termed the "motional impedance," which can be determined by taking the difference between receiver impedance as measured with the diaphragm first blocked, and then free to vibrate. Figure 393 shows the results of such impedance measurements in a typical case. It is possible to deduce most of the essential characteristics of a telephone receiver from a knowledge of the motional impedance at different frequencies.[1]

The ordinary receiver employed in radio work has a resonant frequency in the order of 1000 cycles and possesses relatively little mechanical resistance. Such a receiver shows decided resonance effects, because the response (*i.e.*, the amplitude of vibration) at low frequencies, where the elastance reactance dominates, is relatively independent of frequency, but increases greatly as the resonant frequency is approached and drops rapidly thereafter. The uniformity of response can be increased at the sacrifice of sensitivity by using a stretched air-damped diaphragm having a resonant frequency in the order of 5000 cycles, and this is done where high quality is absolutely necessary.

[1] An exhaustive treatment of the theory of telephone receivers, with particular emphasis on the properties of motion impedance, is given by A. E. Kennelly in "Electrical Vibration Instruments," The Macmillan Company, New York.

Miniature telephone receivers called "phonettes" are commonly used in aircraft radio work and in hearing sets for the deaf. These are very small telephone receivers fitted into the outer ear by means of individually molded ear pieces. Such receivers keep out disturbing noises and are also relatively inconspicuous.

Balanced Armature Receivers.—The possibilities of the magnetic-diaphragm type of receiver that has just been described are limited by the fact that the diaphragm must be of magnetic material and must withstand the strong steady pull of the permanent magnet. These disadvantages are overcome in the Baldwin receiver, illustrated in Fig. 394, which

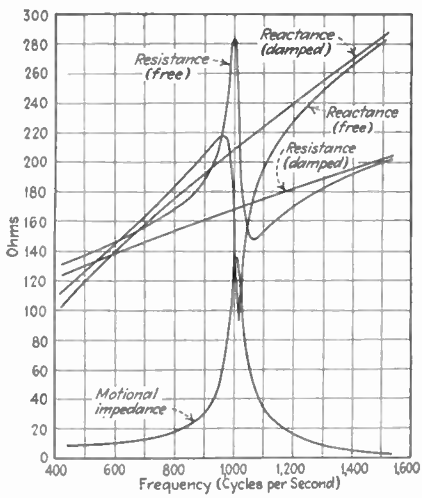


FIG. 393.—Reactance and resistance of a telephone receiver with diaphragm free to vibrate and with the diaphragm blocked, illustrating how the mechanical vibration introduces an impedance into the receiver winding that varies with frequency in much the same way as the impedance of a parallel-resonant circuit. (*After Kennelly.*)

employs a pivoted armature arranged as shown in the illustration and connected to a mica diaphragm by means of a link. The pulls exerted on the magnetic armature are balanced until current is passed through the winding, when the additional flux produced by the winding causes an unbalance that deflects the armature. The Baldwin receiver has the advantages of a low-reluctance magnetic circuit and of a diaphragm which need not be of magnetic material and which is not subjected to a steady stress. The use of a connecting link between the armature and the diaphragm also makes possible a lever action which can be used to improve the efficiency of the receiver. Baldwin receivers have a sensitivity that compares favorably with other types of receivers and are capable of handling unusually large amounts of power before non-linear distortion begins.

Moving-coil Telephone Receivers.[1]—In the moving-coil type of telephone receiver the force driving the diaphragm is obtained by the action of a magnetic field on a coil carrying the audio-frequency currents. This force is exerted at right angles to both the direction of current flow and the magnetic field, and is proportional to the product of the current in the moving coil and the strength of the magnetic field. The constructional details of a commercial design of a moving-coil telephone receiver are shown in Fig. 395. The diaphragm is of light aluminum alloy, the central part of which is formed into a spherical dome to increase its rigidity. The moving coil is composed of edgewise-wound aluminum foil, the adjacent turns of which are insulated by means of a thin layer of enamel which also acts as a binder. This type of construction gives a light self-supporting coil which has excellent heat-radiating powers and can be constructed to have very small tolerances. The magnetic field in which the coil moves is radial and is supplied by a permanent magnet. Frequency distortion

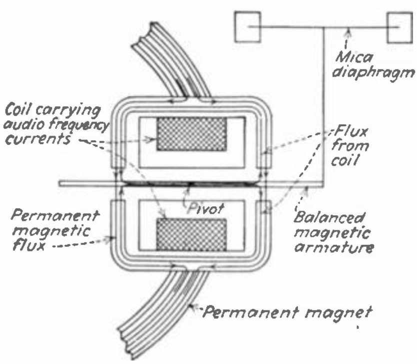


FIG. 394.—Schematic diagram illustrating the operation of a Baldwin receiver. The forces exerted on the armature are balanced unless the coil is carrying current, in which case the flux produced by the coil unbalances the flux distribution in the air gaps.

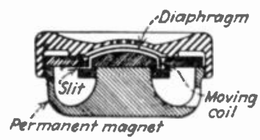


FIG. 395.—Constructional details of a commercial form of moving-coil telephone receiver.

resulting from diaphragm resonance is minimized by enclosing the back side of the diaphragm to give a small air chamber which is connected to outer space by a small slit. By the use of a suitably proportioned air chamber and slit the moving-coil receiver may be made to have a relatively constant response for frequencies up to approximately 1½ times the resonant frequency of the diaphragm.

The moving-coil type of telephone receiver can be constructed to have a substantially flat characteristic up to frequencies as high as 10,000 cycles while retaining a reasonable sensitivity. Compared with the magnetic type of receiver, the moving-coil instrument has a wider frequency range in proportion to its sensitivity, is capable of handling greater power without non-linear distortion, and is relatively permanent in its characteristics.

[1] See E. C. Wente and A. L. Thuras, Moving-coil Telephone Receivers and Microphones, *Bell System Tech. Jour.*, vol. 10, p. 565, October, 1931.

159. Horn-type Loud-speakers.[1]—A loud-speaker is fundamentally a telephone receiver of large power capacity which has been modified in such a way as to radiate sound waves into space rather than to vary the pressure in a small air chamber. This radiation is ordinarily accomplished either by a vibrating piston or by coupling the vibrating diaphragm to a large column of air by means of a horn. Loud-speakers making use of these two principles are described in this and the following section.

A horn is a device for transforming an intense vibration of a small quantity of air into a weaker vibration of a large volume of air. The horn does this by gradually opening up and allowing the sound wave to expand in an orderly manner until large enough to transfer its energy to free space without undue disturbance. The important characteristics of a horn are the size of the mouth, the character of the taper, and the size of the throat.

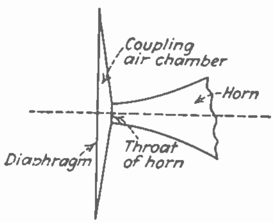


FIG. 396.—Diaphragm, throat, and coupling air chamber of horn-type loud-speaker.

Importance of Throat Area.—The throat area of the horn in relation to the diaphragm area determines the acoustic loading on the diaphragm. Referring to Fig. 396 it is apparent that as the diaphragm vibrates there is a displacement of the air in and out of the small chamber connecting the throat and diaphragm, and that the velocity of the air in the throat is greater than the diaphragm velocity by a ratio determined by the relative areas of the diaphragm and throat. Hence the smaller the throat the greater will be the pressure against which the diaphragm works, and there is no practical difficulty in making this force much greater than the inertia and elastance reactions of the vibrating system. When this is the case resonance effects are damped out, and the velocity of air produced at the throat of the speaker by a constant driving force acting on the diaphragm is practically independent of frequency, resulting in a uniform response.

