# 1962 IRE International Convention Record

## PART 2

Sessions Sponsored by

IRE Professional Groups on

Automatic Control

Circuit Theory

at

the IRE International Convention, New York, N.Y.

March 26-29, 1962

The Institute of Radio Engineers

## 1962 IRE INTERNATIONAL CONVENTION RECORD

An annual publication devoted to papers presented at the IRE International Convention held in March of each year in New York City. Formerly published under the titles CONVENTION RECORD OF THE I.R.E. (1953 & 1954), IRE CONVENTION RECORD (1955 & 1956), and IRE NATIONAL CONVENTION RECORD (1957, 1958, & 1959).

Additional copies of the 1962 IRE INTERNATIONAL CONVENTION RECORD may be purchased from the Institute of Radio Engineers, 1 East 79 Street, New York 21, N.Y., at the prices listed below.

| Part | Sessions | Subject and Sponsoring IRE Professional Group | Prices for Members of Sponsoring Professional Group (PG), IRE Members (M), Libraries and Sub. Agencies (L), and Nonmembers (NM) | | | |
|---|---|---|---|---|---|---|
| | | | PG | M | L | NM |
| 1 | 8, 16, 23 | Antennas & Propagation | $ .70 | $ 1.05 | $ 2.80 | $ 3.50 |
| 2 | 10, 18, 26, 41, 48 | Automatic Control<br>Circuit Theory | 1.00 | 1.50 | 4.00 | 5.00 |
| 3 | 1, 9, 17, 25, 28, 33 | Electron Devices<br>Microwave Theory & Techniques | 1.00 | 1.50 | 4.00 | 5.00 |
| 4 | 4, 12, 20, 34, 49 | Electronic Computers<br>Information Theory | 1.00 | 1.50 | 4.00 | 5.00 |
| 5 | 5, 13, 15, 22, 29, 47, 54 | Aerospace & Navigational Electronics<br>Military Electronics<br>Radio Frequency Interference<br>Space Electronics & Telemetry | 1.20 | 1.80 | 4.80 | 6.00 |
| 6 | 3, 11, 31, 35, 42, 45, 50, 52 | Component Parts<br>Industrial Electronics<br>Product Engineering & Production<br>Reliability & Quality Control<br>Ultrasonics Engineering | 1.40 | 2.10 | 5.60 | 7.00 |
| 7 | 30, 37, 43, 51 | Audio<br>Broadcasting<br>Broadcast & Television Receivers | .80 | 1.20 | 3.20 | 4.00 |
| 8 | 7, 24, 38, 46, 53 | Communications Systems<br>Vehicular Communications | 1.00 | 1.50 | 4.00 | 5.00 |
| 9 | 2, 19, 27, 32, 39, 40, 44 | Bio-Medical Electronics<br>Human Factors in Electronics<br>Instrumentation<br>Nuclear Science | 1.20 | 1.80 | 4.80 | 6.00 |
| 10 | 6, 14, 21, 36 | Education<br>Engineering Management<br>Engineering Writing & Speech | .80 | 1.20 | 3.20 | 4.00 |
| | | Complete Set (10 Parts) | $10.10 | $15.15 | $40.40 | $50.50 |

Responsibility for the contents of papers published in the IRE INTERNATIONAL CONVENTION RECORD rests solely upon the authors and not upon the IRE or its members.

# 1962 IRE INTERNATIONAL CONVENTION RECORD

## PART 2—AUTOMATIC CONTROL; CIRCUIT THEORY

## TABLE OF CONTENTS

Page

### Circuit Theory - I
(Session 10: sponsored by PGCT)

### Circuit Theory - II
(Session 18: sponsored by PGCT)

### Control Systems - Theory
(Session 26: sponsored by PGAC)

Page

Control Systems - Applications
(Session 41: sponsored by PGAC)

Symposium: The Design of Networks with a Digital Computer
(Session 48: sponsored by PGCT)

# SUBNETWORKS

P. S. Castro and W. W. Happ
Microsystems Electronics Department
Lockheed Missiles and Space Company
Sunnyvale, California

## Summary

The logic governing the generation of subnetworks from multiterminal networks can be established by associating with each permissible network operation a suitably defined set of reductions in rank of the indefinite (or equi-cofactor) matrix of the network. Criteria for uniqueness and non-redundancy of subnetworks are defined and applied to evaluate representative large networks in terms of properties associated with generated subnetworks.

## Definition and Terminology for Subnetworks

The methods of network synthesis can be used to prescribe the internal interconnections (or environment) of a "black box" to achieve a given function. An alternative, and sometimes equivalent, method is to begin with a number of circuit elements and prescribe the interconnections between the elements to achieve a given circuit function (or at least as close to it as one can with the given elements). In contrast to network synthesis, however, this method commences with a given "black box" and prescribes the external environment of the "black box". The term "subnetworks" will be used to define a network derived from a given "black box" by specifying a given external environment. Of course, it is then possible to specify an additional environment of the subnetwork and thus generate further subnetworks. It is important to examine the properties of subnetworks in relation to those of the parent network. This investigation is concerned primarily with establishing the number of non-redundant subnetworks with a given classification and the matrix operations required to lead from the parent network to the subnetwork. Since the matrix gives all the network properties and the mechanics of matrix operations is well understood, generation of the desired network function by the methods outlined here is possible in principle, but difficult in practice.

## Statement of Problem

A three-terminal network permits the derivation of the following unique subnetworks: three unique two-ports and six unique one-ports, as shown in Fig. 1 and Fig. 2. To define a unique subnetwork, it will be assumed that an interchange of terminals as shown in Fig. 3 is not considered a novel network configuration, since the method of driving a network will not change its characteristics. Similarly, a one-port may be driven in two ways, as shown in Fig. 4, and is considered only as a single configuration since its network characteristics are unchanged.

In the first row of Fig. 5, one-ports derived from an n-terminal network are listed. To derive these and other subnetworks, techniques must be developed. For example, a two-port network can be constructed in only two distinct ways, either as a three-terminal network or as a four-terminal (also referred to as bridge) network.

Generalizing, a p-port subnetwork can be constructed in p possible ways, using p+1, p+2, p+3, ..., 2p terminals. For example, three-ports with 4 and 5 and 6 terminals are listed in Fig. 5. Properties of unique subnetworks such as those listed in Fig. 5 require an understanding of the operations of generalizing subnetworks. These techniques will be examined by topological methods of analysis and results will be applied to give a systematic account of subnetworks.

## Permissible Operations to Generate Subnetworks

Two operations on a given terminal of a network are distinctly permissible:

1. The operation of floating a terminal specifies that current entering the terminal is zero. No subsequent operation on this terminal is permissible.

2. The operation of shorting a port specifies that two terminals are connected together. If these two terminals are not to be driven by an external generator, then this operation must be followed by either (2.1) floating the combination or by (2.2) shorting the combination to another terminal. If (2.1) is selected, no subsequent operation on the combination is permissible; if (2.2) is selected, operation (2) must follow.

3. The additional operation of grounding a terminal may be added. This operation establishes a reference terminal.

## Matrix Reduction to Subnetwork

### Indefinite Matrices

An n-terminal network can be uniquely defined by an admittance matrix of order n-1 if it exists at all. It is possible to augment this matrix by addition of a row and column to obtain a resultant matrix of order n. It follows that the sum of the elements of each row and of each column are zero. The resultant singular matrix was called by Shekel[1] an indefinite matrix and applied to the analysis of three-terminal networks such as the transistors. Zadeh[2] and Castro and Happ[3] extended the use of indefinite matrices to circuit analysis and to n-terminal networks. Sharpe and Spain[4] showed that all cofactors of an indefinite matrix are equal and coined the term equi-cofactor matrices.

### Matrix Operations

In an n-terminal network, each row of the indefinite admittance matrix corresponds to a terminal current and each column corresponds to a terminal voltage with respect to an arbitrary reference point. For indefinite admittance matrices, three types of operations will be employed which have a one-to-one correspondence to operations generating subnetworks.

1. Cross-off one row and the corresponding column; this operation is the mathematical counterpart of grounding that terminal, that is, using a terminal as reference.

2. Invert one row; this operation is defined by setting one element of the dependent variable equal to zero and eliminating the corresponding independent variable, and corresponds in flow graph terminology (Nisbet and Happ)[5] to path inversion and setting the newly generated independent variable equal to zero. This operation is equivalent to floating a terminal of the network.

3. Adding two rows and two corresponding columns is equivalent to shorting a port which consists of the corresponding terminals.

Successive application of these three operations generates all subnetworks from a given network. A given subnetwork is independent of the order in which the above operations are applied in its derivation, and thus caution must be exerted to insure a systematic, non-redundant enumeration.

A p-port subnetwork has a matrix of order and rank p. To reduce an n-terminal network to a p-port, (n-p) operations are required.

A technique for a systematic enumeration of non-redundant subnetworks will first be illustrated by an example, then to be followed by an investigation of the logic underlying the generation of subnetworks.

### Illustrative Example: Four-terminal Network

Consider the one-port and two-port subnetworks which can be derived from a four-terminal network. The terminals of a four-terminal two-port must be taken two at a time and yield six pairs, one pair serving as input and the other pair as output, resulting in three unique four-terminal two-ports listed in Fig. 6.

Fig. 6 also lists the three-terminal two-ports which can be obtained from a four-terminal network which are of three distinct types:

1. one terminal floating

2. two terminals shorted at input or at output

3. two terminals taken both as reference terminals.

Since all three-terminal two-ports have one terminal as a reference, it is always possible to cross off one row and one column as indicated by operation 1 in Fig. 6, thereby reducing the 4 by 4 matrix to a 3 by 3 matrix. Subsequent reductions may be of one of three types as shown in Fig. 6 resulting in a total of thirty unique two-ports.

Similarly, seven types of one-ports listed in Fig. 7 are obtained by applying matrix operations 1, 2 and 3 in suitable sequence. Sixty-two one-ports result, which are redundant by a factor of two, since it is immaterial which terminal serves as reference and which terminal is driven.

## Topological Techniques of Enumeration

### Scope

The enumeration techniques used in the preceeding illustrative example are valid for networks of arbitrary complexity but do not provide a sufficiently effective approach to be of practical value for networks with more than four terminals. Topological techniques provide a more rigorous method and a more effective technique for large systems. So far, a general topological solution to enumerate subnetworks is not available; however, as will be shown, the laws of subnetwork generation can be expressed in terms of topological theorems. These theorems provide an accurate determination of one class of subnetwork, as well as several recursion formulae reducing the problem to a point where an order of magnitude estimate of the total number of subnetworks is meaningful.

### Applicable Theorems

Theorem 1: If T is number of trees in a (t-1) terminal network, then the total number of p-ports having terminals obtained from a (t+1)-terminal network is

$$N(t + 1,t,p) = (T/2)(t + 1)(t + 2)$$

Proof: The number of trees T(t) in a t-terminal network is equal to the total number of unique (t-1)-ports networks that exist in the t-terminal network. If one more terminal is added to the t-terminal network and if we are to continue to have a (t-1)-port subnetwork, then the added terminal must be either (a) left floating or (b) connected to an existing terminal. For case (a), there will be (t+1)T subnetworks, since there are (t+1) choices of terminals to float. For case (b) there will be (t/2)(t+1)T subnetworks, since there are (t/2)(t+1) choices resulting from taking (t+1) terminals two at a time. The sum of (a) and (b) give the total stated above. This theorem can be generalized in the following:

Theorem 2: If K(s,t) is the number of distinct ways of reducing s terminals to t terminals, and N(t,t,p) is the number of p-ports networks of a t-terminal network, then K(s,t)N(t,t,p) =

$N(s,t,p)$ is the total number of p-port networks having t-terminals obtained from an s-terminal network. The proof of this theorem is similar to that of Theorem 1, but considered beyond the scope of this summary.

Theorem 3: The number of trees in a t-terminal network is given by $T(t) = t^{(t-2)}$ as proven by Trent[6]; thus $N(t,t,t-1) = t^{(t-2)}$.

Theorem 4: A 2p-terminal network generates $N(2p,2p,p) = (2p)!/(2^p p!)$ unique p-ports.

Proof: In a (2p-2)-terminal network the number of (p-1)-ports is $N(2p-1,\ 2p-1,\ p-1)$. If two more terminals are added it is seem that by using one of the added terminals in combination with each of (2p-1) remaining terminals, we have

$$N(2p,2p,p) = (2p-1)N(2p-2,\ 2p-2,\ p-1)$$

Since the product of all odd numbers up to (2p-1) is $(2p)!/(2^p p!)$, it is readily verified that

$$N(2p,2p,p) = (2p)!/(2^p p!)$$

## Illustrative Example: Five-terminal Network

The number $N(5{:}4{:}3)$ of four-terminal three-ports which can be obtained from a five-terminal network is given by theorem 1 as

$$N(5{:}4{:}3) = (T/2)(t+2)(t+1) = 240$$

since $t = 4$ and $T(4) = 16$

theorem 3 yields $N(5{:}5{:}4) = T(t) = 125$ with $t = 5$. Similarly, theorems 1 and 3 yield

$$N(5{:}3{:}2) = (T/2)(t+2)(t+1) = 45, \text{ with } t=4$$

and $T(3) = 3$,

the remaining entries in column 5 of Fig. 5 have to be computed by the methods developed in Section 2. Thus

$N(5{:}2{:}1) = 160$ with the aid of Fig. 8

$N(5{:}3{:}2) = 240$ with the aid of Fig. 9

While the last entry is readily shown by inspection to be $N(5{:}5{:}3) = 30$, however, a formula for $N(t,t,t-2)$ appears to require a different approach from the above and is indicative of problems awaiting solution. The total number of subnetworks obtained from a five-terminal network is 840, as shown in Fig. 10.

## Summary of Results

Computed values of $N(s,t,p)$ are shown in Fig. 5. Additional values are provided by the relationships

$$N(t,t,t-1) = t^{(t-2)} = T(t)$$

$$N(2p,2p,p) = (2p)!/(2^p p!) = P(p)$$

and the recursion formulae

$$N(t+1,t,p) = (1/2)(t+1)(t+2)N(t,t,p)$$

which is a special case for $s = t+1$ of $N(s,t,p) = K(s,t)N(t,t,p)$ where

$$K(t+1,t) = (1/2)(t+1)(t+2)$$

as well as $K(4,2) = 31$ and $K(5,3) = 80$ are known.

An order-of-magnitude estimate of the number of subnetworks is made in Fig. 10 and Fig.11 by extrapolating calculated values and trends indicated by the above formulae. The one-ports increase roughly as $s!$, while the total number of subnetworks must exceed $s^{s-2}$ by a factor of the order of $s$ or $s^2$. Hence, for the total number of subnetworks an order-of-magnitude of $s^2$ appears reasonable. Fig. 11 is a plot of the number of subnetworks as a function of p with s as a parameter. For constant s, the number of subnetworks

$$N(s,p) = N(s,t,p)$$

has two limiting values, namely $s^{s-2}$ and roughly $s!$. Between these limiting values $N(s,p)$ increases and when $p \quad s^{1/2}$ reaches a maximum of the order-of-magnitude

$$N(s,s^{1/2}) \sim s^{s-1}$$

Calculated and estimated number of subnetworks are shown in Fig. 11. The implication of these results are significant and fundamental in the development of circuit design concepts, such as exploration of the distributed parameter networks developed by Castro and Happ[3]. Indeed, an entirely revolutionary concept of network synthesis is foreshadowed by the results here presented. No longer will the circuit designer assemble various and sundry components to obtain a desired circuit response. Instead, it is likely that the circuit designer of the future will shape his own circuit function from a multi-terminal element by generating the subnetwork to fit his specification using the logic of effectively generating the required subnetwork.

## References

1. J. Shekel, "Matrix Representation of Transistor Circuits", Proc. IRE, Vol. 40, pp. 1493-1497, November, 1952.

2. L. A. Zadeh, "Multiple Analysis of Active Networks", IRE Trans. PGCT, Vol. CT4, pp. 97-105, September, 1957.

3. P. S. Castro and W. W. Happ, "Distributed Parameter Circuits and Microsystems Electronics", Proc. NEC, October 1960, Chicago, Ill.

4. G. E. Sharpe and B. Spain, "On the Solution of Networks by Means of the Equicofactor Matrix", IRE Trans. PGCT, Vol. CT7, pp. 230-239, September, 1960.

5. T. R. Nisbet and W. W. Happ, "Flow Graphs Analysis - Visual Engineering Mathematics", Electronic Design, December 9 to January 20 (4 parts), 1959-1960.

6. H. M. Trent, "A Note on the Enumeration and Listing of all Possible Trees in a Connected Linear Graph", Proc. Nat'l. Ac. Sciences, Vol. 40, p. 1004, 1954.

7. J. Shekel, "Voltage Reference Node", Wireless Engineer, Vol. 31, pp. 6-10, January 1954.

8. P. S. Castro, "Microsystem Circuit Analysis", Electrical Engineering, Vol. 80, pp. 535-542, July 1961.

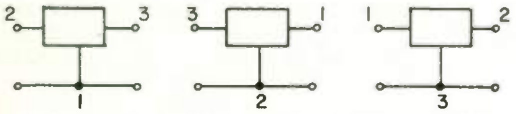


FIG. 1 TWO PORT DERIVED FROM A THREE-TERMINAL NETWORK

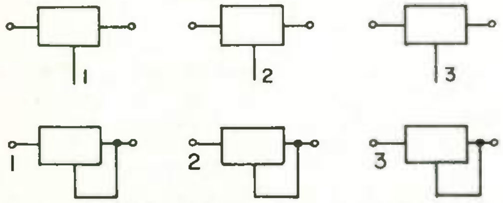


FIG. 2 ONE PORT DERIVED FROM A THREE-TERMINAL NETWORK

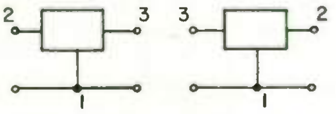


FIG. 3 NON-UNIQUE TWO-PORT

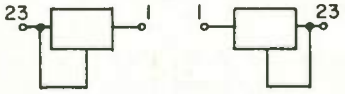


FIG. 4 NON-UNIQUE ONE-PORT

| PORTS (p) / TERMINALS (t) of sub-network | NETWORK TERMINALS (s) | | | | | | |
|---|---|---|---|---|---|---|---|
| p t | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 ONE-PORT | | | | | | | |
| 2 TWO-TERMINAL | 1 | 6 | 31 | 160 | 856 | 4802 | - |
| 2 TWO-PORT | | | | | | | |
| 3 THREE-TERMINAL | | 3 | 30 | 240 | - | - | - |
| 4 FOUR-TERMINAL | | | 3 | 45 | - | - | - |
| 3 THREE-PORT | | | | | | | |
| 4 FOUR-TERMINAL | | | 16 | 240 | - | - | - |
| 5 FIVE-TERMINAL | | | | 30 | 630 | - | - |
| 6 SIX-TERMINAL | | | | | 15 | - | - |
| 4 FOUR-PORT | | | | | | | |
| 5 FIVE-TERMINAL | | | | 125 | - | - | - |
| 6 SIX-TERMINAL | | | | | - | - | - |
| 7 SEVEN-TERMINAL | | | | | | - | - |
| 8 EIGHT-TERMINAL | | | | | | | 105 |

FIG. 5 NUMBER OF SUBNETWORKS N (s; t; p)

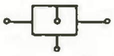
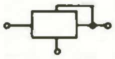
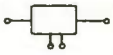
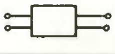
| TYPE | NUMBER TOTAL: 33 | OPERATIONS |
|---|---|---|
|  | 12 | (1) (2) |
|  | 12 | (1) (3) |
|  | 6 | (1) (2) |
|  | 3 | TAKE TERMINALS IN PAIRS |

FIG. 6 NON-REDUNDANT TWO-PORTS DERIVED FROM A FOUR-TERMINAL NETWORK.

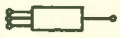
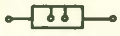
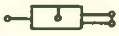
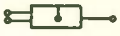
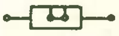
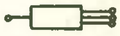
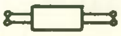
| TYPE | TOTAL: 62 | OPERATIONS |
|---|---|---|
|  | 4 | (1) (1) (1) |
|  | 12 | (1) (2) (2) |
|  | 12 | (1) (2) (3) |
|  | 12 | (1) (1) (2) |
|  | 12 | (1) (3) (2) |
|  | 4 | (1) (3) (3) |
|  | 6 | (1) (1) (3) |

FIG. 7 NON-REDUNDANT ONE-PORTS DERIVED FROM A FOUR-TERMINAL NETWORK.