Taper.—The taper of a loud-speaker horn controls the rate of expansion of the sound wave and should be such that the cross-sectional area increases at a rate proportional to the area. This calls for an exponential taper which can be expressed mathematically as follows:

$$\text{Area at distance } x \text{ from throat} = A_o\epsilon^{Bx} \tag{194}$$

where A_o is the throat area, and B is a constant determining the rate at

[1] For more detailed information on horns and horn-type loud-speakers, see C. R. Hanna and J. Slepian, The Function and Design of Horns for Loud-speakers, *Trans. A.I.E.E.*, vol. 43, p. 393, 1924; C. R. Hanna, Loud-speakers of High Efficiency and Load Capacity, *Trans. A.I.E.E.*, vol. 47, p. 607, April, 1928; E. C. Wente and A. L. Thuras, A High-efficiency Receiver for a Horn-type Loud-speaker of Large Power Capacity, *Bell System Tech. Jour.*, vol. 7, p. 140, January, 1928.

which the horn opens out and is the fundamental constant of the horn. The effectiveness with which energy propagates along an exponential horn is determined by the frequency of the wave in relation to the rate at which the horn expands. At high frequencies the propagation in the horn is almost perfect, but as the frequency is lowered a critical point is reached below which the propagation drops off rapidly. This effect depends upon the ratio $2\pi f/B$, and varies according to the equation[1]

$$\text{Relative power transmitted along horn} = \sqrt{1 - \frac{1}{4}\left(\frac{aB}{\omega}\right)^2} \qquad (195)$$

where a is the velocity of sound. This equation is plotted in Fig. 398, which shows that the radiation drops off rapidly when $\omega/B < 2.5 \times 10^4$, (*i.e.*, $f < 3980\ B$) and becomes zero at a frequency (cut-off frequency) of 2730 B.

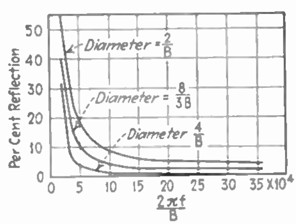


FIG. 397.—Fraction of wave reaching mouth of exponential horn that instead of being radiated is reflected back down the horn (all lengths measured in centimeters).

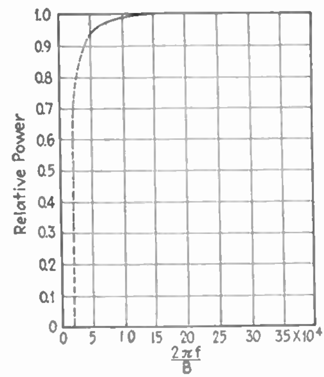


FIG. 398.—Relative power transmitted along an exponential horn with different values of taper (all lengths measured in centimeters).

Size of Mouth.—The mouth of the horn determines the lowest frequency that can be radiated from the horn into space without setting up resonances in the air column of the horn. With an exponential horn the amount of reflection that occurs at the mouth is determined by the frequency, the horn constant B, and the area of the mouth and varies with these factors as shown in Fig. 397.[1] In general the larger the mouth of the horn and the higher the frequency the less will be the reflection at the mouth. If the horn is to radiate the lowest frequency that can propagate satisfactorily along the taper, the diameter of the mouth must be not less than one-quarter wave length for this frequency, or if other than a circular opening is used, an equivalent area must be employed.

Design of Horn Speakers.—In designing a horn the first step is to determine the lowest frequency f to be reproduced. The mouth of the horn is then given an area approximating the area of a circle having a diameter one-fourth wave length at this frequency. Next the taper

[1] See Hanna and Slepian, *loc. cit.*

constant B is assigned the largest value that will permit energy of the frequency f to travel along the horn, which is approximately $B = f/4000$ when the dimensions are in centimeters. The length of the horn is determined by the mouth, the rate of taper, and the throat area. The throat area of the horn determines the acoustic loading on the diaphragm of the horn, which should be as great as possible without going to throat areas so small as to cause large frictional losses. The material used in making the horn is relatively unimportant, the only requirement being a reasonable amount of rigidity.

The air chamber used to couple the horn and diaphragm together should have a small volume, and its radius must not exceed one-fourth wave length at the highest frequency sound to be reproduced. Otherwise the sound waves resulting from the displacement of different parts of the diaphragm will not add up in phase at the throat when the frequency is high, and there will furthermore be considerable loss of output resulting from compression of the air in the chamber.

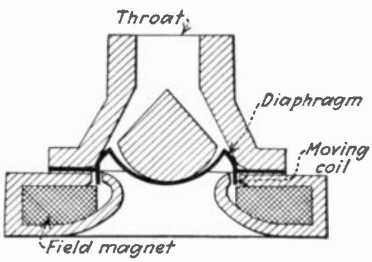


FIG. 399.—Moving-coil type of driving unit used in a commercial horn-type loud-speaker intended for sound-picture projection.

The mechanical stiffness and the mass of the diaphragm should both be as small as possible and should be so related to each other that the resonant frequency is approximately the geometric mean of the highest and lowest frequencies to be reproduced. The acoustic loading which the small throat places on a diaphragm is depended upon to suppress the resonance effects which would otherwise be present.

The most satisfactory means of driving the diaphragm of a horn-type speaker is by the use of a moving coil. The moving coil possesses a linear relation between current and mechanical force over a large range of amplitudes, is capable of handling relatively large amounts of electrical energy, and gives a distributed driving force that makes it possible to obtain piston action with large diaphragms. The ability to handle large amplitudes is particularly important at low frequencies, since Eq. (193) shows that with a resistance-damped vibration (constant velocity) the amplitude of vibration varies inversely with frequency. Horn-type loud-speakers with magnetic diaphragms, or of the balanced-armature type, were at one time used extensively but are not so satisfactory as a moving coil. An example of a moving-coil driving unit for a horn speaker is shown in Fig. 399.

A properly designed horn-type loud-speaker has a relatively constant response over a wide frequency range and an efficiency in the order of 30 to 50 per cent (*i.e.*, 30 to 50 per cent of the electrical energy supplied the moving coil is converted into sound energy). The low-frequency response is limited by the mouth of the horn, while at high frequencies the limiting factor is the size of diaphragm and air chamber. There is no difficulty in obtaining satisfactory response up to about 5000 cycles with a single horn. If the frequency range is to be extended, a second speaker having a small diaphragm permitting satisfactory response up to 10,000 or 12,000 cycles should be used.[1]

The horn-type loud-speaker is almost ideal from all points of view except size, but this feature makes it unsatisfactory for home use in radio receivers.

160. Piston-type Loud-speakers.[2]—All types of loud-speakers employing paper cones, as well as the electrostatic and the Hewlett induction speaker, are of this class.

The sound power radiated from one side of a piston placed in a baffle, as shown in Fig. 400, is:

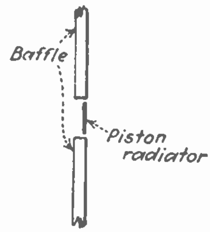


FIG. 400.—Simple piston radiator provided with a baffle to minimize interference between front- and back-side radiations.

For a large piston:

Sound power radiated from one side in ergs per second $= 20.6\ S\mu^2$ (196)

For a small piston:

Sound power radiated from one side in ergs per second $= 2.78 \times 10^{-9}\ (S\omega\mu)^2$ (197)

where S is the piston area in square centimeters, ω is 2π times frequency, and μ the maximum velocity in centimeters per second. It is assumed that all parts of the piston vibrate with the same amplitude and phase. In order to radiate the same sound power at different frequencies the velocity of the large diaphragm must be independent of frequency, while with the small piston the velocity must vary inversely with frequency. In the former case the piston must offer a frictional reaction against the driving force, while in the latter situation the resonant frequency of the piston must be very low so that an inertia reaction is offered.

The size of a piston radiator is determined by the diameter measured in wave lengths. If the diameter is one wave length or more the piston can be considered large, and Eq. (196) applies, while for diameters of one-quarter wave length or less Eq. (197) for small pistons is very nearly correct. With intermediate sizes there is a gradual transition from one law to the other. It is apparent that any particular piston will be

[1] Such a speaker is described by L. G. Bostwick, An Efficient Loud-speaker at the Higher Audible Frequencies, *Jour. Acous. Soc. Amer.*, vol. 2, p. 242, October, 1930.