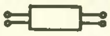
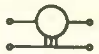
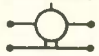
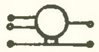
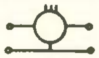
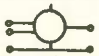
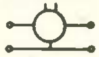
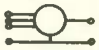
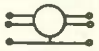
| SUBNETWORKS | OPERATIONS | | | NUMBER |
|---|---|---|---|---|
|  | I | III | II | N |
|  | (1) | (1) | (1) | 10 |
|  | (1) | (1) | (2) | 30 |
|  | (1) | (1) | (3) | 30 |
|  | (1) | (2) | (2) | 30 |
|  | (1) | (2) | (3) | 60 |
|  | (1) | (3) | (2) | 30 |
|  | (1) | (3) | (3) | 20 |
|  | (1) | (3) | (3) | 30 |

FIG. 9 N(5:3:2)=240

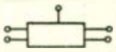
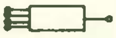
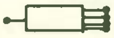
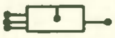
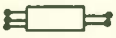
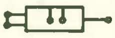
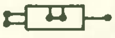
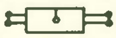
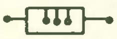
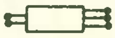
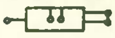
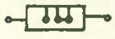
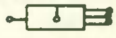
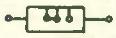
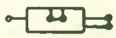
| SUBNETWORK | OPERATIONS | | | | NUMBER |
|---|---|---|---|---|---|
|  | I | II | III | IV | N |
|  | (1) | (1) | (1) | (1) | 5 |
|  | (1) | (3) | (3) | (3) | 5 |
|  | (1) | (1) | (1) | (2) | 20 |
|  | (1) | (1) | (1) | (3) | 10 |
|  | (1) | (1) | (2) | (2) | 30 |
|  | (1) | (1) | (3) | (3) | 30 |
|  | (1) | (1) | (2) | (3) | 30 |
|  | (1) | (2) | (2) | (2) | 20 |
|  | (1) | (1) | (3) | (3) | 10 |
|  | (1) | (2) | (2) | (3) | 30 |
|  | (1) | (2) | (3) | (2) | 60 |
|  | (1) | (2) | (3) | (3) | 20 |
|  | (1) | (3) | (3) | (3) | 20 |
|  | (1) | (3) | (3) | (2) | 30 |

FIG. 8 N(5:2:1)=160

| PORTS (p) | NETWORK TERMINALS (s) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 | s |
| ONE-PORT | 1 | 6 | 31 | 160 | 856 | 4702 | s |
| TWO-PORT | | 3 | 33 | 285 | $\sim 10^3$ | $\sim 10^4$ | $\sim s^{s-1}$ |
| THREE-PORT | | | 16 | 270 | $\sim 10^3$ | $\sim 10^4$ | $\sim s^{s-1}$ |
| FOUR-PORT | | | | 125 | $\sim 10^3$ | $\sim 10^4$ | $\sim s^{s-1}$ |
| SUBNETWORK | 1 | 9 | 80 | 840 | $\sim 10^4$ | $\sim 10^5$ | $\sim s^s$ |

FIG. 10 SUMMARY OF SUBNETWORKS N(s,p)

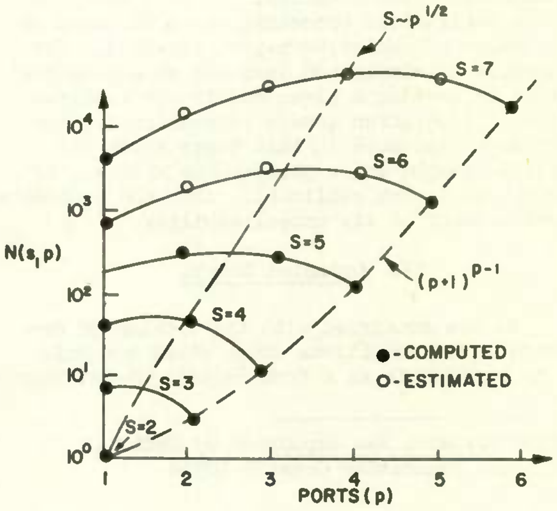


FIG. 11 CALCULATED AND ESTIMATED NUMBER OF SUBNETWORKS

# REALIZATION OF FUNDAMENTAL CIRCUIT AND CUT-SET MATRICES*

C. C. Halkias and W. H. Kim
Department of Electrical Engineering
Columbia University
New York 27, New York

## Summary

A simple procedure for the realization of a fundamental circuit matrix $B_f$ or a fundamental cut-set matrix $C_f$ is given. This procedure constitutes a necessary and sufficient condition for the realization of the matrices $B_f$ or $C_f$ of an oriented, connected graph. It is shown that the problem of realizing $B_f$ or $C_f$ is reduced to the problem of realizing a resistive n-port with exactly n + 1 nodes; theorems and illustrations are provided.

## I. Introduction

The problem of realizing a fundamental circuit matrix or a fundamental cut-set matrix had remained unsolved for a long time. The first solution was offered by Gould[1] in 1957. In 1959 two more solutions were proposed, one by Guillemin[2] and another by Løfgren.[3] In the same year, Auslander and Trent[4] gave an alternate solution and Tutte[5] published his work on matroids and graphs. In 1960 Tutte[6] developed a realization algorithm in an attempt to give practical significance to his highly theoretical results. Other research workers who offered solutions are: Mayeda[7], Okada and Young[8] in 1960. The reader who is familiar with all the above solutions must be aware of their complexity and of the labor involved in attempting to realize any given fundamental cut-set or fundamental circuit matrix. In this paper we give a realization procedure which is based on the theory of resistive n-port networks. This procedure is simple and involves no substantial effort in testing a given matrix for realizability. If a given matrix is realizable, the procedure discussed in this paper gives all possible graphs which satisfy the matrix. If the matrix is not realizable, then the procedure forms a proof of its unrealizability.

## II. Oriented Graphs

We are concerned with the problem of determining the conditions under which a matrix $C_f$ is realizable as a fundamental cut-set matrix of a connected graph. The matrix $C_f$ is assumed to be of order $n$ by $e(n \leq e)$ and rank $n$, to have as elements $\pm 1$ and $0$ and that it may be partitioned in a basic form[10]

$$C_f = [U_1 \quad C_{12}] \tag{1}$$

where $U_1$ is a unit matrix of order n.

An interesting necessary condition on the sign pattern of the matrix $C_f$ follows from the fact that $C_f$ must necessarily be a unimodular** matrix. Let the elements of $C_f$ be $c_{ij}$.

Necessary Condition 1

A necessary condition for the realization of a matrix $C_f$ as a fundamental cut-set matrix is that for any pair of indices $i$, $j$ $(i \neq j)$ we must have:

$$c_{ip}c_{jp} = +1 \text{ or } 0, \quad \text{for } p = 1,2,..,e \tag{2a}$$

or

$$c_{ip}c_{jp} = -1 \text{ or } 0, \quad \text{for } p = 1,2,..,e \tag{2b}$$

Proof: Let $c_{ih}c_{jh}$ and $c_{ik}c_{jk}$ be different from zero and let them have opposite signs, contrary to the above condition,

$$c_{ih}c_{jh} = -c_{ik}c_{jk} \tag{3}$$

Consider the subdeterminant

$$\left|C_{ij}\right| = \begin{vmatrix} c_{ih} & c_{ik} \\ c_{jh} & c_{jk} \end{vmatrix} = c_{ih}c_{jk} - c_{jh}c_{ik} \tag{4}$$

multiply both sides of (4) by $c_{jh}c_{jk}$ which is different from zero, i.e.,

$$c_{jh}c_{jk}\left|C_{ij}\right| = c_{jk}^2 c_{jh}c_{ih} - c_{jh}^2 c_{ik}c_{jk}$$

$$= 2\, c_{jh}c_{ih} = \pm 2 \tag{5}$$

which contradicts the hypothesis that $C_f$ is a unimodular matrix.

One of the implications of the first condition is the requirement that if the signs of the first row of $C_f$ are all (+) or 0 then every element of every other row (with all columns having a zero on the first row crossed out) must be either (+) or 0, or (-) or 0; that is, there cannot be (+) and (-) signs on the same row.

* This research was supported by National Science Foundation Grant G-10354

** A matrix all of whose elements and subdeterminants are 1, -1 or 0 is called a unimodular matrix (or E-matrix). (See Reference 9.)

Necessary Condition 2

A necessary condition that a matrix $C_f$ be realizable as a fundamental cut-set matrix of an oriented graph is that $C_f$ should not include the following submatrix. (See References 1,5, 6, 9, 10, 11, 12)

$$K = \begin{bmatrix} x & x & o & x \\ x & o & x & x \\ o & x & x & x \end{bmatrix} \qquad (6)$$

where the x's stand for non-zero elements. This condition follows from the fact that there cannot exist in a unimodular matrix a submatrix of the form K shown in (6).

Let us now consider a simple approach for the realization of $C_f$. A realizable $C_f$ matrix specifies a connected graph with n + 1 nodes and e edges; moreover, it defines a tree with a specific tree-branch orientation. If we think of the edges as conductances $d_1, d_2, .., d_e$ and if we excite the network by connecting n current generators oriented in the same direction as the tree-branches at the node-pairs specified by the tree-branches, then we obtain the short-circuit admittance matrix Y of an n-port resistive network with (n + 1) nodes which has a tree-port-structure.

It is well known that[10] for an n-port resistive network with (n + 1) nodes

$$Y = C_f D C_f^t \qquad (7)$$

where $C_f$ is the fundamental cut-set matrix with respect to the tree corresponding to the port-structure. $C_f^t$ is the transpose of $C_f$ and D is the diagonal matrix with the positive conductances $d_1, d_2, .., d_e$ as main diagonal elements. If $C_f$ is realizable with (n + 1) nodes and e edges, then Y is realizable; if $C_f$ is not realizable, then Y is not realizable either. Hence our problem of realizing $C_f$ is equivalent to the problem of realizing the n-port network with (n + 1) nodes characterized by the short-circuit admittance matrix Y given in (7).

Let us assume that the conductances $d_1, d_2, .., d_e$ are all unit-conductances; then the expression in (7) is reduced to $Y = C_f C_f^t$, and the product $C_f C_f^t$ is the Grammian of the fundamental cut-set matrix. We have now reduced the problem of realizing the matrix $C_f$ to the problem of realizing the Grammian of $C_f$ as the short-circuit admittance matrix of a resistive network with (n + 1) nodes and e unit-conductances. We now state the above discussions in the next theorem.

Theorem 1.

The matrix $C_f = [U_1 \;\; C_{12}]$ with elements 0, ± 1 is realizable as a fundamental cut-set matrix of an oriented graph if and only if the Grammian $Y = C_f C_f^t$ is realizable as a short-circuit admittance matrix of an n-port resistive network containing (n + 1) nodes and e unit-conductances.

The proof of the above theorem follows because the short-circuit admittance matrix of an n-port network described on the resistive network with (n + 1) nodes is given by

$$Y = C_f D C_f^t \qquad (8)$$

and from the fact shown by Cederbaum[13], that the congruence of Eq. (8) is unique if $C_f$ is a non-redundant unimodular matrix* and D is a positive diagonal matrix.

Our problem has now been reduced to the problem of realizing a Y-matrix of order n as a short-circuit admittance matrix of a resistive n-port network with n + 1 nodes. In order to make the procedure of realizing $C_f$ simple, we shall consider the following theorems and corollaries:

Definition 1

A linear-tree is a tree whose branches are all contained in a single path. (See Fig. 1a).

Theorem 2

If a matrix $C_f$ contains a column with all non-zero elements, then the tree on which $C_f$ is based is a linear-tree.

Proof: The edge corresponding to the column with all non-zero elements links all tree-branches; hence all tree-branches are contained in a single path and thus the tree is linear. The order of the tree branches is, of course, not known yet.

Corollary 1

The matrix $C_f$ of order n by e and with a column of all non-zero elements is realizable if and only if the Grammian $Y = C_f C_f^t$ is realizable as a short-circuit admittance matrix of an n-port resistive network with (n + 1) nodes, e unit resistors, and a linear-tree port-structure.

Proof: The corollary follows the Theorems 1 and 2.

Theorem 3

If the matrix $C_f$ contains a column with non-zero elements in the rows i,j,..,k, then the corresponding tree-branches i,j,..,k form a linear-tree if all other tree-branches are short-circuited.

Proof: The edge corresponding to the column with non-zero elements in the rows i,j,..,k links tree-branches i,j,..,k. If all other tree-branches are short-circuited, then tree branches i,j,..,k are contained in a single path and thus form a linear-tree.

* An E-matrix is called non-redundant if it has no columns with all zero elements and no two columns in which the pattern of zero and non-zero elements is identical. An E-matrix is also called a unimodular matrix.

Corollary 2

If the matrix $C_f$ contains a column with non-zero entries in the rows $i,j,..,k$, then a necessary condition for the realization of $C_f$ is that the matrix $Y_{i,j,..,k}$ be realizable as a short-circuit admittance matrix of a multi-port resistive network with a linear-tree port-structure formed by the ports $i,j,..,k$. (Not necessarily in the order $i,j,..,k$.)

Proof: The proof follows from Theorems 1 and 3.

Corollaries 1 and 2 require that we should be able to determine whether a given Y-matrix of real elements and of order $n$ is realizable with $(n + 1)$ nodes and a linear-tree port-structure consisting of $n$ ports. Of great importance for our purposes is the order of the ports in the linear-tree port-structure; this will enable us to derive the tree on which $C_f$ is based and the complete realization of $C_f$ can follow by inspection. In the following theorem we give the necessary and sufficient conditions for the realization of $Y$ with a linear-tree port-structure. The proof can be found in references (14, 15).

Theorem 4

The necessary and sufficient conditions for the realization of the n'th order matrix $Y = [y_{ij}]$ as a short-circuit admittance matrix of an n-port resistive network with $(n + 1)$ nodes and the ordered linear-tree port-structure of Fig. 1b are:

$$\begin{aligned}&(a)\quad y_{ij} \geq 0\\ &(b)\quad y_{ij} + y_{i-1,j+1} \geq y_{i-1,j} + y_{i,j+1} \qquad (9)\\ &\qquad\qquad \text{for all } i \text{ and } j.\end{aligned}$$

In Eq. (9) all elements of the n by n Y-matrix with an index larger than $n$ or less than one are defined to be identically equal to zero. A matrix which satisfies the conditions of Theorem 4 is called "uniformly tapered"[14,15].

On the basis of the previous theorems and corollaries, we now give a simple procedure for the realization of $C_f$. If $C_f$ is not realizable, then the procedure forms a proof of its unrealizability.

Realization Procedure

Let us first interpret two operations on the matrix $Y$.

I. Changing signs of all elements of row k and column $k$ means that the polarity of the k'th port (and the orientation of the k'th tree-branch) is reversed.

II. Interchanging rows and columns $i$ and $k$ means an interchange in the labeling of ports i and $k$ (an interchange in the numbering of tree-branches $i$ and $k$). We now give the realization procedure.

Case 1: $C_f$ contains a column with all non-zero elements.
(a) From $Y = C_f C_f^t$
(b) Place $Y$ in a uniformly tapered form.*
(If this is not possible, then $C_f$ is not realizable).
(c) From step (b) obtain the order and orientation of the branches of the linear-tree.
(d) With the tree known, the realization of $C_f$ can proceed by inspection. All possible 2-isomorphic graphs[10] realizing $C_f$ can also be obtained by inspection.

Case 2: $C_f$ does not contain a column with all non-zero elements.
(a) From $Y = C_f C_f^t$
(b) Examine the column of $C_f$ with the largest number of non-zero elements in the rows $i,j,..,k$ and obtain the matrix $Y_{i,j,..,k}$.
(c) Place $Y_{i,j,..,k}$ in a uniformly tapered form and thus derive the order and orientation of tree-branches $i,j,..,k$.* (If this is impossible, the matrix $C_f$ is not realizable).
(d) Repeat steps (b) and (c) for the columns with a smaller number of non-zero elements. By combining the various tree parts, obtain the tree on which $C_f$ is based. (If this is impossible, then $C_f$ is not realizable). The realization of $C_f$ can now be completed by inspection.
The procedure is best illustrated with the following example.

* It may be required to multiply some rows and columns by -1 or to interchange some rows and columns.

Example 1.

Realize the following fundamental cut-set matrix.

$$C_f = \begin{array}{ccccccccccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \end{array} \tag{10}$$

Step (a).

$$Y = C_f C_f^t = \begin{bmatrix} 4 & -1 & 2 & 0 & 0 & 0 & -1 & 0 \\ -1 & 5 & 3 & -1 & 0 & -1 & 0 & -1 \\ 2 & 3 & 6 & -1 & 0 & -1 & -1 & -1 \\ 0 & -1 & -1 & 6 & 2 & 3 & -1 & -1 \\ 0 & 0 & 0 & 2 & 4 & -1 & -1 & 0 \\ 0 & -1 & -1 & 3 & -1 & 5 & 0 & -1 \\ -1 & 0 & -1 & -1 & -1 & 0 & 3 & 0 \\ 0 & -1 & -1 & -1 & 0 & -1 & 0 & 3 \end{bmatrix} \tag{11}$$

Step (b).

If we consider the 17th column of the matrix of (10) (this column has the maximum number of non-zero elements) then we have:

$$Y_{2,3,4,6} = \begin{bmatrix} 5 & 3 & -1 & -1 \\ 3 & 6 & -1 & -1 \\ -1 & -1 & 6 & 3 \\ -1 & -1 & 3 & 5 \end{bmatrix} \tag{12}$$

A possible order for tree-branches (or ports) 2, 3, 4, 6 is shown in Fig. 2a.

Step (c)

For the remaining columns with three non-zero elements we have:

$$Y_{1,3,7} = \begin{bmatrix} 4 & 2 & -1 \\ 2 & 6 & -1 \\ -1 & -1 & 3 \end{bmatrix} \qquad Y_{2,3,8} = \begin{bmatrix} 5 & 3 & -1 \\ 3 & 6 & -1 \\ -1 & -1 & 3 \end{bmatrix} \tag{13}$$

(a) (b)

$$Y_{4,5,7} = \begin{bmatrix} 6 & 2 & -1 \\ 2 & 4 & -1 \\ -1 & -1 & 3 \end{bmatrix} \tag{13}$$

(c)

One should note here that $Y_{7,4,5}$ is also uniformly tapered;

$$Y_{7,4,5} = \begin{bmatrix} 3 & -1 & -1 \\ -1 & 6 & 2 \\ -1 & 2 & 4 \end{bmatrix} \tag{13}$$

(d)

The order for tree-branches 1,3,7 2,3,8 and 4,5,7 is shown in Figures 2b,c,d,e.

Step (d)

Combining the sub-trees of Steps (b) and (c), we obtain the tree shown in Fig. 3.

The graph realizing $C_f$ can now be obtained by inspection, which is shown in Fig. 4.

Let us now assume that we are concerned with the problem of determining the conditions under which matrix $B_f$ is realizable as a fundamental circuit matrix of a connected graph. The matrix $B_f$ of a graph is assumed to be of order $n$ by $e(n \leq e)$ and rank n, and to have as elements $\pm 1$ and $0$ and that it may be partitioned into a basic form with respect to a tree of the graph.

$$B_f = [B_{12} \quad U_2] \tag{14}$$

where $U_2$ is a unit matrix of order $n$. The fundamental cut-set matrix based on the same tree is then found to be

$$C_f = [U_1 \quad -B_{12}^t] \tag{15}$$

where $U_1$ is a unit matrix of order $(e - n)$. The matrix $C_f$ can be realized as discussed previously. Thus we reduce the problem of realizing $B_f$ to the problem of realizing the corresponding $C_f$. Since the matrix $B_f$ must be a unimodular matrix, it is obvious that necessary conditions 1, 2 are also valid for fundamental circuit matrices.

Example 2.

Determine if the following matrix, discussed by Guillemin[16] in connection with the realization of an open-circuit resistance matrix, is realizable as a fundamental circuit matrix of an oriented graph.

$$B_f = \begin{array}{c|ccccccccccccccc|}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\
\hline
1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
6 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
7 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
8 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
9 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
10 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\hline
\end{array} \tag{16}$$

The cut-set matrix corresponding to the same tree is:

$$C_f = \begin{array}{c|ccccccccccccccc|}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\
\hline
1 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & -1 & -1 & -1 & -1 \\
2 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & -1 & 0 \\
3 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & -1 & -1 & 0 \\
4 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & -1 & 0 & -1 \\
5 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 1 & -1 & 0 & 0 & -1 \\
\hline
\end{array} \tag{17}$$

We may now form the triple product $Y = C_f\ D\ C_f^t$, where D is a unit matrix of order 15.

$$Y = \begin{bmatrix} 9 & 3 & 3 & 3 & 3 \\ 3 & 5 & 1 & -1 & 1 \\ 3 & 1 & 5 & 1 & -1 \\ 3 & -1 & 1 & 5 & 1 \\ 3 & 1 & -1 & 1 & 5 \end{bmatrix} \quad (18)$$

Applying the methods described in previous sections of this paper we can easily realize the matrix $C_f$. Branches 12 and 13 link the tree paths 1,2,5 and 1,3,4 respectively. The submatrices of Y corresponding to the ports 1,2,5 and 1,3,4 are:

$$Y_{1,2,5} = \begin{bmatrix} 9 & 3 & 3 \\ 3 & 5 & 1 \\ 3 & 1 & 5 \end{bmatrix} \qquad Y_{1,3,4} = \begin{bmatrix} 9 & 3 & 3 \\ 3 & 5 & 1 \\ 3 & 1 & 5 \end{bmatrix} \quad (19)$$

(a) (b)

It is readily recognized that the order of ports in these paths is: 2,1,5 and 3,1,4. Thus the tree is necessarily of the two types shown in Figures 5a, b. The alternative shown in Fig. 5a may be excluded because branch 10 links ports 2 and 4 only. The realization of the matrix $C_f$ and hence of the matrix $B_f$ is shown in Figure 6.

## III. Non-Oriented Graphs

The extension of the realization procedure to non-oriented graphs is basically straightforward. Suppose a fundamental cut-set matrix $C_f$ contains a column with all 1's. Then this matrix must be based on a linear-tree. If $C_f$ is realizable, the orientation of the graph elements can be considered to be as shown in Fig. 7, and hence all 1's of $C_f$ can be taken as +1's. Hence the realization procedure for oriented graphs is valid also for this case. If $C_f$ contains a column with 1's in the rows i,j,..,k, then upon crossing out all other rows of $C_f$, the resulting matrix is based on a linear-tree and its non-zero elements can be considered as +1's. In conclusion we see that the realization procedure for the matrices $C_f$ and $B_f$ of a non-oriented graph is identical to the previously discussed procedure for oriented graphs.

### Example 3.

Determine if the following matrix, given by Mayeda,[7] is realizable as a fundamental cut-set matrix of a graph.

$$C_f = \begin{array}{ccccccccc} a & b & c & d & e & f & g & h & i \\ \hline 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{array} \quad (20)$$

We have:

$$Y_{1,2,3,4} = C_f C_f^t = \begin{bmatrix} 4 & 2 & 1 & 2 \\ 2 & 3 & 1 & 1 \\ 1 & 1 & 4 & 2 \\ 2 & 1 & 2 & 4 \end{bmatrix} \quad (21)$$

A transposition of rows and columns 1,2,3,4 into the sequence 3,4,1,2 results in the uniformly tapered matrix. Ports 1,2,3,4 correspond to tree-branches f,g,h,i.

$$Y_{3,4,1,2} = \begin{bmatrix} 4 & 2 & 1 & 1 \\ 2 & 4 & 2 & 1 \\ 1 & 2 & 4 & 2 \\ 1 & 1 & 2 & 3 \end{bmatrix} \quad (22)$$

The realization of this matrix is shown in Fig. 8. The graph contains exactly 5 nodes and 9 elements.

## References

1. Gould,R., "The Application of Graph Theory to the Synthesis of Contact Networks", Doctorate Thesis, Harvard University, 1957.
2. Guillemin,E.A., "How to Grow Your Own Trees from Given Cut-Set or Tie-Set Matrices", IRE Transactions on Circuit Theory, Special Supplement, vol. CT-6, pp. 110-126, May,1959.
3. Lofgren,L., "Irredundant and Redundant Boolean Branch Networks", IRE Transactions on Circuit Theory, Special Supplement, vol. CT-6, pp.158-175, May, 1959.
4. Auslander,L. and Trent,M., "Incidence Matrices and Linear Graphs", J. Math and Mech., vol. 8, p. 827, 1959.
5. Tutte, W.T., "Matroids and Graphs", Trans. Am. Math. Soc., vol. 90, pp.527-552, 1959.
6. Tutte, W.T., "An Algorithm for Determining Whether a Given Binary Matroid is Graphic", Proc. Am. Math. Soc., vol. 11, pp.905-917, December, 1960.
7. Mayeda,W., "Necessary and Sufficient Conditions for the Realizability of Cut-Set Matrices", IRE Transactions on Circuit Theory, vol. CT-7, pp. 79-81, March, 1960.

8. Okada, Satio and Young, Kwok-Ping, "Ambit Realization of Cut-Set Matrices into Graphs", Research Report, Microwave Research Institute of the Polytechnic Institute of Brooklyn, 1961.

9. Cederbaum, I., "Conditions for the Impedance and Admittance Matrices of N-Ports Without Ideal Transformers", Proc. I.E.E., Monograph No. 276R, January, 1958.

10. Kim, W.H. and Chien,R.T., Topological Analysis and Synthesis of Communication Networks, Columbia University Press, New York, June, 1962.

11. Seshu,S. and Reed,M.B., Linear Graphs and Electrical Networks, Addison-Wesley, 1961.

12. Whitney,H., "On the Abstract Properties of Linear Dependence", Am. J. Math., vol. 57, pp. 509-533, 1935.

13. Cederbaum, I., "Matrices All of Whose Elements and Subdeterminants are 1, -1, or 0", Journal of Math. and Physics, vol. 36, pp. 351-361, 1958.

14. Guillemin, E.A., "On the Analysis and Synthesis of Single-Element-Kind Networks", IRE Transactions on Circuit Theory, vol. CT-7, pp.303-312, September, 1960.

15. Biorci, G. and Civalleri, P.P., "Alcune Considerazioni Sulle Sintesi Dei Multipoli Resistivi", Atti Accad. Sci. Torino, vol.94, 1959-1960, or "On the Conductance Matrices with All-Positive Elements", IRE Transactions on Circuit Theory, vol. CT-8,pp.76-77, March, 1961.

16. Guillemin,E.A., "Realization of an Open-Circuit Resistance Matrix", Quarterly Progress Report No. 60, MIT Research Lab. of Electronics, pp.239-248, January,1961.

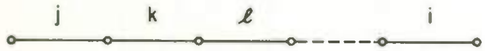


FIG. 1a LINEAR TREE

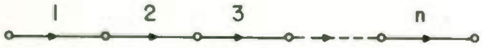


FIG. 1b ORDERED LINEAR TREE

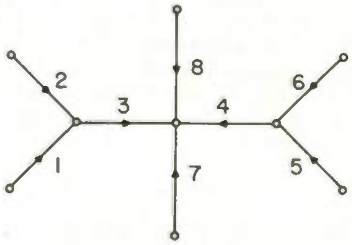
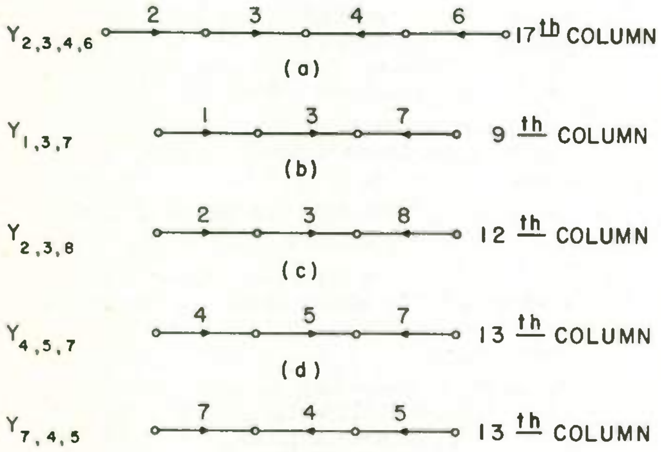


FIG. 3 TREE STRUCTURE



(a)

(b)

(c)

(d)

(e)

FIG. 2 TREE PATHS

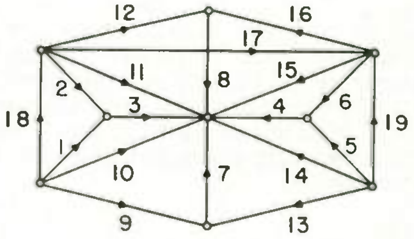
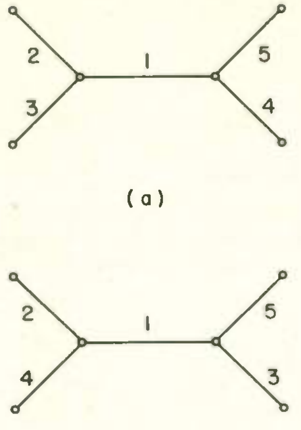


FIG. 4 ORIENTED GRAPH OF EXAMPLE I



(a)

(b)

FIG. 5 POSSIBLE TREE STRUCTURES — EXAMPLE 2

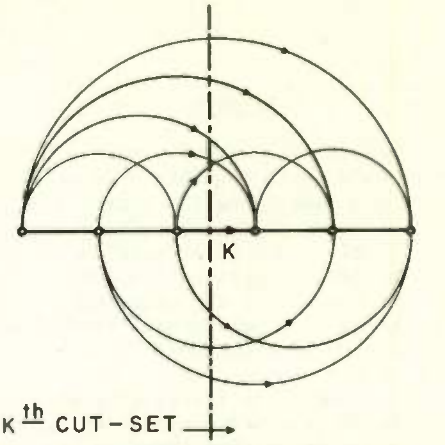


FIG. 7 ORIENTATION OF $K^{th}$ CUT-SET OF A NON-ORIENTED GRAPH

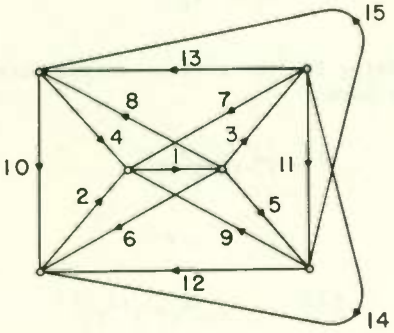


FIG. 6 REALIZATION OF THE FUNDAMENTAL CIRCUIT MATRIX OF EXAMPLE 2

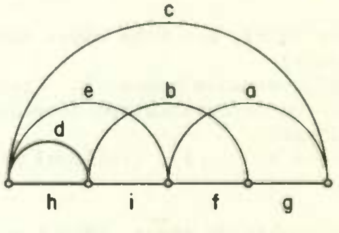


FIG. 8 NON-ORIENTED GRAPH OF EXAMPLE 2

# SOLUTION PROCEDURE FOR SINGLE-ELEMENT-KIND NETWORKS*

S. D. Bedrosian and R. S. Berkowitz
Institute for Cooperative Research and The Moore School of Electrical Engineering
University of Pennsylvania
Philadelphia 4, Pa.

## Summary

The indefinite admittance matrix provides a straightforward means for obtaining admittance parameters for describing the external behavior of a network in terms of the element values. When the element values are considered as the unknowns and the expressions representing short circuit measurements describing its behavior are taken as known constants, the relations obtained become a nonlinear system of equations.

A novel feature in formulating the system of equations is the exclusive use of transfer admittance parameters rather than the use of a reference node. Necessary conditions for "solvability" are given in terms of the "compound" matrix. A general solution procedure is discussed for explicit determination of the element values of single-element-kind networks. Examples are included.

## Introduction

We are concerned with formulation of an adequate mathematical theory for network element value solvability as distinguished from the empirical techniques usually practiced in electronic maintenance. By solvability we mean the ability to determine uniquely the value of all the unknown elements of a given multiterminal network. This concept of theoretical solvability was introduced by Berkowitz.[1]

The networks being treated are considered to have a known configuration and are such that measurements can be made at a limited number of terminals. See Fig. 1. Three types of nodes are permitted:

Accessible nodes, A. (the usual external terminals)
Partially accessible nodes, P. (terminals restricted to application or measurement of voltage)
Inaccessible nodes, I. (internal or "concealed" nodes).

Then the total number of nodes $N = A + P + I$.

## Formulation of Equations

### The Compound Matrix

By suitable labeling of these nodes we can write the indefinite admittance matrix as a "compound" matrix (using subscript $t$ for the transpose).

$$\begin{array}{c|c|c|c|} \text{nodes} & A & P & I \\ \hline A & Y^A & Y^{AP} & Y^{AI} \\ \hline P & Y_t^{AP} & Y^P & Y^{PI} \\ \hline I & Y_t^{AI} & Y_t^{PI} & Y^I \\ \hline \end{array} = \mathcal{M} \qquad (1)$$

This $\mathcal{M}$ matrix provides a direct method of obtaining a "complete set" of parameters representing short circuit measurements for describing the observable external behavior of the network, due to the fact that it facilitates formulation of the transfer admittance from any pair of nodes in the network to any other pair. The set of linear algebraic equations can be represented in matrix form as:

$$\begin{aligned} I^A &= Y^A E^A + Y^{AP} E^P + Y^{AI} E^I \\ I^P &= Y_t^{AP} E^A + Y^P E^P + Y^{PI} E^I \\ 0 &= Y_t^{AI} E^A + Y_t^{PI} E^P + Y^I E^I . \end{aligned} \qquad (2)$$

Then, solving for the current at the accessible nodes, we have

$$I^A = Y^{A'} E^A + Y^{P'} E^P \qquad (3)$$

where

$$\left[Y^A\right]' = Y^A - Y^{AI} (Y^I)^{-1} Y_t^{AI} \qquad (4)$$

$$\left[Y^P\right]' = Y^{AP} - Y^{AI} (Y^I)^{-1} Y_t^{PI} . \qquad (5)$$

If there are no partially accessible nodes in the given network, Eq. 4 simplifies to Kron's reduction formula.[2] Thus, if $P \neq 0$ we have both terms in Eq. 3 and we require the additional equation for defining $\left[Y^P\right]'$. The result given above can be considered as a generalization of Kron's work.

* This paper is based on Chapter IV of a dissertation submitted in partial fulfillment of the requirements for the Ph.D. degree at the University of Pennsylvania, Philadelphia, Pa. The work was supported, in part, by Contract No. DA-36-034-507-ORD-3347RD with the Frankford Arsenal, Army Ordnance Corps.

Since extensive use will be made of equations 4 and 5, we introduce some simplifying notation.

Let,
$$Y_c = Y^{AI}(Y^I)^{-1} \qquad (6)$$

since this matrix product is common to both equations. Then one can write the negative term in equations 4 and 5 respectively as,

$$Y_c Y_t^{AI} = \tilde{Y} \qquad (7)$$

and

$$Y_c Y_t^{PI} = \overset{\approx}{Y}. \qquad (8)$$

Finally,

$$\left[Y^A\right]' = Y^A - \tilde{Y} \qquad (9)$$

and

$$\left[Y^P\right]' = Y^{AP} - \overset{\approx}{Y}. \qquad (10)$$

The $\mathcal{M}$ matrix representation automatically leads to the Q test,[1] i.e. $B \leqslant Q$ is a necessary condition for solvability of a single-element-kind network having B branches.

$$Q = \left(\begin{matrix}\text{independent off-diagonal}\\ \text{element of } Y^A\end{matrix}\right) + (\text{elements in } Y^{AP}) \qquad (11)$$
$$= \tfrac{1}{2}A(A+2P-1).$$

### Conditions on the Matrix Elements

At this point a theorem is given which provides a useful check on setting up the compound matrix representation of a network.

Theorem 1: A set of necessary conditions for solvability in terms of the indefinite admittance matrix $\mathcal{M}$ (Eq. 1) for the network are:

a. no off-diagonal elements of the matrix contain more than a single term.

b. the diagonal elements of the $Y^I$ submatrix must contain $\geqslant 3$ terms each.

c. the $Y^P$ submatrix must be diagonal, i.e. no off-diagonal elements.

d. any $B_f$ branches of the network, not incident on the inaccessible nodes (I), will appear only in the submatrix $Y^A$, or $Y^{AP}$ and $Y_t^{AP}$.

Proof: Parts a, b, and c follow directly from the corresponding parts of Theorem C previously given by Berkowitz.[1]

### The Key Subgraph

With respect to part d of Theorem 1 above, it is observed that the entries in the $Y^I$ submatrix, considered from a topological point of view, represent a special subgraph of the given network. This subgraph consists of the subset $B_k$ of all branches incident on all of the inaccessible nodes. Because of its special importance, we define it as the "key subgraph" $G_k$ of the network. This is shown in Fig. 2, wherein the solid lines are the $B_k$ branches and the broken lines are $B_f$ branches.

Theorem 2: The elements of the matrices $\tilde{Y}$ and $\overset{\approx}{Y}$ defined in equations 7 and 8 respectively only contain terms involving branches appearing in the key subgraph $G_k$ of the network.

Proof: From Theorem 1, part d, we know that the non-key subgraph branches of the network ($B_f$) will appear only in submatrix $Y^A$, or $Y^{AP}$ and $Y_t^{AP}$. The theorem follows directly from the reduction formulas, equations 4 and 5, and the definitions in equations 7 and 8.

### Numbering Convention

Let us be more explicit in terms of equations 9 and 10. In general $\left[Y^A\right]'$ is an A x A symmetric matrix of the form (shown for A = 3; the diagonal elements $D_i$ of the matrix represent the short circuit self admittances)

$$\frac{-1}{\Delta}\begin{bmatrix} -D_1 & C_1 & C_2 \\ & -D_2 & C_3 \\ & & -D_3 \end{bmatrix}$$

and

$\left[Y^P\right]'$ is an A x P matrix of the form (shown for A = 3, P = 2)

$$\frac{-1}{\Delta}\begin{bmatrix} C_4 & C_5 \\ C_6 & C_7 \\ C_8 & C_9 \end{bmatrix}.$$

Observe that the above examples follow a uniform numbering scheme. One proceeds by labeling the $\tfrac{1}{2}A(A-1)$ independent off-diagonal elements of the $\left[Y^A\right]'$ matrix and then continues by labeling the AP elements of the $\left[Y^P\right]'$ matrix. This labeling implies that the measured admittances are being expressed as numerical constants.

### Exclusive Use of Transfer Measurements

Another useful observation has to do with the fact that the sum of the transfer admittances, the C's, equals the self admittance, the $D_i$, for each row in the $\left[Y^A\right]'$ and $\left[Y^P\right]'$ matrices. This permits a novel formulation of the system of $\underline{C}$ equations by use of all of the off-diagonal elements of the "accessible", i.e. the A and P, portion of the matrix $\mathcal{M}$ rather than the conventional approach of selecting an arbitrary reference node with its attendant deletion of row and column. Thus, only the short circuit transfer measurements made on the network are utilized so that in general,

$$\underline{C}_\nu = -Y_{ij}, \quad i \neq j. \qquad (12)$$

Number of Equations

Some remarks are in order regarding the number of equations versus the number of variables, i.e. the B unknown elements of the network. The procedure for formulating equations gives a straightforward and compact way of arriving at a complete set of equations $\underline{C}$. These are analytical equivalents to transfer admittance measurements which can be made on the given single-element-kind network to completely describe its observable external behavior. The total number of equations to be expected is a direct function of the number of accessible and partially accessible nodes. In fact, the number of equations $\underline{C}$ obtained will equal Q, $\frac{1}{2}A(A+2P-1)$, even though in many cases B will be less than Q.

For potentially solvable networks then the number of equations can also be expressed in terms of branches in the graph G;

$$B_q = Q \geqslant B. \qquad (13)$$

Clearly, $B_q = B$ represents a "maximal" condition. Such a network has the maximum number of branches in G for the given number of type A and P nodes. It must be emphasized that by virtue of our technique, the inequality implies that when $B < Q$ we will obtain redundant equations equal in number to $B_q - B$. When there are more equations than unknowns, one cannot say in general that there is a solution to the system of equations. The specified compound matrix method of formulating these equations assures us of "consistency" in the sense that there exist values of the unknown which satisfy all $B_q$ equations.

## General Solution Procedure

Modified System of Equations

The equations derived from single-element-kind networks have some special characteristics. For example, the terms are homogeneous multilinear algebraic forms;

$$\sum_{r} a_{rj} \prod_{k=\varphi_1(r,j)}^{\varphi_{m(I)}(r,j)} x_k = \underline{C}_\nu \qquad (14)$$

where

$$\underline{C}_\nu = \frac{1}{\Delta} C_\nu = -Y_{rj}, \qquad (15)$$

and

$Y_{rj}$ = admittance measurements

$\Delta$ = determinant of the key subgraph

$a_{rj}$ = coefficient (an integer)

$\varphi_{m(I)}(r,j)$ = functional dependence on number of internal nodes

$r = 1, 2, 3, \ldots, m$

$\nu = 1, 2, 3, \ldots, Q$

$k = 1, 2, 3, \ldots, (I+1).$

Each term is linear with respect to each of its variables individually. In general no two or more terms of an equation are alike. This also applies to terms between equations of a set.