[2] An excellent discussion of the fundamental principles involved in speakers of this type is given by Chester W. Rice and Edward W. Kellogg, Notes on the Development of a New Type of Hornless Loud-speaker, *Trans. A.I.E.E.*, vol. 44, p. 461, 1925.

"small" at low frequencies, when the wave length is great, and "large" at high frequencies, when the wave length is small. The high-frequency radiation also tends to be concentrated in a direction normal to the plane of the piston much more than is the low-frequency sound.

The satisfactory operation of a small piston radiator requires the use of a baffle, such as shown in Fig. 400, to prevent the radiation from the front and back sides of the vibrating diaphragm (which are 180° out of phase) from canceling each other. The baffle makes the path lengths of the front and back radiations differ by an appreciable fraction of a wave length and thus shifts the relative phases so that cancellation occurs for only a few critical frequencies, and further reduces interference effects by discriminating against the back-side radiation. The baffle should be large enough to make the shortest distance from front to back not less than one-quarter wave length at the lowest frequency to be

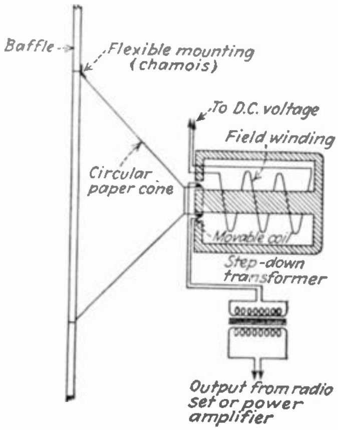


FIG. 401.—Cross section of typical dynamic type of loud-speaker.

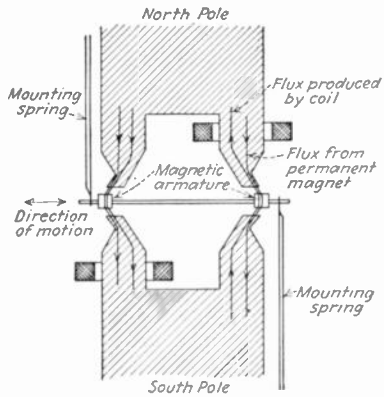


FIG. 402.—Loud-speaker driving mechanism of inductor type.

radiated. In broadcast receivers the cabinet acts as a baffle, while with large radiators, such as employed in the electrostatic and Hewlett induction speakers, the diaphragm acts as its own baffle.

Loud-speakers Employing Small Paper Cones.—The loud-speakers used in broadcast receivers are practically all of the piston type in which a small paper cone acts as the radiator. The cross section of a typical loud-speaker of this type is shown in Fig. 401. The radiator is a paper cone having a central angle of 90° and a diameter of approximately 6 in. The outer edge of the cone is supported by an annular ring of flexible material while the center is supported and guided by a spider. The resonant frequency of the moving system is in the order of 50 cycles, which means that an inertia reaction is offered to the driving force over the essential audio-frequency range. The cone is driven at the apex by a moving coil which is placed in a magnetic field produced by an electromagnet, and a step-down transformer is used to couple the low-impedance coil to the power amplifier.

Different makes of loud-speakers used in broadcast receivers often differ from Fig. 401 in a number of respects. Thus the size, material, and central angle of the paper cone may vary, while the outer edge is sometimes fixed instead of free. An electrical network is also often employed to compensate for variations of response with frequency. The cone is sometimes driven by a balanced-armature arrangement similar to that employed in the Baldwin receiver, and in one make is operated by the driver shown in Fig. 402. Speakers are commonly classified as "dynamic," "magnetic," or "inductor," according to whether a moving coil, magnetic armature, or the arrangement of Fig. 402 is used. The moving coil is generally considered the most satisfactory because it permits the largest amplitude of vibration without non-linear distortion. This is important because with a constant piston velocity (inertia reaction) the amplitude of vibration varies inversely with frequency and is very large at low frequencies.

At low frequencies a loud-speaker, such as illustrated in Fig. 401, approximates a true piston radiator, but as the frequency increases the paper cone ceases to act as a unit, and the different parts of the cone do not vibrate in the same phase. This is an advantage, for if the cone acted as a true piston at all frequencies the high-frequency radiation would be too great. The exact action taking place in the cone is very complicated. Thus investigations show that there is a complicated system of standing waves extending radially from apex to edge and also circularly in a plane perpendicular to the axis of the cone. The vibrational pattern also shows large dissymmetries and is very critical with respect to frequency even at low frequencies. The vibration of the cone also commonly contains frequency components, (particularly second harmonics) which are not present in the current flowing through the moving coil.[1]

The variation of output with frequency of a typical well-made dynamic loud-speaker is shown in Fig. 403. The characteristic shows many small irregularities, but for the most part these are too small to be noticed by the ear. The frequency range of the loud-speaker is sufficient to cover the really important part of the voice range, and while the performance falls far short of perfection the reproduced sound is considered by the average ear to be very natural.

[1] These statements are based largely upon the results obtained by J. S. Low, formerly graduate student at Stanford University, who cemented a large number of very small oscillograph mirrors over a paper cone and observed the patterns formed by reflected light when the cone was set in vibration. Also see M. J. Strutt, On the Amplitude of Driven Loud-speaker Cones, *Proc. I.R.E.*, vol. 19, p. 839, May, 1931.

When compared with a horn-type speaker, the dynamic speaker has an efficiency of only 2 to 4 per cent, approximately one-tenth of that of the horn, but occupies much less space and can be more readily built into cabinets. As a result horn speakers are used in theaters, public address systems, etc., where the problems of space and appearance are not so important, while cone-type speakers are employed in radio receivers because of their more convenient size.

Miscellaneous Types of Piston Radiators.—The double-cone speaker illustrated in Fig. 404 is capable of giving an excellent response characteristic and was widely used before loud-speakers were built into the cabinet with the radio receiver. At low frequencies the front cone acts as a piston, while at the same time the volume displaced by the cone is varied by the driving force, giving a sort of spherical radiator. As the frequency is increased the vibrations in the cone tend to become more and more localized around the front apex, since the high-frequency

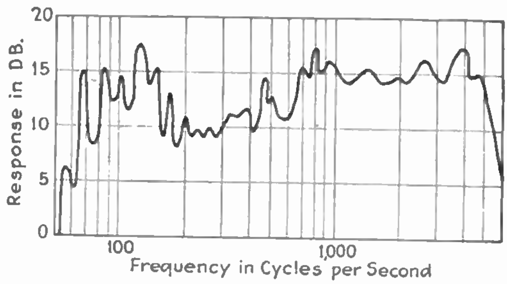


FIG. 403.—Response characteristic of typical well-made dynamic loud-speaker. While far from perfect, a characteristic such as shown is judged by the average ear as being very good.

vibrations are damped out rapidly as they travel toward the edge. The net result is a relatively uniform response over a wide frequency range. Speakers of this type are designed to offer an inertia reactance to the driving force and require no baffle because at low frequencies the action is that of a spherical radiator while at high frequencies the cone acts as its own baffle.

The Hewlett induction speaker and the electrostatic speaker are piston-type radiators which have been experimented with to a considerable extent but have not yet been successful commercially. The Hewlett speaker[1] makes use of the force which a direct-current magnetic field exerts on the audio currents induced in a large diaphragm, and while giving excellent quality of response is not practicable because of the excessive amount of energy required to produce enough direct-current magnetic flux to give reasonable sensitivity. The electrostatic loud-

[1] See C. W. Hewlett, A New Tone Generator and Sound Receiver, *Phys. Rev.*, vol. 19, p. 52, January, 1922.

speaker utilizes the attraction that exists between the plates of a condenser to set one of these plates in vibration and thus produce sound waves by piston action. The practical difficulties in speakers of this type are to obtain sufficient driving force, to obtain a reasonable amplitude of vibration without non-linear distortion, and to avoid rattling. As yet these problems have not been satisfactorily solved.[1]

161. Microphones.—Any device that converts sound energy into electrical energy is termed a microphone. While many types of microphones have been devised the only ones that are now used to an appreciable extent are the carbon, condenser, ribbon, and moving-coil types.[2]

Carbon Microphone.—The carbon microphone makes use of the fact that the resistance which a mass of carbon granules offers to an electrical current depends upon the pressure applied to the carbon. In the carbon microphone a direct current is passed through the granules, which are made from selected and treated anthracite coal and mounted against a diaphragm to form a "button." Sound waves striking the diaphragm vary the pressure exerted on the granules. This produces corresponding changes in resistance of the button and hence causes the current to vary more or less in accordance with the sound pressure exerted against the diaphragm.