With regard to the general form shown in Eq. 14, note also that the number, n, of linear variables in the product is related to the number of inaccessible nodes in the network. For example, with I = 2 each term consists of a product of three variables. The relationship implied by $\varphi_{m(I)}(r,j)$ in Eq. 14 is given by:

$$n = I + 1. \qquad (16)$$

Thus noting Eq. 15, we state:

Theorem 3: The system of nonlinear equations 17 derived from the compound matrix $\mathcal{M}$ by use of the generalized reduction formulas, Eqs. 4 and 5, can be represented by a modified system of equations, Eq. 19, wherein the determinant $\Delta$ of the key subgraph $G_k$ is absorbed as a scale factor.

Proof: We can rewrite Eq. 14 in the form of a functional relationship.

$$\underline{C}_\nu = \frac{1}{\Delta} H_n(x_1, x_2, \ldots) \qquad (17)$$

where

$H_n$ is homogeneous of order n
$\Delta$ is homogeneous of order n-1
$\nu = 1, 2, 3, \ldots, Q.$

Similarly we can write

$$\Delta = \mathcal{D}_{n-1}(x_1, x_2, \ldots). \qquad (18)$$

Now consider the set of equations

$$\underline{C}_\nu = H_n(y_1, y_2, \ldots). \qquad (19)$$

Then

$$H_n(y_1, y_2, \ldots) = \frac{1}{\Delta} H_n(x_1, x_2, \ldots). \qquad (20)$$

Suppose that $y_1, y_2, \ldots$ are unique solutions to Eq. 19. Also let

$$x_j = \sqrt[n]{\Delta}\, y_j \qquad (21)$$

$$\therefore \quad H_n(x_j) = \left(\sqrt[n]{\Delta}\right)^n H_n(y_j)$$

$$= \Delta H_n(Y_j)$$

$$= \Delta \underline{C}_\nu \qquad (22)$$

Thus equations 17 and 20 are satisfied. Using Eq. 21, we can also write,

$$\Delta = \mathcal{D}_{n-1}(x_j) = \left(\sqrt[n]{\Delta}\right)^{n-1} \mathcal{D}_{n-1}(y_j) \qquad (23)$$

$$\therefore \quad (\,\Delta\,)^{1/n} = \mathcal{D}_{n-1}(y_j) \qquad (24)$$

$$\Delta = \left[\mathcal{D}_{n-1}(y_j)\right]^n \qquad (25)$$

Finally then

$$x_j = y_j \mathcal{D}_{n-1}(y_j). \qquad (26)$$

For convenience, we shall indicate the change of variables by means of primes on the variables, i.e. the unknown elements of the network, in the equations. Then the modified system of equations 19 have the alternate form of representation to Eq. 20 namely:

$$\underline{C}_\nu = H_n(x_1', x_2', \ldots). \qquad (27)$$

Note that Eqs. 24 and 26 define what is, in effect, a network element scale factor. To simplify writing it we use the symbol $\lambda$.

<u>Theorem 4</u>: The solution of the modified system of equations, Eq. 27 combined with the equation for $\Delta$, the determinant of the key subgraph, provides a complete solution for the primary variables which are the unknown elements of the given network.

<u>Proof</u>: This follows from Theorem 3. The scale factor then is given by,

$$\lambda = \sqrt[n]{\Delta} \qquad (28)$$

## Topological Implications

Earlier it was indicated that the entires in the $Y^I$ matrix, considered from a topological point of view, represent a special subgraph $G_k$ of the given network. Further implications are given in the next theorem followed by an illustrative example.

<u>Theorem 5</u>: Given a potentially solvable single-element-kind network of B branches, there exists a subset $\underline{C}^k$ of the complete set of the system of equations $\underline{C}$ derived by use of the indefinite admittance matrix $\mathcal{M}$. Solution of this $\underline{C}^k$ subset of the system of equations implies solution of the network.

<u>Proof</u>: Theorem 1 indicates that branches $B_f$ of the network not occuring in the key subgraph can only appear in matrices $Y^A$, or $Y^{AP}$ and $Y_t^{AP}$. From Theorem 2 we know that matrices $\hat{Y}$ and $\tilde{\hat{Y}}$ only contain branches $B_k$ within the key subgraph of the network. Theorem 4 permits absorbing $\Delta$ as a scale factor. Examination of the indefinite admittance matrix $\bar{\mathcal{M}}$ and the matrix manipulations leading to the matrices $[Y^A]'$ and $[Y^P]'$ indicates that there will always exist a subset $\underline{C}^k$ of the $\underline{C}$ system of nonlinear equations for a potentially solvable network $\geqslant$ the number of branches $B_k$ in the key subgraph and which exclude the $B_f$ "free" branches of the network not found in this subgraph $G_k$. On the other hand, the subset of equations $\underline{C}^f$ is equal in number to the non-key subgraph branches $B_f$ in the network. In particular there is one equation for each of these branches. Furthermore, these equations can always be written so as to emphasize the simple relationship of the excluded branch to the branches in the key subgraph. Consequently, explicit solution of the $\underline{C}^k$ subset implies solution of the $\underline{C}$ system of equations which in turn implies explicit solution of the element values of the network itself.

A useful restatement of Theorem 5 is that: the set of transfer admittances, $\underline{C}^k$, remaining in the $[Y^A]'$ and $[Y^P]'$ matrices,after excluding therefrom elements derived from all of the diagonal entries and the nonzero off-diagonal entries of the accessible portion of the $\mathcal{M}$ matrix, must be solvable for the network itself to be solvable.

<u>Corollary</u>: Solution of the system of equations for a maximal network does not entail the use of redundant equations.

<u>Proof</u>: A maximal network is defined as one in which $B_q = B$. For this case the subset $\underline{C}^k$ is equal to $B_k$ in number.

## Examples

### Detailed Example

The following example illustrates the key subgraph and related concepts. In particular the complete set of equations, derived from the compound matrix and the generalized reduction formulas, is divided into two mutually exclusive subsets $\underline{C}^k$ and $\underline{C}^f$.

Consider the twelve branch network ($B_q=B=12$, $B_k=9$, $B_c=1$) of Fig. 3 with all three types of nodes. The indefinite admittance matrix $\mathcal{M}$ for this network is, with zero entries omitted:

$$\begin{array}{cc|ccc|ccc|cc|}
 & & & \text{A} & & & \text{P} & & \text{I} & \\
\text{Nodes} & & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\hline
 & 1 & b+x+y & -y & & -x & & & & -b \\
\text{A} & 2 & -y & g+h+y+z & -z & & & & -h & -g \\
 & 3 & & -z & d+z & & & & -d & \\
\hline
 & 4 & -x & & & a+x & & & & -a \\
\text{P} & 5 & & & & & j+k & & -k & -j \\
 & 6 & & & & & & n & -n & \\
\hline
\text{I} & 7 & & -h & -d & & -k & -n & d+h+k+n+i & -i \\
 & 8 & -b & -g & & -a & -j & & -i & a+b+g+j+i \\
\hline
\end{array}$$

The key steps in obtaining the equations representing admittance measurements on this network from the $\mathcal{M}$ matrix are as follows:

$$\left(Y^I\right)^{-1} = \frac{1}{\Delta}\begin{bmatrix} a+b+g+j+i & i \\ i & d+h+k+n+i \end{bmatrix} \qquad (29)$$

where the determinant for the key subgraph is

$$\Delta = (a+b+g+j)(d+h+k+n)+i(a+b+g+j+d+h+k+n). \quad (30)$$

Let $A = a+b+g+j+i$ and $D = d+h+k+n+i$.

Then

$$Y_c = \frac{1}{\Delta}\begin{bmatrix} bi & bD \\ gi+hA & hi+gD \\ dA & di \end{bmatrix}$$

Then

$$\tilde{Y} = \frac{1}{\Delta}\begin{bmatrix} b^2D & bih+bgD & bid \\ bih+bgD & h^2A+2ghi+g^2D & dgi+dhA \\ bid & dgi+dhA & d^2A \end{bmatrix}$$

and

$$\tilde{\tilde{Y}} = \frac{1}{\Delta}\begin{bmatrix} abD & biK+bjD & bin \\ ahi+agD & hkA+gik+hij+gjD & hnA+gin \\ adi & dkA+dij & dnA \end{bmatrix}.$$

Then we finally get

$$\left(Y^A\right)' = \frac{-1}{\Delta}\begin{bmatrix} -\Delta(b+x+y)+b^2D & y\Delta+bih+bgD & bid \\ y\Delta+bih+bgD & -\Delta(g+h+y+z)+h^2A+2ghi+g^2D & z\Delta+dhA+dgi \\ bid & z\Delta+dhA+dgi & -\Delta(d+z)+d^2A \end{bmatrix} \qquad (31)$$

and

$$\left(Y^P\right)' = \frac{-1}{\Delta}\begin{bmatrix} x\Delta+abD & bik+bjD & bin \\ ahi+agD & hkA+gik+hij+gjD & hnA+gin \\ aid & dkA+dij & dnA \end{bmatrix}. \qquad (32)$$

Then the $\underline{C}^k$ subset of the system of equations for this network is:

$$\begin{aligned}
\underline{C}_2 &= b'i'd' \\
\underline{C}_6 &= b'i'n' \\
\underline{C}_{10} &= a'i'd' \\
\underline{C}_{12} &= d'n'A' \\
\underline{C}_5 &= b'i'k'+b'j'D' \\
\underline{C}_7 &= a'h'i'+a'g'D' \\
\underline{C}_9 &= h'n'A'+g'i'n' \\
\underline{C}_{11} &= d'k'A'+d'i'j' \\
\underline{C}_8 &= h'k'A'+g'i'k'+h'i'j'+g'j'D'
\end{aligned} \qquad (33)$$

where

$$\begin{aligned}
A' &= a'+b'+g'+j'+i' = b'\alpha+i' \\
D' &= d'+h'+k'+n'+i'' = d'\beta+i'.
\end{aligned}$$

The $\underline{C}^f$ subset of the system of equations is:

$$\begin{aligned}
\underline{C}_1 &= y'\Delta'+b'i'h'+b'g'D' = y'\Delta'+b'\underline{C}_7/a' \\
\underline{C}_3 &= z'\Delta'+d'h'A'+d'g'i' = z'\Delta'+d'\underline{C}_9/n' \\
\underline{C}_4 &= x'\Delta'+a'b'D' = x'\Delta'+(\underline{C}_5-\eta_5\underline{C}_6)/\eta_4.
\end{aligned} \qquad (34)$$

After appropriate manipulation of the system of equations for this network we find that:

$$d' = \left( \frac{\dfrac{\gamma^2 \underline{C}_2(\underline{C}_5-\eta_5\underline{C}_6)}{\eta_4\underline{C}_{10}} - \dfrac{\underline{C}_2\underline{C}_{12}}{\underline{C}_6}}{\beta - \dfrac{\alpha}{\gamma}} \right)^{\frac{1}{3}} \qquad (35)$$

where

$$\alpha = 1+\underline{C}_{10}/\underline{C}_2+\eta_1+\eta_4\underline{C}_{10}/\underline{C}_2$$

$$\beta = 1+\eta_3+\underline{C}_6/\underline{C}_2+\eta_5\underline{C}_6/\underline{C}_2$$

$$\gamma = \frac{\alpha+\beta\underline{C}_{12}/\underline{C}_6}{\dfrac{\alpha(\underline{C}_5-\eta_5\underline{C}_6)}{\eta_4\underline{C}_{10}} + \beta} = \frac{d'}{b'}$$

and

$$\eta_1 = \frac{\underline{C}_9\underline{C}_{11}-\underline{C}_8\underline{C}_{12}}{\underline{C}_6\underline{C}_{11}-\underline{C}_5\underline{C}_{12}} = \frac{g'}{b'}$$

$$\eta_3 = \frac{\underline{C}_6\underline{C}_8 - \underline{C}_5\underline{C}_9}{\underline{C}_6\underline{C}_{11}-\underline{C}_5\underline{C}_{12}} = \frac{h'}{d'}$$

$$\eta_4 = \frac{\underline{C}_9\underline{C}_{11}-\underline{C}_8\underline{C}_{12}}{\eta_1\underline{C}_6\underline{C}_{10}-(\underline{C}_7-\eta_3\underline{C}_{10})\underline{C}_{12}} = \frac{j'}{a'}$$

$$\eta_5 = \frac{\underline{C}_{11}-\eta_4\underline{C}_{10}}{\underline{C}_{12}} = \frac{k'}{n'}$$

Then it follows that,

$$\begin{aligned} b' &= d'/\gamma \\ n' &= d'\underline{C}_6/\underline{C}_2 \\ h' &= d'\eta_3 \\ k' &= n'\eta_5 \\ a' &= b'\underline{C}_{10}/\underline{C}_2 \qquad (36) \\ j' &= a'\eta_4 \\ g' &= b'\eta_1 \\ i' &= \underline{C}_6/b'n'. \end{aligned}$$

These nine elements of the key subgraph permit one to evaluate the scale factor $\lambda$ for the individual elements. Actual admittance measurements used in the solution equations yield the primed values. The primed values are proportional to the desired original element values by the scale factor. The equation defining the determinant of the key subgraph yields the scale factor $\lambda$ when the primed element values are substituted therein. Subsequently, the unprimed element values are used in the same equation to yield the value of the determinant itself. With this we can solve for the three non-key-subgraph elements x, y, and z.

$$\begin{aligned} x &= \underline{C}_4-(\underline{C}_5-\eta_5\underline{C}_6)/\eta_4 \\ y &= \underline{C}_1-b\underline{C}_7/a \qquad (37) \\ z &= \underline{C}_3-d\underline{C}_9/n\ . \end{aligned}$$

Numerical Results

In Fig. 3 let the element values in mhos be:

a = b = g = j = 1

d = h = k = n = 2

i = 3, x = 6, y = z = 4.

$\therefore$ A = 7, D = 11, $\Delta$ = 68.

Then the computed results are as tabulated below:

Transfer Measurements

| Notation | | Measured Value | | |
|---|---|---|---|---|
| Numbering Convention | Short circuit Admittance | Nominal | Assumed | Assumed Error |
| $\underline{C}_2$ | $Y_{13}$ | .088235 | .09265 | +5% |
| $\underline{C}_5$ | $Y_{15}$ | 0.25000 | 0.2625 | " |
| $\underline{C}_8$ | $Y_{25}$ | 0.75000 | 0.7875 | " |
| $\underline{C}_9$ | $Y_{26}$ | 0.50000 | 0.5250 | " |
| $\underline{C}_{10}$ | $Y_{34}$ | .088235 | .09265 | " |
| $\underline{C}_6$ | $Y_{16}$ | .088235 | .08382 | -5% |
| $\underline{C}_7$ | $Y_{24}$ | 0.25000 | .23750 | " |
| $\underline{C}_{11}$ | $Y_{35}$ | 0.50000 | .47500 | " |
| $\underline{C}_{12}$ | $Y_{36}$ | 0.41179 | .39120 | " |
| $\underline{C}_1$ | $Y_{14}$ | 4.25000 | 4.2500 | none |
| $\underline{C}_3$ | $Y_{12}$ | 4.50000 | 4.5000 | " |
| $\underline{C}_4$ | $Y_{23}$ | 6.16176 | 6.1618 | " |

The two sets of calculated transfer admittance measurements given above are used with the set of solution formulas, equations 35 through 37, for the twelve branches of this network given in the previous section. The resulting element values are listed below. Note that the first column of results is based on the ideal condition of perfect measurements and calculations. The second column is based on introduction of assumed measurement errors of ±5% for the $\underline{C}^k$ subset only.

| | Calculations | | |
|---|---|---|---|
| Network Branch | Ideal Case | Meas. Error | Deviation in % |
| a | 1.000 | 0.959 | - 4 |
| b | 1.000 | 0.959 | - 4 |
| d | 2.000 | 2.064 | + 3 |
| g | 1.000 | 0.895 | -11 |
| h | 2.000 | 2.357 | +18 |
| i | 3.000 | 3.498 | +17 |
| j | 1.000 | 1.271 | +27 |
| k | 2.000 | 1.681 | -16 |
| n | 2.000 | 1.868 | - 7 |
| x | 6.000 | 6.021 | + 0 |
| y | 4.000 | 4.013 | + 0 |
| z | 4.000 | 3.920 | - 2 |

Observe that the maximum deviation does not occur for the concealed branch i as may have been expected. This is a function of the magnitude of the errors in the specific measurements.

## Networks Related by Key Subgraph[3]

Further significance of the key subgraph of a solvable network can be illustrated. The totality of branches in the network is divided into the key subgraph and its complement. Hence, the number of branches in a given network can be written as the sum

$$B = B_k + B_f \tag{38}$$

If one can solve a sufficiently general case having a given key subgraph, one can find many related networks which can be considered as members of this family of networks. The number of networks in the family is given by

$$2^{B_f} \tag{39}$$

For purposes of illustration, we can use the network shown in Fig. 3. Here we have, $B = 12$, $B_k = 9$ and $B_f = 3$.

Consequently, using Eq. 39 we get,

| Total Branches | Number of Networks |
|---|---|
| $B_k + 3$ | 1 |
| $B_k + 2$ | 3 |
| $B_k + 1$ | 3 |
| $B_k$ | 1 |

The eight resulting networks are shown in Fig. 4. It is emphasized that the initial detailed solution of the general case (in this example $B = 12$) readily yields the remaining indicated solutions by virtue of the fact that the non-key subgraph $B_f$ branches of the network occur as individual solution equations. In order to be able to use this simple deletion procedure no special case is required in setting up the indefinite admittance matrix $\mathcal{Y}_i$ .

## Conclusions

A systematic procedure has been described for formulating the system of equations representing the observable external behavior of a network with partially accessible as well as accessible terminals. The concept of the key subgraph was introduced. It was shown that it is possible to select a subset of the system of equations which include only branches of the key subgraph as variables. These are found to be adequate for determining the element values of the network.

Details of the systematic explicit method used to solve the system of equations for the example presented will be the subject of a future paper. In addition to this explicit method of solution, attention is invited to a forthcoming paper[4] on an implicit method using maximum likelihood estimation procedures. Both of these techniques are suitable for digital computer use.

## Appendix

### List of Symbols

$B$ Actual number of elements in the network or branches in its graph.

$B_c$ Number of concealed branches in the key subgraph $G_k$. These join internal nodes.

$B_f$ Number of non-key subgraph (or "free") branches in the network or its graph. These join two accessible or an accessible and a partly accessible node.

$B_k$ Number of branches in the key subgraph $G_k$ of the network.

$B_q$ Maximum number of branches permitted by the Q test, i.e. $B_q = Q$.

$\underline{C}$ General transfer admittance measurement. Also the complete set of measurements for the external behavior of the network.

$\underline{C}^f$ The subset of measurements involving non-key subgraph as well as key subgraph branches as unknowns. There is one such measurement for each of the $B_f$ branches.

$\underline{C}^k$ The subset of measurements involving only branches of the key subgraph as the unknowns.

$G_k$ The key subgraph for the network or its corresponding linear graph.

Q Number of admittance functions which specify a passive linear bilateral network.

$\Delta$ Determinant of $Y^I$ submatrix of the indefinite admittance matrix, $\mathcal{M}$ , expressed as a compound matrix.

$\lambda$ Scale factor related to the determinant of the key subgraph.

## References

1. R. S. Berkowitz, "Conditions for Network Element Value Solvability," IRE Trans. on Circuit Theory, Vol. 9, Mar. 1962.

2. G. Kron, Tensor Analysis of Networks, J. Wiley and Sons, New York, 1939, Chapt. 10.

3. S. D. Bedrosian, "On Element Value Solution of Single-Element-Kind Networks," Ph.D. Dissertation, Univ. of Pennsylvania, 1961.

4. R. S. Berkowitz, "Statistical Considerations in Network Element Value Solution," to appear in IRE Trans. on Military Electronics, 1962.

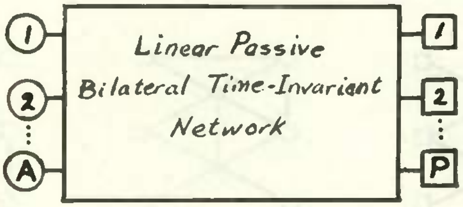


Fig. 1. General network.

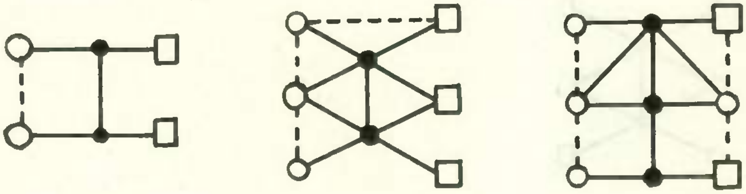


Fig. 2. Networks illustrating key subgraph.

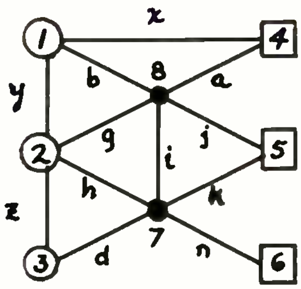


Fig. 3. Network for detailed example.

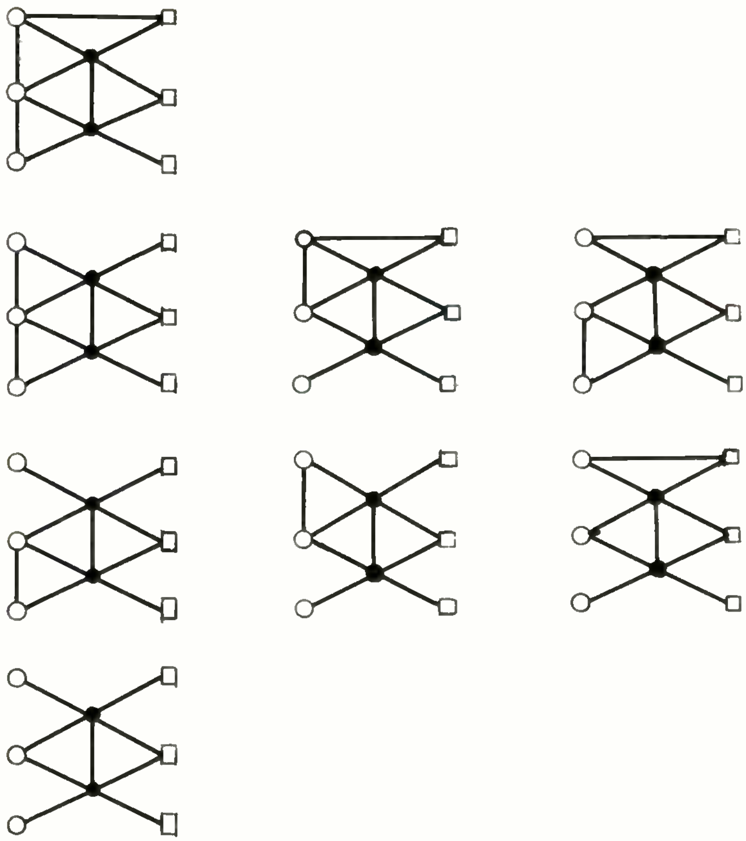


Fig. 4. Family of solvable networks related by a common key subgraph.

# THE IMAGE-PARAMETER DESIGN OF THE GENERAL TWO-SECTION ELLIPTIC-FUNCTION FILTER

W. N. Tuttle
General Radio Company
West Concord, Mass.

## Summary

Elliptic-function filters, which are optimum designs for many practical requirements, have found limited use because of the difficulties of design. It is shown that electrical symmetry requires that two-section filters be matched at the internal junction on an image-impedance basis. For the range of designs which can be realized in the ladder structure, therefore, the general fifth-order elliptic-function filter is a pure Zobel design with symmetrically modified end reactances. This fact provides the basis for simplified design formulas for filters with any desired ripple level and any separation between the pass and stop bands.

## Introduction

A previous paper [1] showed that a limited class of two-section elliptic-function filters can be realized as pure Zobel designs by proper choice of the m values and the resistance termination ratio. This class of filters is restricted in that only a single level of pass-band ripple is available for each width of the cutoff region. It is frequently desirable to obtain much lower ripple levels when a low reflection coefficient is needed, or much higher ripple levels when maximum discrimination in the stop band is the prime specification. When the full range of elliptic-function designs is available, two-section filters are versatile enough to satisfy a large variety of practical requirements. However, no simple design procedure has been available to handle the general case, and inferior designs are still used extensively because of the cost and difficulty of obtaining an optimum design.

The present paper describes an extension of the previous results to the general case, and is based on the fact, which has not previously been reported, that the larger class of filters, although not purely Zobelian, must nevertheless have their sections matched at the internal junction on an image-impedance basis. This requirement, which is necessary for electrical symmetry, applies to fifth-order Butterworth and Tchebycheff filters as well as to elliptic-function filters. All three types of two-section filters, therefore, can be realized as pure Zobel or prototype filters with the end reactances symmetrically modified. The quantity determining the end reactance modification constitutes an additional design constant which permits these more general types of filters to be realized by image-parameter methods. The design formulas are considerably simpler than those of the usual synthesis procedures and lend themselves readily to routine filter design with a desk calculator, or, using auxiliary charts, with a slide-rule.

It should be emphasized that in the case of three or more sections matched junctions are not necessary for electrical symmetry. In these cases successive mismatches can compensate one another so that over-all electrical symmetry is possible without matching. The method of the present paper, therefore, can not be directly extended to three or more sections.

## Proof that the Sections Must be Matched

The proof will be given for the circuit shown in Fig. 1, which is the usual configuration for the two-section low-pass filter, but the proof is similar for the corresponding case of mid-series terminations. It is assumed that this filter is electrically symmetrical, as is the two-section elliptic-function filter. The first step is to split off equal amounts from the two end capacitances so that the sum of the remainders is equal to the center capacitance. Then the center part can be divided into two sections each having equal end capacitances. The original filter has then been broken down as shown in Fig. 2.

The equal end capacitances g can be either positive or negative, so that the capacitances $C_A$ and $C_B$ of the component sections can be less or greater, respectively, than the end capacitances of the original filter.

In a Zobel filter section, the three image parameters $R_1$, $\omega_c$ and m determine the three element values, $C_A$, $L_2$ and $C_2$. The usual equations can be inverted and the elements of the general symmetrical filter of Fig. 2 can be described in terms of the image parameters. For the first section the values are

$$\omega_c^2 = \frac{1}{L_2\ (C_2 + \frac{C_A}{2}\ )} \tag{1}$$

$$R^2 = \frac{L_2}{2C_A} \tag{2}$$

$$m_1^2 = \frac{1}{1 + \frac{2C_2}{C_A}} \qquad (3)$$

So far this is a purely formal description of the original filter. This was assumed to be electrically symmetrical, however, and since the capacitances g which have been removed from the ends are equal, the remainder in the middle must still be symmetrical. The central portion, then, is electrically symmetrical and consists of two Zobel sections in tandem. It remains to show that this is possible only if the sections are matched.

The last point can be proved from the general symmetry requirement that the open- and short-circuit impedances must be the same from both ends. If the branch impedances of Fig. 1 are replaced by $z_1$ $z_2$ $z_3$ $z_4$ and $z_5$, the open- or short-circuited impedances can readily be written down and equated. For either case the symmetry requirement can be shown to be,

$$z_4 - z_2 = (z_2z_3 + z_3z_4 + z_2z_4)(\frac{z_1 - z_5}{z_1z_5}) \qquad (4)$$

The reactance values of the arms of the Zobel sections can be computed from the usual equations and put in the form

$$x_1 = \frac{-R_1\omega_{c1}}{m_1\omega} \qquad (5)$$

$$x_2 = \frac{2R_1\omega_{c1}\omega}{\omega_{c1}^2 - (1-m_1^2)\omega^2} \qquad (6)$$

$$x_3 = \frac{-R_1R_2\omega_{c1}\omega_{c2}}{\omega(R_1\omega_{c1}m_2 + R_2\omega_{c2}m_1)} \qquad (7)$$

$$x_4 = \frac{2R_2\omega_{c2}\omega}{\omega_{c2}^2 - (1-m_2^2)\omega^2} \qquad (8)$$

$$x_5 = \frac{-R_2\omega_{c2}}{m_2\omega} \qquad (9)$$

Equation (4) gives

$$(x_4 - x_2)x_1x_5 = (x_2x_3 + x_3x_4 + x_2x_4)(x_1 - x_5) \quad (10)$$

and the reactance values can be substituted from (5) to (9). Since the equation must hold for all frequencies, the terms in the various powers of $\omega$ can be equated to determine the conditions on the image parameters necessary for electrical symmetry. It is found that after cancellation there are only constant terms and terms in $\omega^2$. Equating the $\omega^2$ terms and simplifying the resulting expression finally yields

$$R_1^2\omega_{c1}^2 = R_2^2\omega_{c2}^2 \qquad (11)$$

Equating the constant terms gives

$$R_1^2 = R_2^2$$

Hence from (11)

$$\omega_{c1}^2 = \omega_{c2}^2$$

It has thus been shown that the two Zobel sections in the central portion of Fig. 2 must be matched at their junction on an image-impedance basis. Hence the whole filter, which is the general symmetrical two-section filter including all the elliptic-function designs realizable in the ladder configuration, is either a pure Zobel filter or a Zobel filter with symmetrically modified end reactances.

## Derivation of the Design Equations

The insertion loss of an electrically symmetrical reactance filter can be expressed in terms of the arms of the equivalent lattice by the formula

$$L = 10 \log_{10} \left[1 + (\frac{1 + uv}{u - v})^2\right] \qquad (14)$$

where the loss is in decibels and $u = X_x/R_t$ and $v = X_y/R_t$ are the ratios of the lattice-arm reactances to the terminating resistances. For the pure Zobel filter with the end reactances unmodified Saraga[2] gives the formulas

$$u = -r \frac{m_1 + m_2}{1 + m_1m_2} \quad \frac{x}{x^2 - \frac{1}{1 + m_1m_2}} \qquad (15)$$

$$v = -r \frac{1 + m_1m_2}{m_1 + m_2} \quad \frac{x^2 - \frac{1}{1 + m_1m_2}}{x\,(x^2 - 1)} \qquad (16)$$

where r is the ratio of the design resistance to the terminating resistance, x the ratio of the frequency to the image-parameter cutoff frequency, and $m_1$ and $m_2$ are the m values of the two Zobel sections.

The additional end capacitances g can be included by taking

$$\frac{1}{u'} = \frac{1}{u} - gx \qquad (17)$$

$$\frac{1}{v'} = \frac{1}{v} - gx \qquad (18)$$

and the value of the characteristic function becomes

$$E = \frac{1 + u'v'}{u' - v'} = \frac{1 - gx(u + v) + (1+g^2x^2)uv}{u - v} \qquad (19)$$

The reactance values for the two-section Zobel filter can be taken from (15) and (16) giving

$$E = \frac{\frac{1}{r} + gx\left[\frac{a}{b}\,\frac{x}{x^2 - \frac{1}{b}} + \frac{b}{a}\,\frac{x^2 - \frac{1}{b}}{x(x^2 - 1)}\right] + (1+g^2x^2)\frac{r}{x^2-1}}{\frac{b}{a}\,\frac{x^2 - \frac{1}{b}}{x(x^2 - 1)} - \frac{a}{b}\,\frac{x}{x^2 - \frac{1}{b}}} \qquad (20)$$

where a and b are abbreviations for $m_1 + m_2$ and $1 + m_1m_2$, respectively.

If the numerator and denominator are multiplied by

$$abx\,(x^2 - 1)(x^2 - \frac{1}{b})$$

the denominator becomes

$$b^2(x^2 - \frac{1}{b})^2 - a^2x^2(x^2 - 1)$$

which reduces to

$$(1 - m_1^2)(1 - m_2^2)(x^2 - \frac{1}{1 - m_1^2})(x^2 - \frac{1}{1 - m_1^2})$$

Finally, collecting the terms of the numerator by powers of x, the characteristic function is obtained as

$$E = \frac{-abx\,(Ax^4 - Bx^2 + C)}{r(1-m_1^2)(1-m_2^2)(x^2 - \frac{1}{1 - m_1^2})(x^2 - \frac{1}{1 - m_2^2})} \qquad (21)$$

where

$$A = 1 + r^2g^2 + rg\,(\frac{a}{b} + \frac{b}{a}) \qquad (22)$$

$$B = \frac{1 + r^2g^2}{b} + rg\,(\frac{a}{b} + \frac{2}{a}) + 1 - r^2 \qquad (23)$$

$$C = \frac{rg}{ab} + \frac{1 - r^2}{b} \qquad (24)$$

In (21) the insertion-loss poles appear in the denominator as the familiar rejection peaks of the m-derived sections, and the finite zeros in the numerator as the roots of the biquadratic expression.

Equation (21) gives the characteristic function for two matched Zobel m-derived sections with the end reactances modified in accordance with the new parameter g. To realize the elliptic-function response, giving equal pass-band ripples and equal stop-band valleys of insertion loss, the procedure of the earlier paper [1] can be followed, and the poles and zeros located according to the expression,

$$E = \frac{Hy(y^2 - a_2^2)(y^2 - a_4^2)}{a_2^2a_4^2\,(y^2 - \frac{1}{a_4^2})(y^2 - \frac{1}{a_2^2})} \qquad (25)$$

where $a_2$ and $a_4$ are the Cauer parameters and the frequency variable y is with respect to $\sqrt{f_1f_2}$, the geometric mean of $f_1$, the end frequency of the pass band, and $f_2$ the beginning frequency of the stop band. It will be noted that the poles and zeros are symmetrically located about the mean frequency of the cutoff region, so that their distribution is specified by only the two constants, $a_2$ and $a_4$. These are defined in the usual terminology of elliptic functions, as follows:

$$a_2 = \sqrt{k}\ \operatorname{sn}(\frac{2\bar{K}}{5}, k) \qquad (26)$$

$$a_4 = \sqrt{k}\ \operatorname{sn}(\frac{4\bar{K}}{5}, k) \qquad (27)$$

where $k = \sin f_1/f_2 = \sin\theta$.
The Cauer parameters have been tabulated by Glowatzki [3] or can be obtained conveniently from tables of elliptic functions [4]. A short table at 5° intervals of θ is given as Table I. The entire design of the elliptic-function filter, can be carried through in terms of image parameters and the additional quantity g if the constants A, B and C of (22)-(24) can be chosen to give the pole and zero locations specified by (25), and if the constant H can be given the value corresponding to the desired ripple level.

The frequency variable y in (25) is related to the corresponding quantity x in the image-parameter expression (21) by

$$y = x\,f_{co} \qquad (28)$$

where $f_{co} = f_c / \sqrt{f_1f_2}$. This change of variable in (25) gives

$$E = \frac{Hx\,f_{co}(x^2 - \frac{a_2^2}{f_{co}^2})(x^2 - \frac{a_4^2}{f_{co}^2})}{a_2^2\,a_4^2\,(x^2 - \frac{1}{f_{co}^2\,a_4^2})(x^2 - \frac{1}{f_{co}^2\,a_2^2})} \qquad (29)$$

Comparing (29) with (21) it is evident that

$$\frac{H\,f_{co}}{a_2^2\,a_4^2} = \frac{abA}{r(1 - m_1^2)(1 - m_2^2)}$$

and since $1 - m_1^2 = f_{co}^2\,a_4^2$ and $1 - m_2^2 = f_{co}^2\,a_2^2$ the constant H is given by

$$H = \frac{Aab}{rf_{co}^5} \qquad (30)$$

If the roots of the biquadratic expression in the numerator of (21) are equated to the values specified in (29), the following conditions are obtained on the coefficients A, B and C.

$$B = Ac \qquad (31)$$

$$C = Ad \qquad (32)$$

$$\text{where} \quad c = \frac{a_2^2 + a_4^2}{f_{co}^2} \quad \text{and} \quad d = \frac{a_2^2 \, a_4^2}{f_{co}^4}$$

Taking the values of the coefficients from (22)(23) and (24) conditions (31) and (32) become respectively

$$(1 + r^2g^2)(\frac{1}{b} - c) + rg\left[\frac{a}{b} + \frac{2}{a} - c(\frac{a}{b} + \frac{b}{a})\right] + 1 - r^2 = 0 \qquad (33)$$

and

$$(1 + r^2g^2)bd + rg \left[bd(\frac{a}{b} + \frac{b}{a}) - \frac{1}{a}\right] - (1 - r^2) = 0 \qquad (34)$$

The sum of (33) and (34) gives the quadratic in the product rg:-

$$(1 + r^2g^2)(c - bd - \frac{1}{b}) - rg\left[(\frac{a}{b} + \frac{1}{a}) - (c-bd)(\frac{a}{b} + \frac{b}{a})\right] = 0 \qquad (35)$$

This can be put in the form,

$$r^2g^2 - 2Frg + 1 = 0 \qquad (36)$$

where

$$F = \frac{\frac{a}{b} + \frac{1}{a} - (c - bd)(\frac{a}{b} + \frac{b}{a})}{2\,(c - bd - \frac{1}{b})} \qquad (37)$$

The solution is

$$rg = F(1 - \sqrt{1 - \frac{1}{F^2}}). \qquad (38)$$

For large values of F the solution is conveniently obtained from the series

$$rg = \frac{1}{2F} + \frac{1}{8F^3} + \frac{1}{16F^5} + \cdots \qquad (39)$$

and for F greater than 9 the first two terms give the result within about $10^{-6}$. When rg has been obtained, substitution in (33) gives r and hence also g. The product rg can be substituted in (22) to yield A, which, with r, gives the constant H from (30). The constant H with a third elliptic-function parameter, $\Delta$, determines the maximum pass-band ripple $\alpha_p$ and the stop-band valley height from the expressions,

$$\alpha_p = 10 \log_{10}(1 + H^2 \Delta^2) \qquad (40)$$

$$\alpha_a = 10 \log_{10}(1 + H^2/\Delta^2) \qquad (41)$$

The quantity $\Delta$ is defined as

$$\Delta = k^{5/2} sn^2 (\frac{K}{5}, k) \, sn^2 (\frac{3K}{5}, k) \qquad (42)$$

Values of $\Delta$ are included in Glowatzki's tables [3] and in Table I.

In the procedure above outlined it will be noted that $f_1$ and $f_2$ determine $\theta$ and hence the Cauer parameters $a_2$, $a_4$ and $\Delta$. These quantities with $f_{co}$ determine the entire filter design. If $f_{co}$ is fixed with respect to $f_1$ and $f_2$, then only one elliptic function filter with a particular ripple level $\alpha_p$ is possible. The quantity $f_{co}$, which is usually thought of as the end of the pass band, actually is an additional parameter which must be varied in its location between $f_1$ and $f_2$ to obtain the desired ripple level or value of the constant H. It has not been found feasible to compute the required value of $f_{co}$ directly from H or from $\alpha_p$. But by computing for each of a series of values of $\theta$ a group of filter designs with different values of $f_{co}$ a chart can be prepared giving the required value of $f_{co}$ for any desired ripple level and any separation between the pass and stop bands.

Such a chart is shown in Fig. 3 covering ripple levels down to .001 db, corresponding to 1.52 per cent reflection factor. The dashed line at the upper right shows the limit of designs realizable in the ladder structure. The limit corresponds to one of the end capacitances going negative, that is, to the negative additional capacitance g being equal to the unmodified end capacitance of the section with the lower value of m. Designs at the limit can be used and require one less element in the high- or low-pass case and two less elements in the band-pass case. They are not usually of interest because the minimum stop band loss is less than 20 db. The design with .001 db ripple, for example, gives 19.5 db minimum loss in a frequency ratio of 1.4 to 1.

The need for using a chart to determine $f_{co}$ at the start of the design does not imply an approximation in the computation as will be seen in the following section. An error in selecting $f_{co}$ will cause the pass-band ripple to depart more or less from the desired value but the filter will still be a perfect elliptic-function design. The chart shows that the variation of $f_{co}$ is considerable in going from one ripple level to another so that it is easy to meet practical specifications on ripple level.

## Design Procedure

A performance chart for two-section filters is generally preferable to formulas for arriving at the best design compromise to meet given specifications. The chart of Fig. 4 gives the frequency ratio as a function of the minimum stop-band loss for various constant pass-band ripple levels. The chart shows, at the left, the limit of ladder realizibility, discussed above, and, by the dashed line, the designs realizable as pure Zobel filters by the method of the previous paper [1]. The ripple level of the Zobel design is generally between 0.02 and 0.1 db, which is in the middle of the useful range. When ripple levels of this order are satisfactory, advantage can be taken of the simpler design procedure, the filters being identical.

The general design procedure will be illustrated by a numerical example. Assume that a low-pass filter is needed to work between 500-ohm resistances, pass frequencies up to 1 kc and provide at least 50 db insertion loss at 1.7 kc and beyond. A ripple level of about 0.3 db is acceptable. Fig. 4 shows that 50 db can be obtained in a frequency ratio 1.48 for a ripple level of 0.5 db or 1.59 for 0.2 db, so the specifications are within the capabilities of a two-section filter. Taking $\theta = 40°$ gives a frequency ratio of 1.56, between the above limits, and permits the use of the Cauer parameters directly from Table I. For this angle the $f_{co}$ chart, Fig. 3, shows that $f_{co} = 0.90$ gives a ripple level of about 0.3 db, and will be taken for the design.