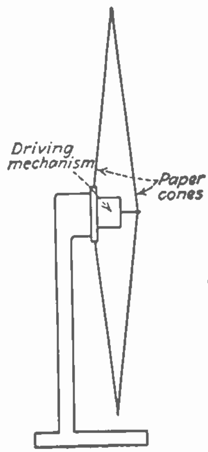


FIG. 404.—Cone type of loud-speaker widely used before speakers were built into the cabinet with the radio receiver.

The constructional features of a typical high-quality carbon-type microphone,[3] such as is sometimes employed in broadcasting stations, are shown in Fig. 405. The microphone diaphragm is made of a thin aluminum alloy which is stretched until its resonant frequency is approximately 5700 cycles. A plate is placed 0.001 in. behind the diaphragm for the purpose of introducing frictional resistance (air damping) which tends to flatten out the characteristic near the resonant frequency. This damping plate is provided with a single concentric groove, which is for the purpose of controlling the amount of damping introduced. The bridge across the front of the diaphragm is carefully proportioned with rounded corners in order to prevent the formation of resonant air pockets.

[1] Probably the most successful of the electrostatic loud-speakers is that invented by Kyle, which is described by V. Ford Greaves, F. W. Kranz, and W. D. Crozier, The Kyle Condenser Loud-speaker, *Proc. I.R.E.*, vol. 17, p. 1142, July, 1929.

[2] A discussion of different types of microphones is given by H. A. Frederick, The Development of the Microphone, *Bell Telephone Quart.*, vol. 10, p. 164, July, 1931.

[3] For a more detailed discussion of the carbon microphone see W. C. Jones, Condenser and Carbon Microphones—Their Construction and Use, *Bell System Tech. Jour.*, vol. 10, p. 46, January, 1931.

The microphone of Fig. 405 employs a button on each side of the diaphragm, with the outputs of the separate buttons combined by means of a transformer provided with a center-tapped primary. This arrangement balances out the even harmonics generated by non-linearity in the carbon buttons and practically eliminates amplitude distortion when the buttons are balanced. The resistance of the buttons is in the order of 50 to 200 ohms and is matched to the output circuit by a transformer with a suitable turn ratio. A double-button carbon microphone gives an

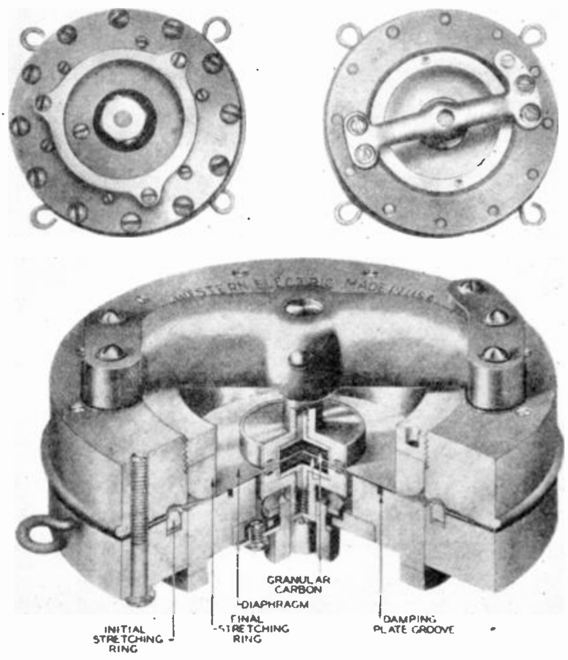


FIG. 405.—Constructional features of high-grade double-button carbon microphone.

electrical output very nearly proportional to the sound pressure up to the resonant frequency of the diaphragm, as is apparent from the typical calibration curve shown in Fig. 406. The sensitivity of a carbon microphone is very high, because the amount of electrical energy which is controlled by the pressure of the sound wave on the diaphragm is considerably greater than the energy of the sound.

The chief disadvantages of the carbon microphone are the tendency "to pack," the background noise that is always present, and general instability. Packing is a cohering of the carbon granules most commonly caused by suddenly breaking the direct-current circuit. Severe packing is accompanied by a lowering in resistance and sensitivity, which persists until the microphone is tapped or agitated mechanically. The most satisfactory way of minimizing the tendency to pack is to insert a simple filter in the microphone circuit as shown in Fig. 407. This protects the microphone button when the direct-current circuit is opened, and has a negligible effect on the voice-frequency currents. The output of a carbon microphone always contains a continuous hissing or frying sound which is caused by minute variations in the resistance of the carbon granules. The sensitivity of the carbon microphone depends markedly upon the condition of the granules in the microphone buttons and is affected by vibration, microphone position, etc.

The carbon microphone finds its chief usefulness where its high sensitivity is of value, and where the requirements as to background noise, permanency of calibration, and frequency range are not unduly severe. The double-button microphone was at one time widely used in broadcast work but has been displaced by condenser and other types of microphones having less background noise.

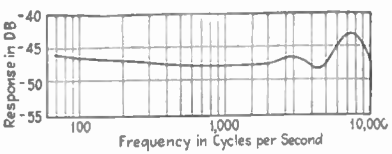


FIG. 406.—Pressure calibration curve of high-grade double-button carbon microphone.

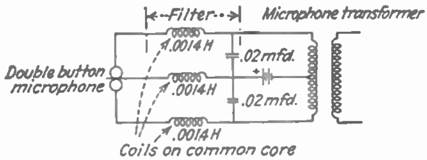


FIG. 407.—Circuit of double-button microphone showing a filter to protect the buttons from surges arising when the direct-current is opened.

The telephone transmitter used in the ordinary house telephone is a single-button type of carbon microphone in which a highly damped aluminum diaphragm having a resonant frequency in the middle of the voice range is employed. Such a microphone is considerably more sensitive than the double-button microphone of Fig. 405, because of the lower resonant frequency, but does not give as uniform a response.

Condenser Microphone.[1]—The condenser microphone is a condenser in which one plate is fixed, while the other is a diaphragm against which the sound waves act. A direct-current potential of several hundred volts is applied between the plates of the condenser, and as the capacity is varied by the vibrations which the sound waves produce in the flexible plate, a corresponding voltage drop is produced in the high resistance that is placed between the direct-current voltage and the microphone, as shown in Fig. 409.