For $\theta = 40°$ Table I gives

$$a_2 = .5116709$$

$$a_4 = .7713694$$

Then $m_1^2 = 1 - a_4^2 f_{co}^2 \qquad m_1 = .7197509$

$m_2^2 = 1 - a_2^2 f_{co}^2 \qquad m_2 = .8876577$

$$\begin{aligned} a &= 1.607409 \\ b &= 1.638892 \\ c &= 1.057800 \\ d &= .2374303 \end{aligned}$$

From (37) and (38)

$$\begin{aligned} F &= 2.267231 \\ rg &= .232449 \end{aligned}$$

Hence from (33)

$$\begin{aligned} r &= .743994 \\ g &= .312434 \end{aligned}$$

The values of r and g are the two final quantities required to complete the design of the filter, but it is useful as a check to compute H and from H the values of $\alpha_p$ and $\alpha_a$ to compare with the specification limits.

From (30) and Table I

$$\begin{aligned} H &= 9.108732 \\ \Delta &= .02841309 \end{aligned}$$

and these values in (40) and (41) give

$$\begin{aligned} \alpha_p &= .281567 \text{ db} \\ \alpha_a &= 50.1188 \text{ db,} \end{aligned}$$

which satisfy the specification limits.

The quantities $f_{co}$, $m_1$, $m_2$, r and g, which have been determined, fix the design of the filter normalized for an image-parameter cutoff frequency of 1 radian per second and for terminating resistances of 1 ohm. The element values are obtained from the usual Zobel design equations modified by the addition of the quantity g to the end capacitances. These are

$$\begin{aligned} C_1 &= \frac{m_1}{r} + g = 1.27985 \text{ f} \\ L_2 &= 2m_1 r = 1.07098 \text{ h} \\ C_2 &= \frac{1 - m_1^2}{2m_1 r} = 0.45002 \text{ f} \\ C_3 &= \frac{m_1 + m_2}{r} = 2.16051 \text{ f} \\ L_4 &= 2m_2 r = 1.32082 \text{ h} \\ C_4 &= \frac{1 - m_2^2}{2m_2 r} = .16055 \text{ f} \\ C_5 &= \frac{m_2}{r} + g = 1.50553 \text{ f} \end{aligned}$$

To obtain the final design from the normalized design it is necessary to locate the end of the pass band with respect to the highest frequency to be transmitted, in this case 1 kc. If $f_1$ is taken as 1.07 kc then $f_2 = f_1 \csc \theta = 1.665$ kc, which satisfies the specification limit of 1.7 kc. It is desirable to place $f_1$ slightly outside the desired pass band because dissipation effects are most pronounced near the end of the band. It must be emphasized that the performance figures which have been given are for ideal filters with no dissipation in the elements.

Since $f_c^2/f_1 f_2 = f_{co}^2$ and $f_1/f_2 = \sin\theta$ it follows that

$$f_c = f_1 f_{co} / \sqrt{\sin\theta}$$

The quantity $\sqrt{\sin\theta}$ is the ratio of the end of the pass band to the mean frequency of the cutoff region, and is included in the tabulated Cauer parameters as $a_5$. In this case the value is obtained from Table I as .8017404.

Hence $f_c = 1.07 \times 0.9 \div .8017404 = 1.20114$ kc.

The conversion factors are therefore,

$$C_o = \frac{1}{R_t 2\pi f_c} = 0.26501 \ \mu\text{f}$$

$$L_o = \frac{R_t}{2\pi f_c} = .066252 \text{ h}$$

and the final element values are

$$\begin{aligned} C_1 &= 0.3392 \ \mu\text{f} \\ L_2 &= 0.07095 \text{ h} \\ C_2 &= 0.1193 \ \mu\text{f} \\ C_3 &= 0.5726 \ \mu\text{f} \\ L_4 &= 0.08754 \text{ h} \\ C_4 &= 0.04255 \ \mu\text{f} \\ C_5 &= 0.3990 \ \mu\text{f.} \end{aligned}$$

The computation as above outlined is very simple except for the steps of determining the coefficients of equations (33) and (35) and solving for r and g. These steps can be eliminated by using additional charts. Figs. 5 and 6 give r and g, respectively, as functions of $\theta$ for the same constant ripple levels used in the performance chart, Fig. 4, and the $f_{co}$ chart, Fig. 3. To avoid interpolation error one of these ripple levels should be selected if possible. With these charts filter designs accurate enough for most purposes can be obtained by slide-rule computation. The steps are the same as above outlined as far as the determination of $m_1$ and $m_2$. At this point the charts of Figs. 5 and 6 are entered to obtain r and g and the element values then obtained from the modified Zobel design equations as before.

## Discussion

Examination of the $f_{co}$, r and g charts shows how the ripple level is varied in filters with the same width of the cutoff region. Fig. 6 shows that the filters with large ripple are those with the end capacitances substantially increased over those of the pure Zobel design. This decreases the damping effect of the terminating resistances, giving stronger reactive control and hence increased ripple and increased rejection. Also Fig. 5 shows that these filters, at least for small $\theta$,have a lower design resistance relative to the terminating resistances, further reducing the damping. The filters with low ripple have the end capacitance reduced below the Zobel value, and, as pointed out in the body of the paper, cease to be realizable when one of the required end capacitances becomes negative.

Fig. 3 shows that in the filters with large ripple the image-parameter cutoff frequency is close to the end of the pass band but that in the flat filters it is far removed. In the filter for $\theta = 45°$ and $\alpha_p = .001$ db, for example, the cutoff frequency is 14 per cent beyond the middle of the cutoff region. In the sharp-cutoff filters, those with large $\theta$, the critical frequencies are so close together that little variation is possible and $f_{co}$ is always beyond the middle point.

## References

(1) W. N. Tuttle, "The design of two-section symmetrical Zobel filters for Tchebycheff insertion loss", Proc. IRE vol. 47, pp. 29-36, Jan. 1959.

(2) W. Saraga, "Insertion loss and insertion phase shift of multi-section Zobel filters with equal image impedances", P.O. Elect. Eng. J., vol. 39, pp. 167-172; January, 1947.

(3) E. Glowatzki, "Sechsstellige Tafel der Cauer-Parameter", Abhandlungen der Bayerischen Akademie der Wissenschaften, Neue Folge, Heft 67; 1955.

(4) G.W. and R.M. Spenceley, "Smithsonian Elliptic Functions Tables", Washington, D.C., The Smithsonian Institution, 1947.

TABLE I CAUER PARAMETERS

| $\theta$ | $f_2/f_1$ | $a_2$ | $a_4$ | $a_5$ | $\Delta$ |
|---|---|---|---|---|---|
| 5 | 11.4737 | .1737433 | .2808234 | .2952215 | .0001408278 |
| 10 | 5.7588 | .2461629 | .3966041 | .4167111 | .0008004651 |
| 15 | 3.8637 | .3024182 | .4846348 | .5087426 | .002223661 |
| 20 | 2.9238 | .3507266 | .5578199 | .5848249 | .004617248 |
| 25 | 2.3662 | .3943513 | .6210843 | .6500910 | .008187876 |
| 30 | 2.0000 | .4350349 | .6769012 | .7071068 | .01316095 |
| 35 | 1.7434 | .4738790 | .7266982 | .7573483 | .01979737 |
| 40 | 1.5557 | .5116709 | .7713694 | .8017404 | .02841309 |
| 45 | 1.4142 | .5490344 | .8115034 | .8408964 | .03940573 |
| 50 | 1.3054 | .5865130 | .8474973 | .8752396 | .05329416 |
| 55 | 1.2208 | .6246260 | .8796190 | .9050702 | .07078198 |
| 60 | 1.1547 | .6639159 | .9080421 | .9306049 | .09286628 |
| 65 | 1.1034 | .7050027 | .9328646 | .9520020 | .1210395 |
| 70 | 1.0642 | .7486638 | .9541161 | .9693774 | .1577036 |
| 75 | 1.0353 | .7959824 | .9717525 | .9828153 | .2071458 |
| 80 | 1.0154 | .8486984 | .9856278 | .9923748 | .2783823 |
| 85 | 1.0038 | .9103780 | .9954034 | .9980955 | .3973593 |

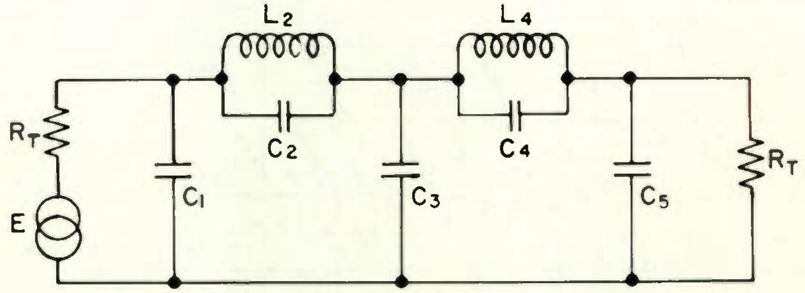


Fig. 1. Structure of two-section low-pass filter.

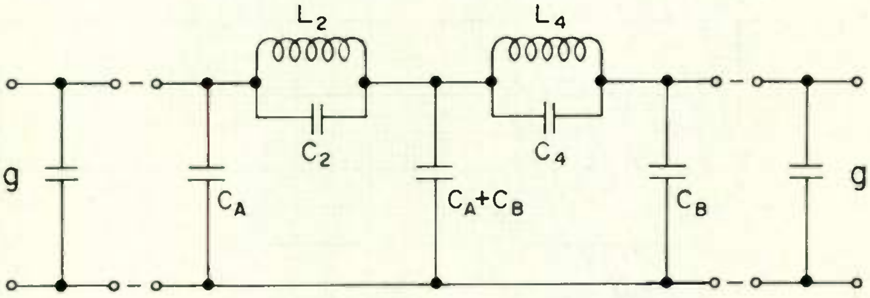


Fig. 2. Breakdown of filter into two symmetrical sections by use of supplementary end capacitances.

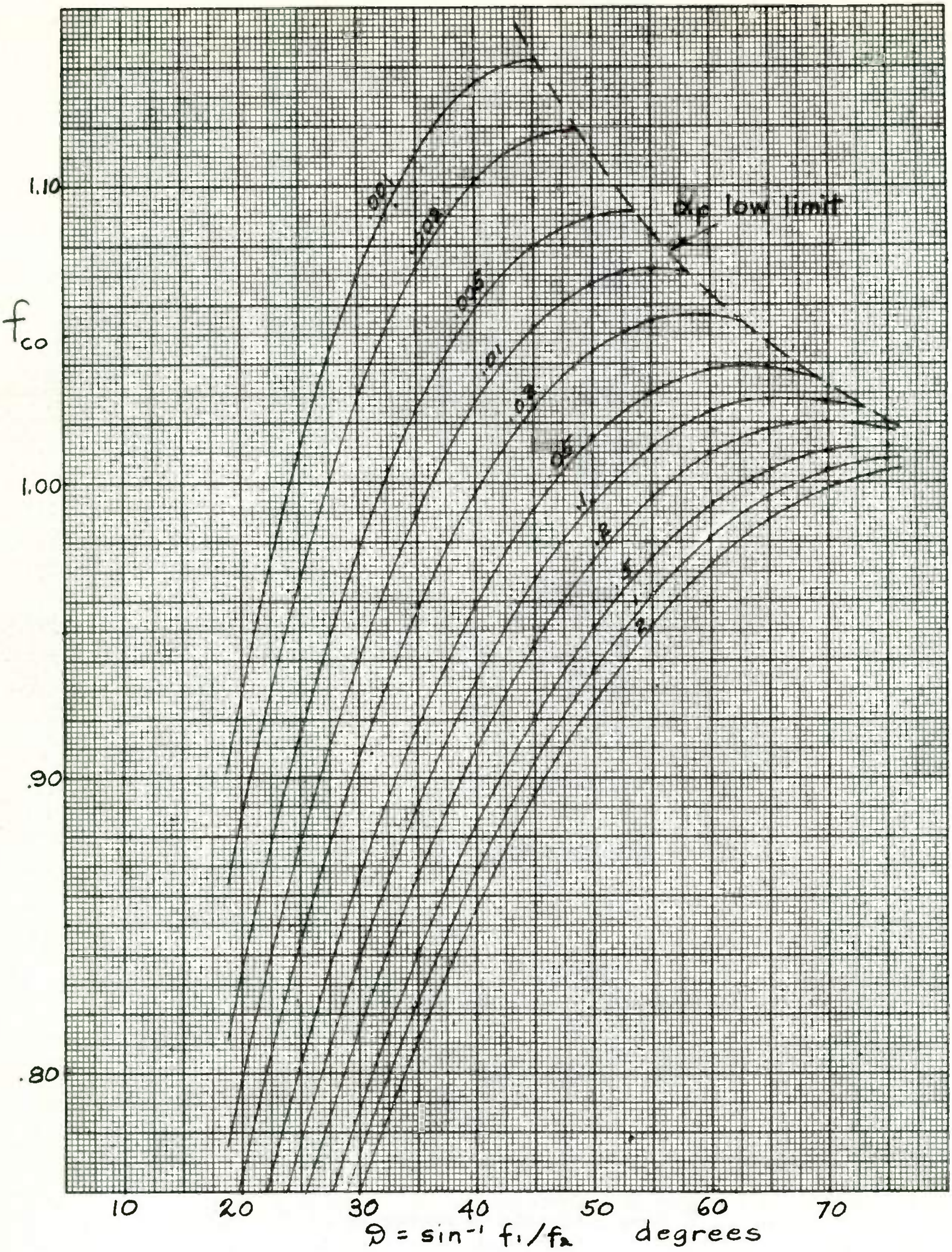


Fig. 3. Chart showing required location of the cutoff frequency for different ripple levels and different widths of the cutoff region.

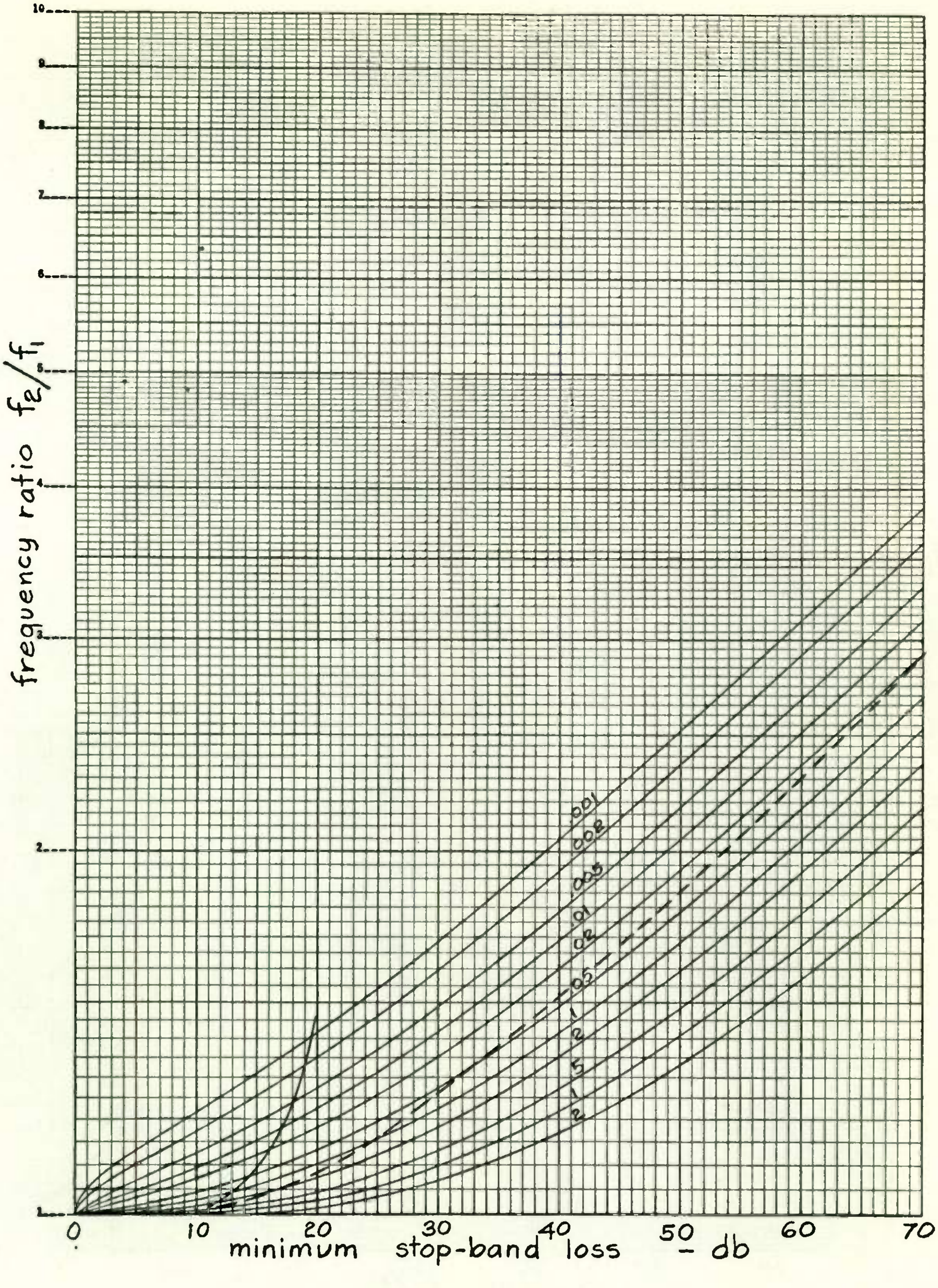


Fig. 4. Performance chart for two-section filters, giving the frequency ratio as a function of the minimum stop-band loss for various ripple levels. The limit of ladder realizability is indicated at the left and the characteristics of the optimum Zobel designs by the dashed line.

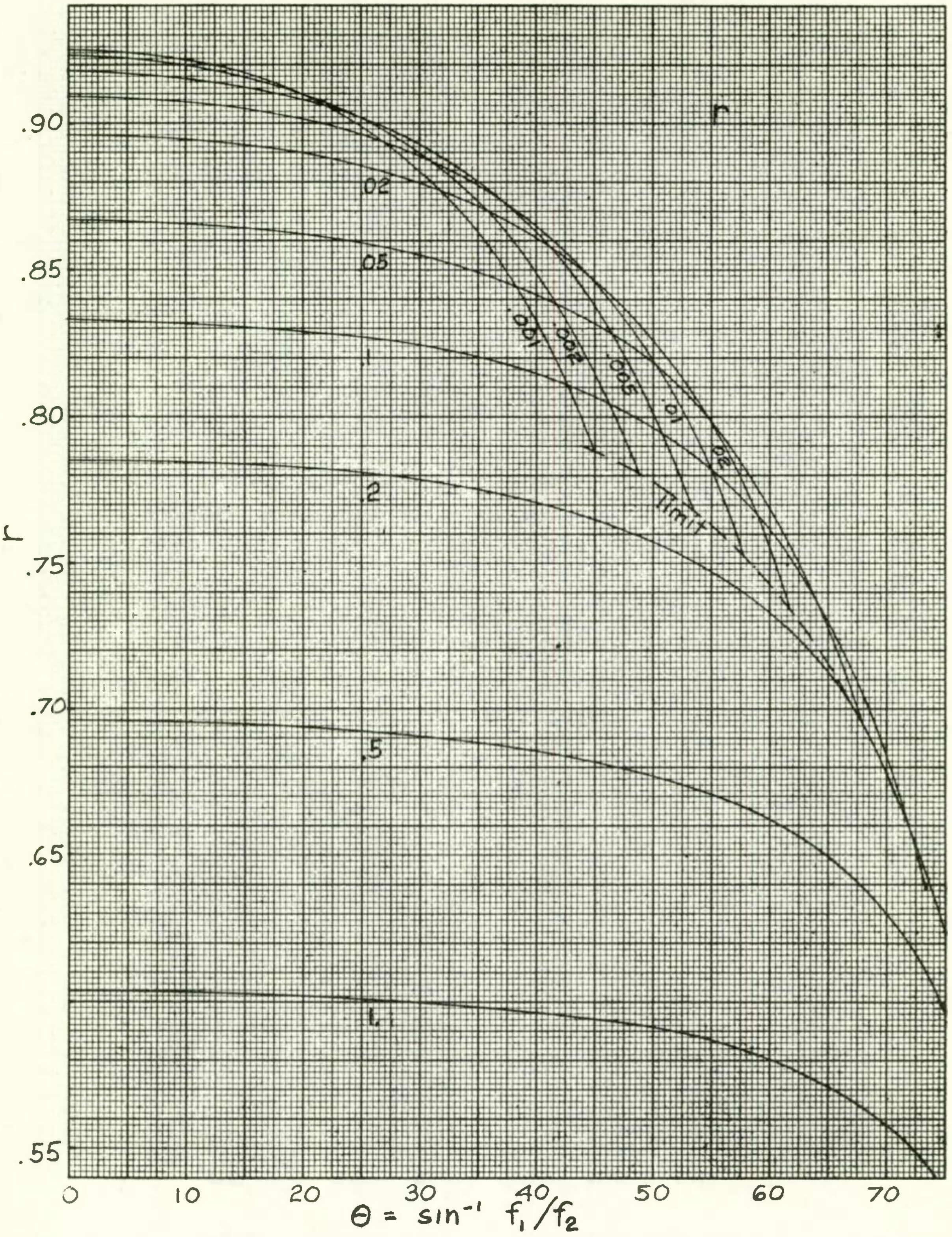


Fig. 5. Design chart for r, the ratio of the design resistance to the terminating resistances.

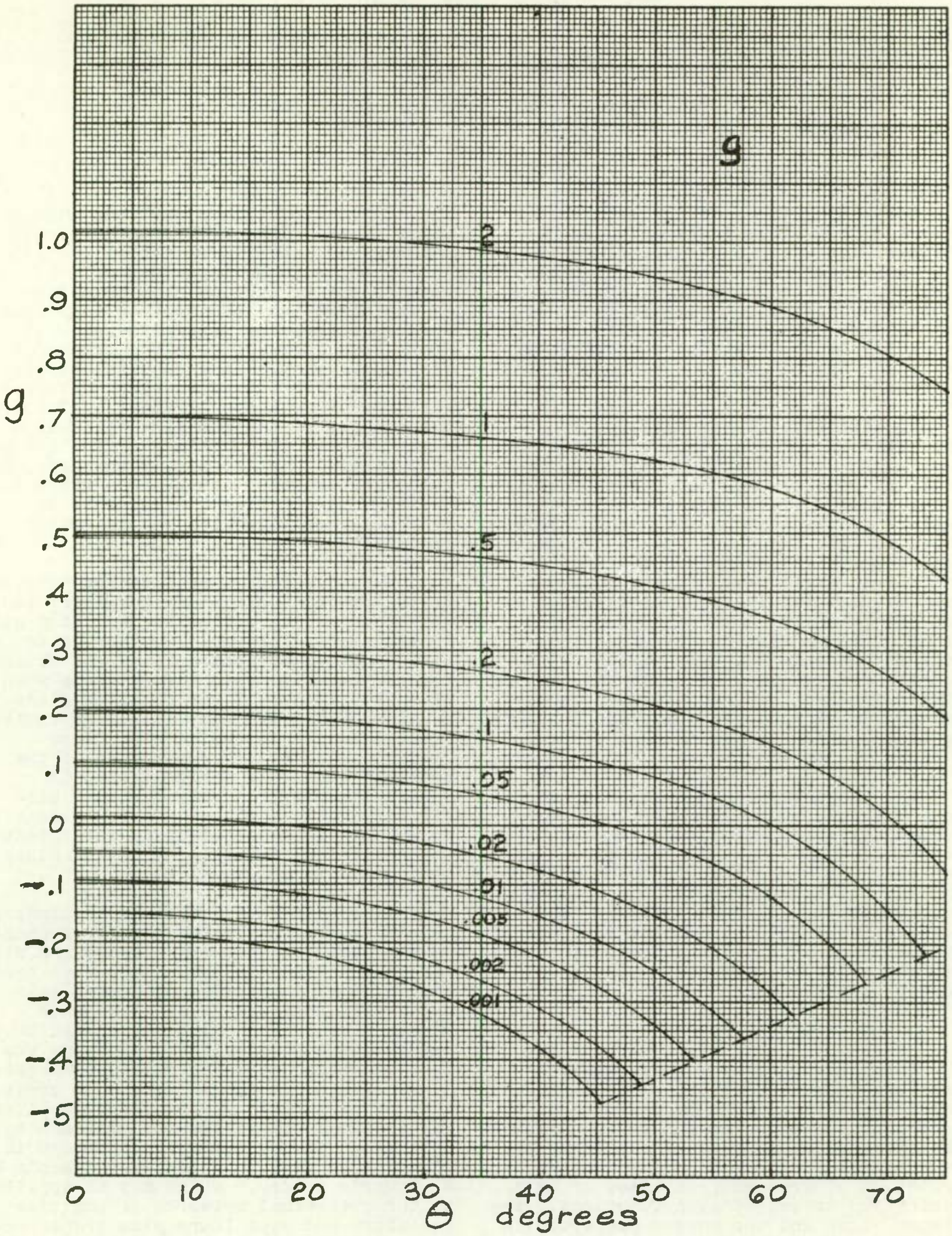


Fig. 6. Design chart for g, the supplementary end capacitance.

# SOME PROPERTIES OF MULTITERMINAL RC NETWORKS

Sidney Darlington
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey

## Summary

This paper is concerned with transformerless, multiterminal, grounded RC networks, for which the complete admittance matrices are assumed to be specified. Most of the remarks refer to 3-terminal networks, but suggest generalizations for n terminals. Certain properties of the admittance matrices are noted, and from these some canonical configurations are conjectured.

Specifically, it is easily shown that admittance poles at finite frequencies and the corresponding residues, can easily be realized without transformers, in any of many ways, but specific realizations impose lower bounds on the admittances at zero frequency and on the residues of poles at infinity. Thus poles at finite frequencies may be said to be bought at a price in terms of behavior at zero and infinity. By means of simple circuits, it is easily shown that the price may be paid entirely at zero or at infinite frequencies, or it may be divided between the two. The conjectured canonical configurations are composed of building blocks which appear to cost the least in terms of behavior at zero and infinity.

The conjectured canonical configurations are consistent with the so called series-parallel theorem, which was conjectured in 1955 but has not been proved or disproved. Furthermore, if the configurations are indeed canonical, they represent a substantially stronger theorem. They also suggest that direct specifications of necessary and sufficient conditions on the over-all admittance matrices must be extremely complicated.

## Introduction

This paper concerns a so-called classical network realization problem. Each such problem concerns networks of a particular class or type, and some particular external property of those networks. The external property is characterized by a function of frequency or time, or by a combination or set of such functions. The network class and the chosen property, in combination, determine a class of functions, or of (finite) sets of functions.

In most of this paper, the network class is the class of all three terminal networks of positive resistors and capacitors, but with no transformers. The results may be applied also to networks of resistors and inductors or inductors and capacitors, by means of very well known transformations on the frequency variable. Many of the results are easily generalized to transformerless networks with more than three terminals. Throughout the paper, the function class is a class of sets of functions of the frequency variable, $s = i\omega$, namely the driving-point and transfer admittances which completely characterize the currents in external short circuits due to voltage excitations.

One object of a realization problem is to determine necessary and sufficient conditions which define the function class in mathematical terms. Another is to find a method of constructing, or designing a network in the network class when given any member of the function class. Usually, there are many so called equivalent networks in the network class corresponding to any one member of the function class. A design method is usually appropriate only for some particular subclass of the network class. When the network subclass is sufficient for realizing the entire function class it is said to be canonical.

Many different classical realization problems have been attacked, corresponding to different network classes and different function classes. Many have been solved, some in various different ways using different canonical networks.[1] Others have defied solution, thanks to mathematical difficulties. Perhaps the most important of these relate to passive networks without transformers. A great deal of classical network theory achieves mathematical simplicity and elegance by including networks with transformers in the network class. A familiar example is W. Cauer's theory,[2] which may be applied to three terminal networks of positive resistors and capacitors plus transformers. In practical applications, transformers are highly undesirable. While some progress has been made toward a comprehensive realization theory for transformerless passive networks, severe mathematical difficulties have kept it very slow.

This paper reports a small, but perhaps significant advance in the general area. Simple relationships are developed which give a much better understanding of networks in the network class. They lead to a conjectured canonical three-terminal network subclass made up of quite simple building blocks. On the other hand, the function class is not delimited in terms of both necessary and sufficient mathematical conditions, nor is the canonical nature of the network subclass established rigorously. If the network subclass is in fact canonical, the mathematical boundaries of the function class must be very complicated indeed.

A three terminal network is commonly viewed as a two-port, or two-terminal-pair network in which the two ports share a common terminal. Fig. 1 is a block diagram. Two driving-point admittance functions, $Y_{11}$, $Y_{22}$, and one transfer admittance function, $Y_{12}$, completely determine the external current-voltage relationships, in accordance with

$$\begin{aligned} I_1 &= Y_{11}E_1 - Y_{12}E_2 \\ I_2 &= - Y_{12}E_1 + Y_{22}E_2 \ . \end{aligned} \qquad (1)$$

The separate admittances may of course be represented collectively as the admittance matrix:

$$Y = \begin{vmatrix} Y_{11} & - Y_{12} \\ -Y_{12} & Y_{22} \end{vmatrix} \ . \qquad (2)$$

We are concerned here with finding a network when given a complete set of the admittance functions. In a related, but somewhat simpler problem, only one or two of the functions is given, and at least one may be chosen arbitrarily (within the general function class). Solutions to this problem are already known,[3] but they are not easily modified to fit our present problem.

The present problem was brought to the author's attention by H. M. Lucal, in 1955.[4] He proposed a specific synthesis technique for networks in our network class, using a so called series-parallel decomposition. He did not claim that his method works more than some of the time. However, the author conjectured that Lucal's method is, in fact, canonical,[5] and in the six intervening years the conjecture has been neither proved nor disproved.

The series-parallel decomposition is roughly as follows: Figs. 2A and 2B illustrate respectively the parallel and series connection of subnetworks. When the subnetworks are connected in parallel, their short circuit admittances may be added. When they are connected in series, their open circuit impedances may be added. For a series-parallel decomposition, one separates the admittance functions or the impedance functions into parts appropriate for parallel connected or series connected subnetworks. Then one decomposes each subnetwork in a similar way (after transformation between admittance and impedance representations), and so on until the subnetworks are single branches.

The present paper supports the series-parallel conjecture, although it does not prove it. It also goes further, and conjectures a canonical configuration in which more specific subnetworks are connected in parallel. The subnetworks themselves have a quite special series-parallel configuration. It is shown that they can always be used to realize the finite poles which can be realized when transformers may be included, but generally at a price in terms of restrictions on the behavior at $s = 0$ or $\infty$. The price depends in a complicated way on the residues at the finite poles.

## Driving-Point and Transfer Admittances

We shall derive new insight from an analysis which is almost, but not quite conventional. To establish the point of departure, and its utility, we must review some well known network theory of an extremely elementary sort.

In Fig. 1, the two ports of the network share terminal 3. Given the network with three external terminals and no further instructions, one can form two other two-ports by using terminal 1 or 2 as the common terminal. The new admittance functions are quickly established as linear transformations on the old. Thus, all that can be establsihed about the network from its external properties can be established, in principle, from the properties of any one of the three possible two-ports. However, with any such representation one is likely to miss important relationships which stem from the inherent three-way symmetry of the network as a three-terminal device.

A well known representation which retains the three-way symmetry is the so called indefinite matrix Y:

$$Y = \begin{vmatrix} Y_{11} & -Y_{12} & -Y_{13} \\ -Y_{12} & Y_{22} & -Y_{23} \\ -Y_{13} & -Y_{23} & Y_{33} \end{vmatrix} \qquad (3)$$

It relates the currents into all three of the terminals to appropriate voltages. It is necessarily singular, for Kirchoff's current law makes the three currents sum to zero. To obtain the matrix $Y_j$, for input and output terminal pairs sharing terminal j, simply remove row j and column j from Y.

Corresponding to the general three-terminal network there is an equivalent $\pi$ network. It is represented in the usual two-port form in Fig. 3A, and rearranged in a symmetrical delta in Fig. 3B. It is an imagined network which has the same external properties as the general network. It is useful as an aid to understanding the external properties. It is not generally a physical embodiment of the admittance matrix, for the three two-terminal branches cannot generally be constructed out of positive physical components.

The indicated relations between the admittance of the two-terminal branches and the off-diagonal elements in matrix Y are easily established. Then it is quickly shown that

$$\begin{aligned} Y_{11} &= Y_{12} + Y_{13} \\ Y_{22} &= Y_{12} + Y_{23} \qquad (4) \\ Y_{33} &= Y_{13} + Y_{23} \end{aligned}$$

Thus Y is not only singular. It is singular in a particular way. The direct sum of the rows or columns is identically zero, not just a weighted sum with unspecified coefficients.

From (4), the external behavior of the three-terminal network may be specified in terms of $Y_{12}$, $Y_{13}$, $Y_{23}$ (as functions of frequency) rather than $Y_{11}$, $Y_{22}$, $Y_{12}$. The three way symmetry achieved in this way we shall find very illuminating.

## Frequency Functions

Much of classical network theory stems from partial fraction expansions of frequency functions. For present purposes, the appropriate expansions are

$$Y_{ij} = \sum_{\sigma=1}^{n} \frac{K_{\sigma ij}\, s}{s + s_\sigma} + K_{oij} + K_{\infty ij}\, s \qquad (5)$$

$K_{\infty ij}$ is the residue of $Y_{ij}$ at a pole at $s = \infty$. We will also refer to $K_{\sigma ij}$ and $K_{oij}$ as residues, although in strict fact they are residues at poles of $Y_{ij}/s$. Either or both of $K_{oij}$ and $K_{\infty ij}$ may be zero.

In classical two-port theory[2], $Y_{11}$, $Y_{22}$, $Y_{12}$ are covered by i,j = 1,2. Then necessary and sufficient conditions for realization with positive resistors and capacitors plus ideal transformers are

$$s_\sigma = \text{real and} > 0$$

$$K_{\rho ij} = \text{real}$$

$$\begin{vmatrix} K_{\rho 11} & -K_{\rho 12} \\ -K_{\rho 12} & K_{\rho 22} \end{vmatrix} = \text{nonnegative definite} \qquad (6)$$

$$\rho = \sigma,\ o,\ \infty\ .$$

The condition on the residue matrix can be broken into:

$$\begin{aligned} K_{\rho 11},\ K_{\rho 22} &\geq 0 \\ K_{\rho 11}K_{\rho 22} - K^2_{\rho 12} &\geq 0 \qquad (7) \\ \rho = \sigma,\ o,\ \infty\ . \end{aligned}$$

The same set of $s_\sigma$'s is used in the expansions of $Y_{11}$, $Y_{22}$, $Y_{12}$. It is understood that $K_{\rho ij} = 0$ may be used to remove some of the poles from certain of the admittance functions, but the second condition in (7) requires a nonzero $K_{\rho 11}$ and $K_{\rho 22}$ for every nonzero $K_{\rho 12}$.

To avoid reference to a specific choice of common terminal, we may simply let i, j in (5) run over 1,2,3. Then, from (4)

$$K_{\rho ii} = K_{\rho ij} + K_{\rho ik} \qquad (8)$$

Using this relationship in (7), and in the corresponding condition on $K_{\rho 33}$ gives

$$\begin{gathered} K_{\rho 12} + K_{\rho 13} \geq 0 \\ K_{\rho 12} + K_{\rho 23} \geq 0 \\ K_{\rho 13} + K_{\rho 23} \geq 0 \\ K_{\rho 12}\, K_{\rho 13} + K_{\rho 12}\, K_{\rho 23} + K_{\rho 13}\, K_{\rho 23} \geq 0 \\ \rho = \sigma,\ o,\ \infty\ . \end{gathered} \tag{9}$$

This is an alternative set of necessary and sufficient conditions for realization with positive resistors and capacitors plus ideal transformers. (Actually, only two of the first three conditions need be stated, for then the third is implied by the fourth condition. The three are included here to retain the three-way symmetry.)

## Behavior at Zero and Infinite Frequencies

The constants $K_{o1j}$ determine the admittances at s = 0:

$$Y_{1j}(0) = K_{o1j}\ . \tag{10}$$

We shall refer to the corresponding matrix, $K_o$, as the behavior of the network at s = 0. Similarly, if there is a pole at s = ∞,

$$Y_{1j}(s \rightarrow \infty) = K_{\infty 1j}\ s \tag{11}$$

We shall refer to the corresponding residue matrix $K_\infty$ as the behavior of the network at s = ∞. Changes in $K_o$ and $K_\infty$ do not change the finite poles, or the corresponding residues, but they do change the admittances at all frequencies except the poles.

Well known properties of transformerless networks include the following: When a three terminal network of positive resistors and capacitors includes no transformers,

$$\begin{gathered} K_{o1j} \geq 0 \\ K_{\infty 1j} \geq 0 \\ i,j = 1,2,3 \end{gathered} \tag{12}$$

The behavior at s = 0 is the same as for the simple resistance network illustrated in Fig. 4A. The behavior at s = ∞ (pole and residues) is the same as for the simple capacitance network illustrated in Fig. 4B. Thus, the behaviors at zero and infinity can be realized, by themselves, with positive resistors and capacitors, whenever the admittances are appropriate for any transformerless network of positive resistors and capacitors.

We shall assume hence forth that conditions (12) are satisfied by our admittance functions.

## Rank-One Residue Matrices at Finite Poles

Important concepts will be clearer if we impose, temporarily, the following arbitrary restriction: The equal sign is to apply to the last condition of (9) for all finite poles (ρ=σ), but not necessarily for ρ = 0 or ∞. The corresponding residue matrices have rank one. Equations (9) become

$$\begin{gathered} K_{\rho 12} + K_{\rho 13} \geq 0 \\ K_{\rho 12} + K_{\rho 23} \geq 0 \quad \rho = \sigma,\ o,\ \infty \\ K_{\rho 13} + K_{\rho 23} \geq 0 \\ K_{\sigma 12}\, K_{\sigma 13} + K_{\sigma 12}\, K_{\sigma 23} + K_{\sigma 13}\, K_{\sigma 23} = 0 \\ K_{o12}\, K_{o13} + K_{o12}\, K_{o23} + K_{o13}\, K_{o23} \geq 0 \\ K_{\infty 12}\, K_{\infty 13} + K_{\infty 12}\, K_{\infty 23} + K_{\infty 13}\, K_{\infty 23} \geq 0 \end{gathered} \tag{13}$$

We shall remove the restriction later on, but only at a cost of some quite subtle complications (in conditions for realizability without transformers).

Consider the implications of these equations regarding the signs of the three residues $K_{\sigma 12}$, $K_{\sigma 13}$, $K_{\sigma 23}$ corresponding to a single pole. Because of the first three equations, no more than one of the three residues may be negative. Because of the fourth equation, either two are zero or at least one is negative. Hence

> Theorem (proved): When the residue matrix $K_\sigma$ has rank one, out of $K_{\sigma 12}$, $K_{\sigma 13}$, $K_{\sigma 23}$, either two are zero and one is positive, or else one is negative and two are positive.

When a residue matrix need not have rank one (for example our $K_o$ and $K_\infty$) it is still true that no more than one of the three transfer admittance residues can be negative. However, all three can now be positive, or one can be zero with the other two positive.

Given a set of admittances we can now classify, or characterize the set by a pattern of 0's, +'s, and -'s characterizing the residues $K_{\sigma 1j}$ of the three transfer admittances. The pattern may be displayed as a table such as the following:

| Pole Number | $Y_{12}$ | $Y_{13}$ | $Y_{23}$ |
|---|---|---|---|
| 1 | + | 0 | 0 |
| 2 | + | - | + |
| 3 | + | + | - |
| 4 | 0 | + | + |
| 5 | + | 0 | 0 |
| 6 | + | - | + |

(14)

If only the finite poles are represented, and are subject to the rank one residue condition, each row contains either two 0's and one + or else one - and two +'s.

## Realization of Single Poles with T Networks

The partial fraction expansions suggest, of course, W. Cauer's celebrated canonical network for two-ports made up of resistors and capacitors plus ideal transformers.[2] In Cauer's network, a number of subnetworks are connected in parallel. Each admittance, $Y_{ij}$, of the combination is simply the sum of the corresponding admittances of the parallel connected parts. There is a separate subnetwork corresponding to each admittance pole. More exactly a single subnetwork realizes a set of partial fractions, in the various admittance functions but all corresponding to a single pole. In Cauer's network, most of the subnetworks include ideal transformers. We now examine alternative subnetworks, which avoid the transformers but usually include negative resistors or capacitors. When all the subnetworks are connected together, the negative components can frequently be cancelled by positive components, as we shall see.

Consider the simple T network illustrated in Fig. 5, in which two branches are resistors and one is a capacitor. The corresponding transfer admittance functions may be arranged as follows, in which $g_a$ and $g_b$ are conductances and c is capacitance.

$$Y_{ab} = - \frac{g_a\, g_b}{g_a + g_b}\,\frac{s}{s + s_o} + \frac{g_a\, g_b}{g_a + g_b}$$

$$Y_{ac} = + g_a \frac{s}{s + s_o} \qquad (15)$$

$$Y_{bc} = + g_b \frac{s}{s + s_o}$$

$$s_o = \frac{g_a + g_b}{c} \quad .$$

The three functions have a common finite pole, with a rank one residue matrix. This suggests using a separate T as a subnetwork for realizing each finite pole of a set of admittance functions of general degree. The application is complicated, however, by the added constant term in $Y_{ab}$.