The most important constructional details of a typical condenser microphone are shown in Fig. 408. The diaphragm is of aluminum alloy 0.0011 in. in thickness and stretched until its resonant frequency is in the order of 5000 cycles. Acoustic damping is provided by the back

[1] For additional information on condenser microphones see W. C. Jones, *loc. cit.*

plate and is controlled by a series of grooves which intersect each other at right angles, with holes drilled through the back plate at the intersections. The spacing between the diaphragm and back plate must be as small as possible in order that movements of the diaphragm will change the capacity appreciably. In the microphone shown the spacing is 0.001 in., and since there is a high direct-current potential difference between the plates it is necessary to keep out all dust and dirt. This is done by sealing the microphone from the outside air and filling it with nitrogen in order to prevent corrosion. A compensating diaphragm of

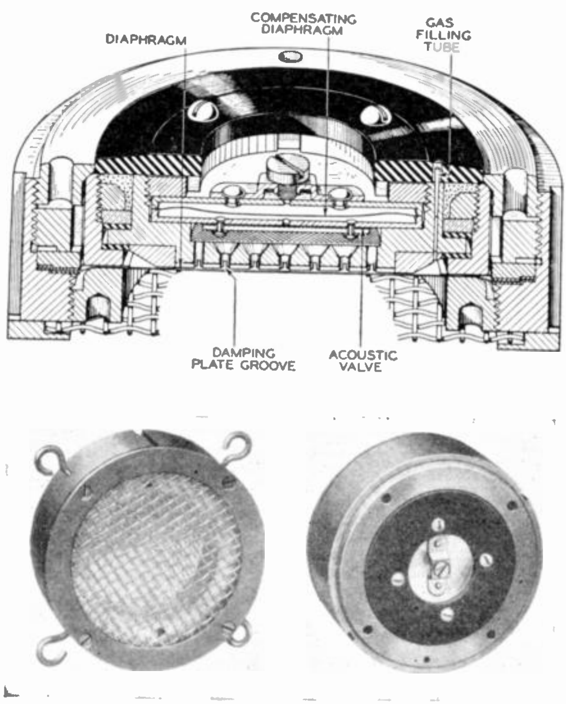


FIG. 408.—Constructional features of a typical condenser microphone.

organic material and possessing low stiffness is employed as part of this seal in order to equalize the pressures. Resonances in the air spaces of the microphone are avoided by connecting the space behind the back plate to the remaining air spaces through an acoustic valve provided by a disk of silk clamped between two aluminum rings.

The condenser microphone is used in the circuit shown in Fig. 409, which should be so proportioned that the capacity shunted across the microphone by the leads and amplifier tube is small in comparison with the microphone capacity, while the equivalent resistance formed by R and R_{gl} in parallel is at least as great as the reactance which the capacity formed by microphone, its leads, and the amplifier tube has at the lowest frequency to be reproduced. A low shunting capacity increases the sensitivity, since, as the diaphragm vibrates and changes the capacity of the microphone, the resulting potential variations are proportional to the change in capacity divided by the total capacity provided the resistances R and R_{gl} are large enough to prevent appreciable change in the charge on the microphone plates. If these resistances are not large enough there will be enough charge flowing in and out of the condenser at low frequencies to reduce the potential variations appreciably. The amplifier tube into which the microphone output feeds should be placed as close as possible to the microphone in order to shorten the leads and also to reduce the stray pick-up. This latter point is particularly important because the condenser microphone has a relatively small output and is a high-impedance device. The usual practice is to place the first amplifier tube in the microphone stand. This tube should also be non-microphonic in order to avoid the introduction of extraneous noise through electrode vibration.

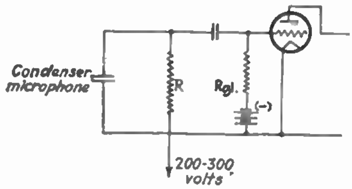


FIG. 409.—Circuit of condenser microphone.

Compared with the carbon microphone, the condenser microphone requires two additional stages of amplification to make up for its lower sensitivity, but has a much better signal-to-background-noise ratio. The latter feature makes the condenser microphone preferable in all applications where high-grade performance is desired. The condenser microphone is in turn, however, being displaced in some applications by the moving-coil and ribbon-type microphones, which, being low-impedance devices, are less susceptible to troubles from stray pick-up.

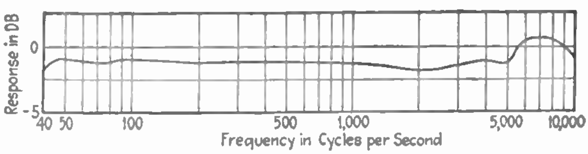


FIG. 410.—Pressure calibration curve of a high-grade moving-coil type of microphone.

Moving-coil Microphone.[1]—The moving-coil microphone is similar in construction to the moving-coil type of telephone receiver discussed in Sec. 158. The dynamic characteristics of the two instruments differ, however, because in the microphone the voltage induced in the moving coil is proportional to the velocity of the coil, which must therefore be

[1] See E. C. Wente and A. L. Thuras, Moving-coil Telephone Receivers and Microphones, *Bell System Tech. Jour.*, vol. 10, p. 565, October, 1931.

proportional to the sound pressure, while in the telephone receiver the best performance is obtained when the amplitude of vibration varies directly as the sound pressure being reproduced. The requirements of the moving-coil microphone can be met by modifying the moving-coil receiver in several respects and particularly by making the resonant frequency of the diaphragm moderately low, such as 600 cycles, and using resonant air chambers to increase the amplitude at lower frequencies and higher frequencies. When this is done a very excellent microphone results, as is illustrated by the calibration curve of Fig. 410, which is practically flat from 40 to 10,000 cycles. In addition to its very satisfactory frequency response the moving-coil microphone has somewhat greater sensitivity than a condenser microphone of the same frequency range, is not subject to stray pick-up because it is a low-impedance instrument, and does not require a source of energy for polarization because of the permanent magnet. These considerations make the moving-coil microphone more desirable than the condenser microphone in many applications.

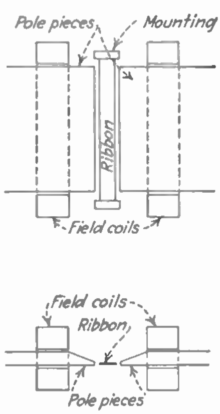


FIG. 411.—Constructional features of a ribbon microphone (shown schematically).

The Ribbon Microphone.—The ribbon microphone is a special form of moving-coil microphone in which the moving coil consists of a flat piece of aluminum-alloy foil that is acted upon directly by the sound waves, and has a resonant frequency below the audible range. The construction of such a microphone is shown in Fig. 411. The ribbon is 0.0001 in. thick, $\frac{3}{16}$ in. wide, and 2 in. long, and is placed between the pole pieces of an electromagnet. Sound waves striking the ribbon cause it to flutter between the pole pieces, inducing voltages in the ribbon which can be stepped up by a suitable transformer to match any desired load impedance. The magnitude of these voltages is proportional to the velocity of the ribbon, which must therefore offer a friction reaction to a driving force. This is accomplished by making the ribbon very light and mounting it with negligible tension. Under these conditions the air pressure on the back side of the wide ribbon gives the necessary damping. The ribbon microphone has a frequency characteristic comparable with that of the condenser microphone and has the advantage of being a low-impedance instrument.

162. Measurements, with Particular Reference to the Determination of Microphone and Loud-speaker Characteristics.—The method universally employed in making sound measurements is to calibrate a condenser microphone and then use this microphone as a measuring instrument. This procedure is possible because the condenser microphone is relatively permanent in its characteristics and has a substantially flat response with respect to frequency.

The calibration curve obtained for a condenser microphone depends upon the way in which the calibration is made because the pressure of a sound wave traveling in free space is not necessarily the pressure which this same wave produces against the diaphragm of a condenser microphone. At low frequencies, where the microphone dimensions are small compared with a wave length, the sound wave diffracts around the microphone with negligible reflection and produces a pressure against the diaphragm which is a true measure of the pressure of the wave. At higher frequencies, however, reflections occur at the microphone and may cause the pressure exerted on the diaphragm to reach twice the pressure of the wave. In addition the front of the microphone forms a shallow air pocket which introduces a resonance that will still further increase the pressure against the diaphragm at high frequencies.

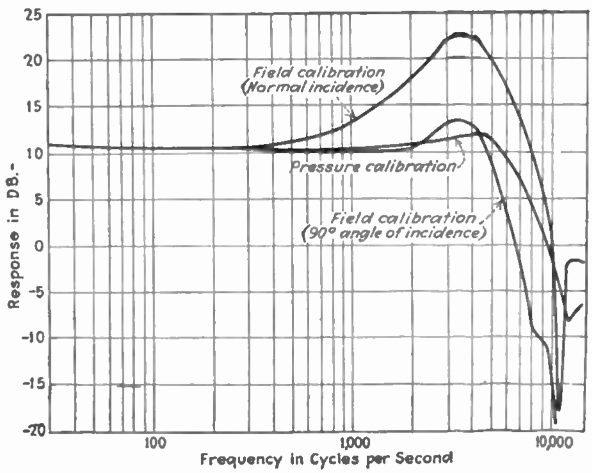


FIG. 412.—Typical pressure and field calibrations of a condenser microphone.

The relation between the output voltage of a condenser microphone and the pressure exerted against the diaphragm is called the pressure calibration, while the relation between the pressure of a free sound wave and the output voltage which this wave develops when striking the microphone is called the field calibration. The two are the same at low frequencies, while at high frequencies the field calibration depends upon the angle of incidence with which the wave strikes the microphone, as well as upon the frequency. Typical pressure- and field-calibration curves of a condenser microphone are shown in Fig. 412.

Absolute Calibration of Microphones by Rayleigh Disk, Thermophone, and Actuator.—The absolute field calibration of a microphone can be most conveniently obtained by using a Rayleigh disk to measure the intensity of the free wave.[1] The Rayleigh disk, which actually measures the velocity component of the free sound wave, consists of a very light silvered disk hung vertically on a fine suspension and arranged with an optical system so that small rotations can be accurately measured by means of an optical lever. Sound waves in striking the disk tend to turn it perpendicular to the direction of travel of the wave. With a plane wave the pressure can be calculated from the behavior of the Rayleigh disk according to the following formula:[2]

$$\text{Pressure in bars} = \left(\frac{3\pi^2 mc^2 P}{4T^2 a \sin 2\theta}\right)^{\frac{1}{2}} \Phi \qquad (198)$$

where

m = mass of disk
c = velocity of sound
P = density of air
T = period of free vibration of the disk
a = radius of disk
θ = angle between direction of sound and equilibrium position of the normal to the disk
Φ = angle in radians through which disk turns.

Absolute pressure calibrations can be made by means of the thermophone, the Rayleigh disk, or an actuator.[3] The thermophone method produces calculable sound pressures in a small enclosed space in the front of the microphone by passing an alternating current superimposed upon a direct current through a strip of gold or platinum foil which is so thin that the temperature varies at the frequency of the alternating current. In the Rayleigh disk method the microphone is placed at the closed end of a resonance tube and the Rayleigh disk used to measure the velocity at a pressure node. In the actuator method a calculable force is applied to the diaphragm of the condenser microphone electrostatically by means of a screen placed a known distance in front of the diaphragm and excited by a suitable audio-frequency voltage.

Relative Pressure Calibration.—The relative sensitivity of a condenser microphone to pressures of different frequencies can be readily obtained

[1] An excellent description of the details involved in such a calibration is given by Harry F. Olsen and Stanford Goldman, The Calibration of Microphones, *Electronics*, vol. 3, p. 106, September, 1931.

[2] Other methods sometimes employed to obtain field calibrations include the use of a very small microphone to determine the intensity of the free sound wave, and placing an ordinary microphone in a spherical or stream-line housing for which the increase in pressure due to reflection effects can be calculated. See L. J. Sivian, Absolute Calibration of Condenser Transmitters, *Bell System Tech. Jour.*, vol. 10, p. 96, January, 1931; Stuart Ballantine, Note on the Effect of Reflection by the Microphone in Sound Measurements, *Proc. I.R.E.*, vol. 16, p. 1639, December, 1928.

[3] For a more detailed discussion of pressure calibration see L. J. Sivian, *loc. cit.*

with the arrangement shown in Fig. 413. Here an audio-frequency voltage superimposed upon a direct-current potential is applied between the plate and diaphragm and used to apply force on the diaphragm. The resulting capacity variations can be determined by having the microphone capacity control the frequency of a high-frequency oscillator.

Sound Measurements with Calibrated Condenser Microphone.—After a condenser microphone has been calibrated, the intensity of sound waves can be readily determined by observing the output voltage of the microphone. Other microphones can be tested by comparison, while the frequency response of loud-speakers may be determined by supplying a known electrical input and determining the sound output by means of the microphone. In making sound measurements it is necessary to pay particular attention to the conditions under which the measurements are made, for otherwise the results will be difficult to interpret. In loud-speaker measurements, for example, it is necessary to avoid all reflections if errors from interference effects are to be avoided. The result is that loud-speaker measurements must either be made outdoors, with the speaker mounted on a stand, or, better yet, on ropes some distance from all objects (including the earth), or in a room having walls lined with sound-absorbing material. The microphone should furthermore be placed close to the speaker in order to make the ratio of direct to indirect sound waves as great as possible. In order to avoid errors from diffraction effects the microphone must not be closer than $D^2f/4500$ ft. from the loud-speaker, where D is the diameter of the radiator in feet.[1]

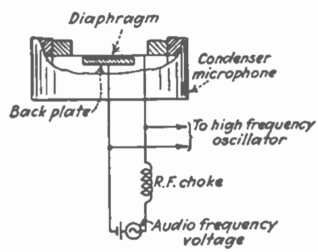


FIG. 413.—Simple method of obtaining a relative calibration of a condenser microphone. The diaphragm is driven by an audio-frequency voltage superimposed upon a direct-current potential, and the resulting capacity variations, which are proportional to the response, are determined by noting the amount of frequency modulation produced.

The total power radiated by a loud-speaker can be determined by measuring the sound radiation in different directions and integrating the resulting energy flow. A rapid method of accomplishing this integration approximately is to rotate a microphone and note the average response. Sound power can also be determined from the average intensity produced in a room having known reverberation characteristics.

[1] For detailed information on the problems involved in measuring loud-speaker characteristics see L. G. Bostwick, Acoustic Considerations Involved in Steady State Loud-speaker Measurements, *Bell System Tech. Jour.*, vol. 8, p. 135, January, 1929; I. Wolff and A. Ringel, Loud-speaker Testing Methods, *Proc. I.R.E.*, vol. 15, p. 363, May, 1927; I. Wolff, Sound Measurements and Loud-speaker Characteristics, *Proc. I.R.E.*, vol. 16, p. 1729, December, 1928; Benjamin Olney, Notes on Loud-speaker Response Measurements and Some Typical Response Curves, *Proc. I.R.E.*, vol. 19, p. 1113, July, 1931.

The response of most loud-speakers contains so many irregularities that a point-by-point calibration taken at regular frequency intervals is extremely tedious. The most satisfactory method of testing a loud-speaker is to drive it with a beat-frequency oscillator, hunt for the peaks and valleys in the characteristic, and make observations at these points irrespective of the exact frequency intervals that result. Even with such a procedure the time required to obtain a complete calibration is very great. When a large number of characteristics are to be obtained

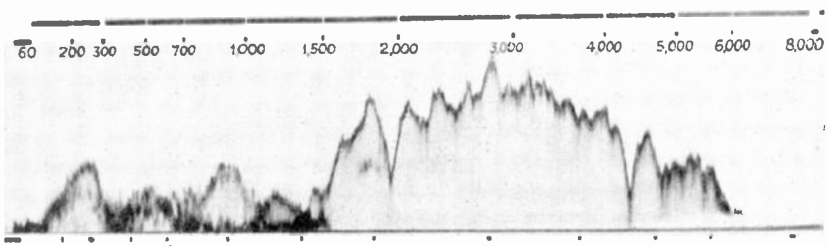


FIG. 414.—Loud-speaker characteristic as recorded on an oscillograph film.

a more satisfactory test procedure is to employ a two-element oscillograph arranged so that one element records the output level of the microphone amplifier, while the other is associated with the dial of the beat-frequency oscillator in such a way as to give indications at certain known frequencies by interconnecting the beat-frequency oscillator and oscillograph drum mechanically. It is possible in this way to take a complete calibration in a fraction of a minute. An example of such a calibration is shown in Fig. 414.[1]

[1] This technique for the rapid recording of loud-speaker characteristics was developed by C. R. Skinner, former graduate student at Stanford University.

APPENDIX A

FORMULAS FOR CALCULATING INDUCTANCE, MUTUAL INDUCTANCE, AND CAPACITY

163. Formulas for Calculating Inductance, Mutual Inductance, and Capacity.[1]—This section gives formulas for calculating inductance, mutual inductance, and capacities for the cases commonly encountered in radio work. Most of these formulas involve small approximations but will give results sufficiently accurate for all ordinary requirements.

Inductance of Single-layer Solenoid.—The equation applying in this case has already been discussed in Sec. 6 and is

$$\text{Inductance in microhenrys} = n^2dF \tag{199}$$

where n is the number of turns, d the diameter of the coil measured to the center of the wire, and F is a constant determined by the ratio of length to diameter, and given in Fig. 7. This formula can also be used to obtain the inductance of multilayer coils provided that the radial depth of the winding is small compared with the radius and length of the coil. In such cases the equivalent diameter is taken as the diameter measured to the center of the winding. Equation (199) can also be used to calculate the inductance of polygonal coils, when the number of the sides of the polygon is fairly large, by assuming that the coil is equivalent to a helix whose mean radius is the mean of the radii of the circumscribed and inscribed circles of the polygon.

Inductance of Single-layer Rectangular Coil.

$$\text{Inductance in microhenrys} = an^2[G + H] \tag{200}$$

where

a = length of long side in inches
a_1 = length of short side in inches
b = axial length of coil in inches
n = number of turns
G = factor determined by a_1/a and b/a and given at Fig. 415
H = factor determined by number of turns and $\dfrac{\text{diameter of wire}}{\text{length of coil}}$ and given at Fig. 415.

[1] Most of the formulas given here are taken from *Bur. Standards* Circular 74, Radio Instruments and Measurements. This book is the standard authority on the subject and contains formulas for making calculations of any desired accuracy for almost every case that can be encountered in practice.

This formula finds its chief application in calculating the inductance of loop antennas.

Inductance of Multilayer Coils with Winding Having a Rectangular Cross Section.

$$\text{Inductance in microhenrys} = \frac{l^{5/3}}{D^{2/3}} I \tag{201}$$

or

$$\text{Inductance in microhenrys} = an^2J \tag{202}$$

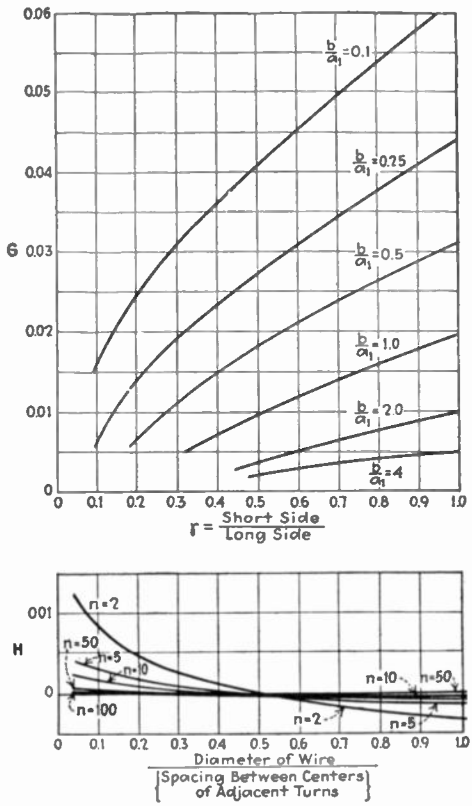


Fig. 415.—Factors for use in Eq. (200).

where l is the length of wire in inches, D the distance between centers of adjacent wires, a the mean coil radius in inches, n the number of turns, and I and J factors given by Fig. 416. It will be observed that the maximum inductance is obtained from a given length wire when the cross section of the winding is square, and the side of the cross section is 0.662 times the mean coil diameter.

Inductance of a Flat Spiral.—A flat spiral is a special case of a multilayer coil having a winding of rectangular cross section and so can be handled by Eq. (202). In the case of a pancake one turn wide and wound of small wire, b/a can be taken as zero. With a fixed width of coil, the maximum inductance with a given length of wire is obtained when the radial depth of the coil is three-fourths of the mean radius.

Inductance of Multilayer Rectangular Coils Having Windings of Rectangular Cross Section.

$$\text{Inductance in microhenrys} = an^2G \qquad (203)$$

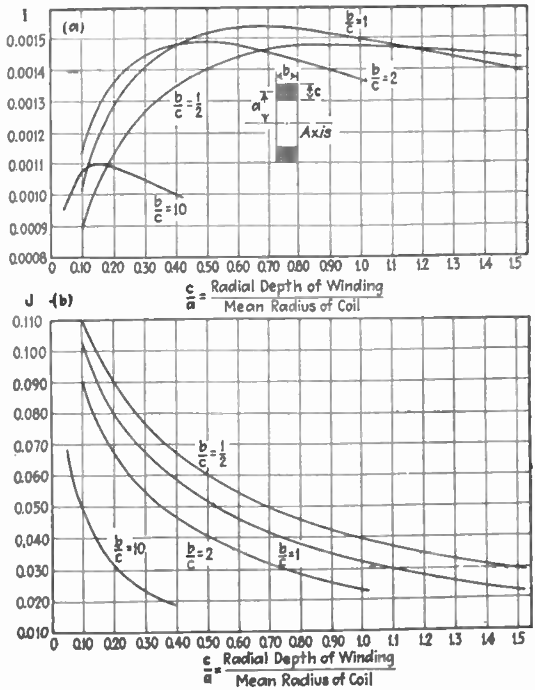


FIG. 416.—Factors for use in Eqs. (201) and (202).

where a is the length in inches of the longer side of the rectangle measured between centers of the rectangular winding cross section, n is the number of turns, and G is given by Fig. 415, where a/a_1 is taken as the ratio of short side to long side of the rectangle, and b/a is considered to be

$$\frac{\text{circumference of winding cross section}}{2a}.$$

Inductance of Toroidal Coils with Single-layer Windings.—When the toroid is a circular ring of circular cross section (doughnut shape)

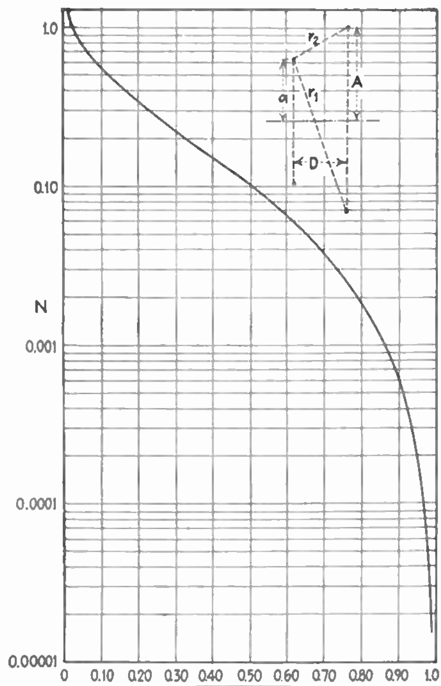
$$\text{Inductance in microhenrys} = 0.0319n^2R\left[1 - \sqrt{1 - \left(\frac{a}{R}\right)^2}\right] \qquad (204)$$

where R is the distance in inches from the axis to center of cross section of winding, a is the radius of the turns of the winding, and n the number of turns.

When the toroid is a ring of rectangular cross section

$$\text{Inductance in microhenrys} = 0.0117n^2h \log_{10} \frac{r_2}{r_1} \tag{205}$$

where n is the number of turns, h the axial length of the ring in inches, and r_1 and r_2 the inner and outer radii of the ring.



$$\frac{r_2}{r_1} = \sqrt{\frac{\left(1-\frac{a}{A}\right)^2+\frac{D^2}{A^2}}{\left(1+\frac{a}{A}\right)^2+\frac{D^2}{A^2}}}$$

FIG. 417.—Factor for use in Eq. (208).

Inductance of Parallel-wire Transmission Line.

$$\text{Inductance in microhenrys per foot} = 0.281 \log_{10} \frac{b}{a} + 0.030 \tag{206}$$

where b is the spacing between centers, and a the radius of the wire.

Inductance of Cable with Sheath Serving as Return Conductor.

$$\text{Inductance in microhenrys per foot} = 0.140 \log_{10} \frac{b}{a} + 0.015 \tag{207}$$

where b is the radius of the inner side of the sheath, and a the radius of the central conductor. This formula neglects the small contribution to the inductance made by the flux within the sheath metal.

Mutual Inductance between Two Parallel Coaxial Circles.—Using the notation shown in Fig. 417, with the dimensions in inches:

$$\text{Mutual inductance in microhenrys} = N\sqrt{Aa} \tag{208}$$

where N depends upon r_2/r_1 and is given by Fig. 417.

Mutual Inductance between Two Coaxial Circular Coils of Rectangular Cross Section.—If the windings are of square or approximately square cross section, then

$$\text{Mutual inductance in microhenrys} = n_1 n_2 M_0 \tag{209}$$

where n_1 and n_2 are the number of turns in the two coils, and M_0 is the mutual inductance as calculated by Eq. (208) for two coaxial circles which are located at the centers of the cross sections of the two coils. Equation (209) will hold with good accuracy even when the cross section departs widely from a square provided the coils are not close together.

Mutual Inductance between Coaxial Solenoids.—There are three cases to distinguish, as illustrated in Fig. 418.

Coaxial solenoids not concentric:

$$M = 0.02505\frac{a^2A^2n_1n_2}{2x \times 2l}[K_1k_1 + K_3k_3 + K_5k_5] \tag{210}$$

where

a = the smaller radius, measured from the axis of the coil to the center of the wire, in inches
A = the larger radius, measured in the same way, in inches
$2l$ = length of the coil of smaller radius = number of turns times the pitch of winding, in inches
$2x$ = length of the coil of larger radius, measured in the same way, in inches
n_1 and n_2 = total number of turns on the two coils
D = axial distance between centers of the coils in inches.

$$x_1 = D - x \qquad r_1 = \sqrt{x_1^2 + A^2}$$

$$x_2 = D + x \qquad r_2 = \sqrt{x_2^2 + A^2}$$

$$K_1 = \frac{2}{A^2}\left(\frac{x_2}{r_2} - \frac{x_1}{r_1}\right), \quad k_1 = 2l$$

$$K_3 = \frac{1}{2}\left(\frac{x_1}{r_1^5} - \frac{x_2}{r_2^5}\right), \quad k_3 = a^2l\left(3 - 4\frac{l^2}{a^2}\right)$$

$$K_5 = -\frac{A^2}{8}\left[\frac{x_1}{r_1^9}\left(3 - 4\frac{x_1^2}{A^2}\right) - \frac{x_2}{r_2^9}\left(3 - 4\frac{x_2^2}{A^2}\right)\right]$$

$$k_5 = a^4l\left(\frac{5}{2} - 10\frac{l^2}{a^2} + 4\frac{l^4}{a^4}\right)$$

Concentric Solenoids (Figs. 418*b* and 418*c*).—The formulas for these two cases are the same provided g is defined for each case as shown in the figure.

$$\text{Mutual inductance in microhenrys} = 0.0501\frac{a^2 n_1 n_2}{g}\left[1 + \frac{A^2 a^2}{8g^4}\left(3 - 4\frac{x^2}{a^2}\right)\right] \tag{211}$$

Capacity of Parallel-plate Condenser.

$$\text{Capacity in micromicrofarads} = 0.2244K\frac{S}{t} \tag{212}$$

where

K = dielectric constant
S = area of dielectric
t = thickness of dielectric.

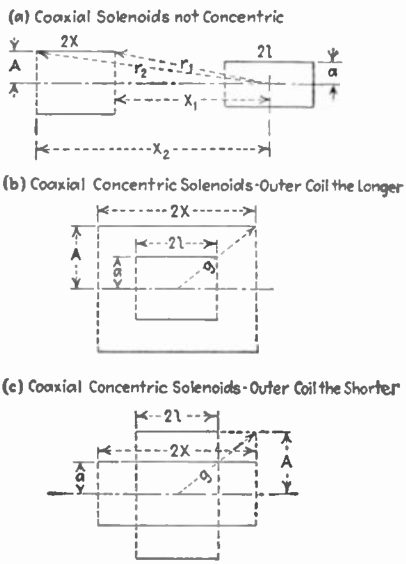


FIG. 418.—Possible arrangements of coaxial solenoids.

Capacity of Two-wire Transmission Line.

$$\text{Capacity in micromicrofarads per foot} = \frac{3.680}{\log_{10}(b/a)} \tag{213}$$

where b is the spacing between wire centers, and a is the radius of the wire.

Capacity of Round Wire in a Concentric Sheath.

$$\text{Capacity in micromicrofarads per foot} = \frac{7.354K}{\log_{10}(b/a)} \tag{214}$$

where K is the dielectric constant of the insulation ($K = 1$ for air), b is the radius of the inner side of the sheath, and a is the radius of the central conductor.

APPENDIX B

THE DECIBEL

The decibel (abbreviated db) is a logarithmic unit used in communication work to measure power ratios. If the powers being compared are P_1 and P_2, then the ratio P_1/P_2 as expressed in decibels is given by the equation

$$\text{Power ratio in decibels} = 10 \log_{10}\left(\frac{P_1}{P_2}\right) \tag{215}$$

The power ratio corresponding to different decibel values is given in Table XVI.

TABLE XVI.—POWER RATIOS AS EXPRESSED IN DECIBELS

Number of decibels	Corresponding power ratio	Number of decibels	Corresponding power ratio	Number of decibels	Corresponding power ratio
0.1	1.023	1.0	1.259	10	10
0.2	1.047	2.0	1.585	20	100
0.3	1.072	3.0	1.995	30	1,000
0.4	1.096	4.0	2.51	40	10,000
0.5	1.122	5.0	3.16	50	100,000
0.6	1.148	6.0	3.98	60	1,000,000
0.7	1.175	7.0	5.01	70	10×10^6
0.8	1.202	8.0	6.31	80	100×10^6
0.9	1.230	9.0	7.94	90	$1,000 \times 10^6$
1.0	1.259	10.0	10.00	100	$10,000 \times 10^6$

NOTE:—Interpolation can be obtained by breaking the decibel equivalent into a sum and multiplying the corresponding power ratios. Thus 25.3 db = 20 + 5 + 0.3, which represents a power ratio of $100 \times 3.16 \times 1.072 = 339$.

The practical value of the decibel arises from its logarithmic nature. Thus the enormous ranges of power involved in communication work can be expressed in terms of decibels without running into inconveniently large numbers, while at the same time small ratios can still be readily expressed. Thus 1 db represents a power ratio of approximately 5 to 4, while 60 db represents a ratio of 1,000,000 to 1. The logarithmic character of the decibel also makes it possible to express the ratio of input-to-output powers of a complicated circuit as the sum of the decibel equivalents of the ratios of input-to-output powers of the different parts of the circuit that are in series. The decibel is also the natural unit for expressing sound intensities since the effect which sound waves have on the ear is roughly proportional to the logarithm of the intensity.

When the two powers P_1 and P_2 being compared are dissipated in equal resistances the power ratio is proportional to the square of the voltage ratio or to the square of the current ratio. Under these conditions it is possible to express a voltage or current ratio in decibels by using the relation

$$db = 20 \log \left(\frac{E_1}{E_2}\right) = 20 \log \left(\frac{I_1}{I_2}\right) \tag{216}$$

This relation must be used with caution, however, since it holds only when the resistance associated with E_1 (and I_1) is the same as the resistance associated with E_2 (and I_2).

INDEX

A

B