The constant term disappears if a suitable negative resistor is connected between the external terminals a and b. The combination of the T and associated negative resistor is illustrated in Fig. 6. We may use it as a subnetwork in synthesizing a network of positive components provided we can introduce, eventually, positive resistors which cancel all such negative resistors. Appropriate synthesis formulas for the subnetwork are as follows:

$$Y_{\sigma ij} = \frac{K_{\sigma ij}\; s}{s + s_\sigma} , \qquad i,j = a,b,c$$

$$g_a = K_{\sigma ac}$$

$$g_b = K_{\sigma bc} \qquad (16)$$

$$c = \frac{K_{\sigma ac} + K_{\sigma bc}}{s_\sigma}$$

$$-g_{ab} = K_{\sigma ab} = - \frac{K_{\sigma ac}\; K_{\sigma bc}}{K_{\sigma ac} + K_{\sigma bc}} \quad .$$

The last equation is a simple rearrangement of our condition for a residue matrix of rank one.

If, out of $K_{\sigma 12}$, $K_{\sigma 13}$, $K_{\sigma 23}$ one is $< 0$ and two are $> 0$, terminals a, b, c can be identified with 1, 2, 3 in such an order that $K_{\sigma ac}$ and $K_{\sigma bc}$ are $> 0$. Then (16) defines a corresponding network in which $g_1$, $g_2$, c are $> 0$. The corresponding negative resistor (Fig. 6) is bridged between terminals 12, 13, or 23, whichever corresponds to $K_{\sigma ij} < 0$.

If, out of $K_{\sigma 12}$, $K_{\sigma 13}$, $K_{\sigma 23}$, two are zero and one is positive, a, b, c can be identified with 1, 2, 3 in such an order that $K_{\sigma ac} > 0$ and $K_{\sigma bc} = 0$. Then $g_a$ and c are $> 0$ while $g_b = 0$ and also $g_{ab}$. Since a zero admittance is an open circuit, the network of Fig. 6 now degenerates into the single branch illustrated in Fig. 7. Tne single branch is connected between terminals 12, 13 or 23, whichever corresponds to $K_{\sigma ij} > 0$.

The properties of these networks are further illustrated by a comparison with Cauer's canonical subnetworks. Fig. 8

indicates exact equivalences. Note that the orientation of our networks depends on the voltage ratio φ of Cauer's ideal transformer, but that the various orientations cover all values of φ.

From Fig. 8, T networks (and degenerate T's) may be used as subnetworks in a transformerless counterpart of Cauer's canonical network, provided the negative resistors can somehow be absorbed (or else may be tolerated as such). However, the chance of avoiding negative components is much improved if one has available also a second T configuration.

Consider the T network illustrated in Fig. 9, in which two branches are capacitors and one is a resistor. As before, all the admittance functions share a common finite pole. Now, however, $Y_{\sigma ab}$ has also a pole at $s = \infty$, instead of a nonzero behavior at $s = 0$. The pole at infinity may be removed by bridging a negative capacity between terminals a and b, as illustrated in Fig. 10.

The networks illustrated in Figs. 6 and 10 are externally equivalent, in the usual network sense. Appropriate synthesis formulas for the two capacitor T are as follows (corresponding to (16) for the two resistor T):

$$Y_{\sigma ij} = \frac{K_{\sigma ij}\, s}{s + s_\sigma}, \quad i,j = a,b,c$$

$$c_a = \frac{K_{\sigma ac}}{s_\sigma}$$

$$c_b = \frac{K_{\sigma bc}}{s_\sigma} \qquad (17)$$

$$g = K_{\sigma ac} + K_{\sigma bc}$$

$$-c_{ab} = \frac{K_{\sigma ab}}{s_\sigma} = -\frac{K_{\sigma ac}\, K_{\sigma bc}}{(K_{\sigma ac} + K_{\sigma bc})\, s_\sigma} .$$

The components of the two T configurations are positive under exactly the same conditions. They degenerate into the same single branch illustrated in Fig. 7, under the same conditions. A choice between them may depend upon whether a negative resistor or capacitor is more easily absorbed.

More generally, two T networks, one of each kind, may be connected in parallel, to realize a single finite admittance pole. Both a negative conductance and a negative capacitance are associated with the combination, but each is smaller than it would be in the absence of the other. The configuration is illustrated in Fig. 11. While it is more complicated than the equivalent single T subnetworks, it is important for general synthesis techniques which we shall consider. Appropriate formulas for the components are as follows (combining (16) and (17)):

$$Y_{\sigma ij} = q\,\frac{K_{\sigma ij}\, s}{s + s_\sigma} + (1-q)\,\frac{K_{\sigma ij}\, s}{s + s_\sigma}$$

$$0 \le q \le 1$$

$$g_a = qK_{\sigma ac}\ , \quad c_a = (1-q)\,\frac{K_{\sigma ac}}{s_\sigma}$$

$$g_b = qK_{\sigma bc}\ , \quad c_b = (1-q)\,\frac{K_{\sigma bc}}{s_\sigma} \qquad (18)$$

$$c = q\,\frac{K_{\sigma ac} + K_{\sigma bc}}{s_\sigma}\ , \quad g = (1-q)(K_{\sigma ac}+K_{\sigma bc})$$

$$-g_{ab} = q\,K_{\sigma ab}\ , \quad -c_{ab} = (1-q)\,\frac{K_{\sigma ab}}{s_\sigma} .$$

A well known theorem requires q to be the same for all ij, so long as the residue matrix of the combination is to have rank one.

Theorem (known): When two parallel connected networks of resistors and capacitors have a common admittance pole, the residue matrix $K_\sigma$ for the combination has rank one if and only if:

a. The residue matrices for the two networks have rank one,

b. $K'_\sigma = qK''_\sigma$, where q is a scalar.

The second condition requires, of course, $K'_{\sigma ij} = qK''_{\sigma ij}$, and this requires the same sign sequences as the $K_{\sigma ij}$'s in a table like (14).

## General Synthesis in Terms of T Networks

We can now put together a network realization of the complete admittance functions. Corresponding to each nondegenerate finite pole (a pole of all the admittance functions) one may choose a T circuit of either kind, or a parallel combination of the two, accompanied by the appropriate negative component or components (Figs. 6 and 10). Corresponding to each degenerate finite pole, there is a single two-terminal branch

(Fig. 7). Corresponding to the behavior at $s = 0$ and $\infty$, there are the subnetworks of Fig. 4. The complete network is formed by connecting all these subnetworks in parallel.

The negative components which come with the T networks are connected between external terminal pairs, 12, 13, 23. So are the positive components which correspond to $K_{o1j}$ and $K_{\infty 1j}$ and represent behavior at $s = 0$ and $\infty$. All the positive and negative components of any one kind (resistors or capacitors) across any one terminal pair may be replaced by a single component. Then the complete network takes the form illustrated in Fig. 12 (in which single 2-terminal branches (Fig. 7) may be included as degenerate T's). It is a complete realization of the admittance functions, without transformers. All the components are positive if and only if

$$K_{oij} \geq g_{ij}$$
$$K_{\infty ij} \geq c_{ij} \qquad (19)$$
$$ij = 12,\ 13,\ 23$$

where $-g_{ij}$ and $-c_{ij}$ are the total negative conductance and capacitance across terminals ij associated with the T networks.

For the most general choice of the T networks

$$g_{ij} = \sum_{\eta} q_\eta(-K_{\eta ij})$$
$$c_{ij} = \sum_{\eta} (1-q_\eta)\ \frac{-K_{\eta ij}}{s_\eta}\ . \qquad (20)$$

For any one choice of ij, index $\eta$ takes on only those integer values such that $K_{\eta ij} < 0$. The constants $q_\eta$ may be chosen arbitrarily (and independently) in the range $0 \leq q_\eta \leq 1$.

The restrictions on $K_{oij}$ and $K_{\infty ij}$ are a minimum for choices of the $q_\eta$'s of a special sort. Let the partial fractions of the admittance functions be so numbered that $s_\eta$ increases with $\eta$. Then make every $q_\eta = 0$ or 1 except for one, in such a way that

$$g_{ij} = \sum_{\eta<\mu} (-K_{\eta ij}) + q_\mu(-K_{\mu ij})$$
$$c_{ij} = (1-q_\mu)\ \frac{-K_{\mu ij}}{s_\mu} + \sum_{\eta>\mu} \frac{-K_{\eta ij}}{s_\eta}\ . \qquad (21)$$

Given any other choice, there is always a choice of this sort which reduces both $g_{ij}$ and $c_{ij}$. The reason stems from the factor $1/s_\eta$ in the second equation, which reduces contributions to $c_{ij}$ at larger $s_\eta$.

In (21), $\mu$ may be any $\eta$ from (20), and then $0 \leq q_\mu \leq 1$. The permissible choices establish a relation between $c_{ij}$ and $g_{ij}$, which is illustrated in Fig. 13. Note its concave upward, broken straight line character. The values of $K_{oij}$, $K_{\infty ij}$ may be represented by a point in the same capacitance-conductance plane. Then (19) requires the point to be in the positive quadrant and above the $c_{ij}$-$q_{ij}$ curve.

There are three such conditions, corresponding to ij = 12, 13, 23. Because $K_{\sigma ij} < 0$ for no more than one ij, the three conditions correspond to nonoverlapping subsets of the admittance poles. Thus $\mu$ and $q_\mu$ may be chosen independently for each.

When added to the necessary and sufficient conditions on networks which include transformers, the new conditions complete a necessary and sufficient set for passive network synthesis in terms of our parallel combination of T subnetworks.

## Residue Matrices of Rank Two

In the above, we simplified the argument by assuming, temporarily, that all finite poles have rank one residue matrices. Now suppose a pole at $s = -s_\sigma$ has a rank two residue matrix. This requires

$$K_{12}K_{13} + K_{12}K_{23} + K_{13}K_{23} > 0\ . \qquad (22)$$

All three of the transfer admittance residues may now be > 0. Then the corresponding partial fractions may be realized as a physical $\pi$ network, like Fig. 4 except that each branch is like Fig. 7.

On the other hand, one of the transfer residues may still be negative, with the other two positive and too large for the rank one condition. A portion of one of the partial fractions with positive residues may now be split off and realized as in Fig. 7; -- just enough so that the reduced residue matrix has rank one.

A more subtle alternative uses two T subnetworks in parallel. Recall that the parallel combination can have a rank one residue matrix only if $Y'_{\sigma ij} = q\ Y''_{\sigma ij}$, where q is the same for all ij. By choosing q differently for different ij, one can obtain a parallel T combination with a rank 2 matrix.

In either case, the conditions on $Y_{0ij}$ and $Y_{\infty ij}$ are determined entirely by the negative residue $K_{\sigma ij} < 0$, and in exactly the same way as before. Thus our previous conclusions are unaffected by rank two residue matrices. There may be additional branches across external terminals, but condition (19), equation (21), and Fig. 13 remain unchanged. However, more sophisticated subnetworks, which we must now consider, may be affected in much more subtle ways.

## More Complicated Building Blocks

When combined with the previously known condition $K_{0ij}$, $K_{\infty ij} \geq 0$, our networks of parallel T's established the following:

> Theorem (Proved): When a three-terminal network of positive resistors and capacitors is used to realize a given admittance matrix, the realizability of the finite poles is the same whether or not transformers are used, but omission of transformers must be paid for by restrictions on the behavior at s = 0 and/or ∞. Any restriction beyond $K_{0ij}$, $K_{\infty ij} \geq 0$ may be paid entirely in terms of behavior at s = 0, or at s = ∞, or partly in terms of each.

The curve in Fig. 13 established an upper bound on the price of no transformers, in terms of minimum permitted $K_{0ij}$, $K_{\infty ij}$. However, it is not necessarily a least upper bound. It is easy to find circuits which pay lower prices under more restricted conditions.

A convenient generalization retains the parallel connection of subnetworks, but adds new kinds of subnetworks to our T's and two-terminal branches. The new subnetworks usually include one or more negative components between external terminals. They are useful when the negative admittances are smaller than those required with equivalent combinations of our previous subnetworks. The partial fractions corresponding to a single pole may be divided into portions assigned to several subnetworks.

Two subnetworks which come at reduced prices are illustrated in Figs. 14A and B. Each has two finite poles, with rank one residue matrices, and also $K_{0ij}$ and $K_{\infty ij} = 0$. In terms of restrictions on the behavior of the complete network at s = 0 and ∞, one costs nothing, and the other costs less than the equivalent pair of T networks. Thus, when they can be used, these subnetworks ease our previous restrictions. Unfortunately, they can be used only under rather restricted conditions.

Recall our discussion of the signs of residues, illustrated by the table in (14). When a set of admittance functions is given, a pattern of residue signs is established. Suppose the residue matrices have rank one. Then, in any division of partial fractions into separately realizable parts, the parts must retain the same sign pattern. The sign pattern for any two nondegenerate poles may take either of the following two forms

| Pattern | Pole | $K_{\sigma ij}$ | $K_{\sigma ik}$ | $K_{\sigma jk}$ |
|---|---|---|---|---|
| A | 1 | + | - | + |
|  | 2 | - | + | + |
| B | 1 | - | + | + |
|  | 2 | - | + | + |

(23)

On the other hand, our new subnetworks can have only the first pattern (A), and simply cannot be used even for portions of rank one partial fractions which follow the second. There are other restrictions on their residues, but these are the most important.

There are many other possible subnetworks with two finite poles. There are many more with more than two poles. When given residue matrices have rank one, only those subnetworks may be used whose residues match the sign pattern of the corresponding complete partial fractions. When residue matrices have rank two, it may or may not be necessary to match sign patterns, depending on residue magnitudes. Then restrictions on synthesis applications become complicated in the extreme.

On the basis of laborious studies of specific configurations, which need not concern us here, the author has arrived at two conjectures. First

> Theorem (Conjectured): The parallel T configuration is a canonical network for the subclass of our network class such that the residue sign pattern (14) contains only -'s and 0's in one of the three columns.

Second, consider the general ladder illustrated in Fig. 15. Each series branch may be a resistor or a capacitor or the two in series. Each shunt branch may be a resistor or a capacitor or the two in parallel. It follows that

$$Y_{ab} = \frac{s^{\lambda}}{a_o + a_1 s + \ldots a_{\nu} s^{\nu}} \tag{24}$$

$$0 \leq \lambda \leq \nu + 1 \quad , \quad 0 \leq \nu \leq n \ .$$

The terminals a b c may be connected to the terminals 1 2 3 in any order.

Theorem (Conjectured): Parallel connected subnetworks of the sort defined by Fig. 15 and Eq. (24) constitute a canonical transformerless three terminal network of positive resistors and capacitors.

## Networks with More than Three Terminals

The synthesis of three-terminal networks in terms of parallel T subnetworks is easily generalized for networks with n terminals. The short circuit admittances may be collected in an indefinite matrix, like (3) except of order n. They may be expanded in partial fractions, like (5) except that i and j now run through 1,..., n. The counterpart of the T subnetwork is the star configuration illustrated in Fig. 16.

In the star, each branch may be a resistor or a capacitor (but not both) or an open circuit. Then the admittance in branch i, to terminal i, is

$$y_i = g_i \quad \text{or} \quad c_i s \quad \text{or} \quad 0 \ . \tag{25}$$

The typical corresponding transfer admittance is

$$Y_{ij} = \frac{Y_i Y_j}{\sum\limits_{k=1}^{n} Y_k} = \frac{Y_i Y_j}{\sum g_k + \sum c_k s} \ . \tag{26}$$

It may be rearranged as follows:

$$Y_{ij} = \frac{-1}{\sum g_k} \, \frac{Y_i(-s_o) Y_j(-s_o) s}{s + s_o} + Q_{ij}$$

$$s_o = \frac{\sum g_k}{\sum c_k} \ . \tag{27}$$

The added term $Q_{ij}$ depends on the components i and j, in branches i and j. If they are both resistors $Q_{ij}$ is a constant. If they are both capacitors $Q_{ij}$ is proportional to s. If one is a resistor and one a capacitor, $Q_{ij} = 0$. When $Q_{ij} \neq 0$, it may be removed by associating a negative resistor or capacitor with the star, connected between external terminals i and j. When there are more than three nonzero $Y_i$, more than one pair of external terminals require negative components.

The admittances again have a single finite pole, and the residue matrix has again rank one. When components i and j are similar (both resistors or both capacitors) $K_{ij}$ is again negative. When they are dissimilar, $K_{ij}$ is again positive. As before, there is a one to one correspondence between negative $K_{ij}$ and terminal pairs which get negative associated components.

Given a set of partial fractions corresponding to a single admittance pole and with a rank one, nonnegative-definite residue matrix, one can always find a corresponding star of positive components (plus associated negative components across certain of the external terminals). There are, in fact, two corresponding stars, with capacitors and resistors appearing in one where there are resistors and capacitors in the other.

Star networks corresponding to all the partial fractions in a given set of admittances may be connected in parallel, in a transformerless counterpart of Cauer's canonical network for (n-1)-ports.[6] All components are at once positive except for possibly negative components across external terminal pairs. All components will be positive provided the behaviors at s = 0 and ∞ meet conditions which depend in a complicated way on the residues at the finite admittance poles.

The penalty for no transformers is somewhat flexible, in regard to division between restrictions on $K_{oij}$ and $K_{\infty ij}$, but it is not so arbitrary as before. The trouble is, a single star may require an associated negative resistor and negative capacitor (across different terminal pairs). Then the choice is between contributions to restrictions on $K_{oij}$ and $K_{\infty kl}$ or $K_{\infty ij}$ and $K_{okl}$.

Speculation regarding subnetworks with a lower penalty than stars, like the ladders for three-terminal networks, is still too vague to justify further remarks here.

## References

1. Darlington, S., "A Survey of Network Realization Techniques". TRANS. IRE, Vol. CT2, (1955), pp. 291-297.

2. Cauer, W., "Untersuchungen über ein Problem, das drei positiv definite quadratishe Formen mit Streckenkomplexen in Beziehung Setzt". Math. Ann. Vol. 105 (1931), pp. 86-132.

3. Fialkow, A.D. and Gerst, I., "The Transfer Function of General Two-Terminal-Pair RC Networks". Quarterly Applied Mathematics, Vol. 10 (1952), pp. 113-127.

4. Lucal, H. M., "Synthesis of Three-Terminal RC Networks". TRANS. IRE, Vol. CT2 (1955), pp. 308-316.

5. Reference 1, p. 295.

6. Cauer, W., "Äquivalenz von 2n-Polen ohne Ohmsche Viderstande". Nachrichten von der Gesellschaft der Vissenshaften (N.F.), Vol. 1 (1934), pp. 3-33.

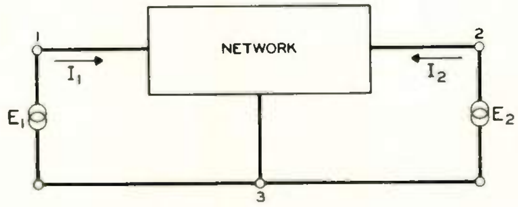


FIG. 1. A 3-TERMINAL NETWORK AS A 2-PORT

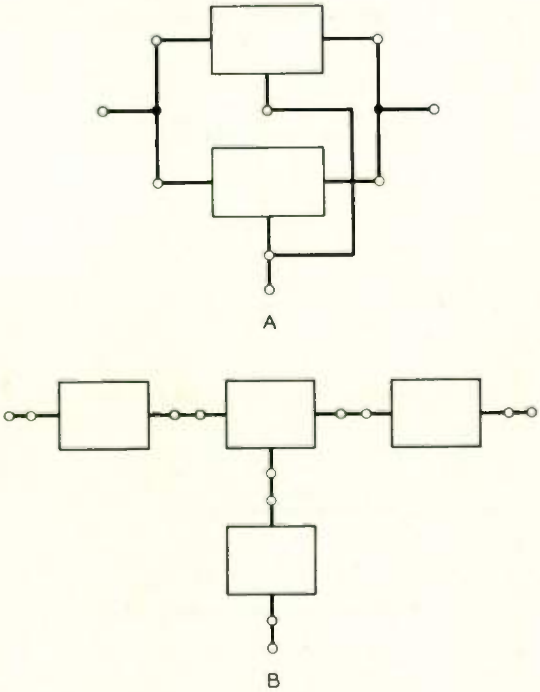


FIG. 2. PARALLEL (A) AND SERIES (B) CONNECTION OF SUBNETWORKS

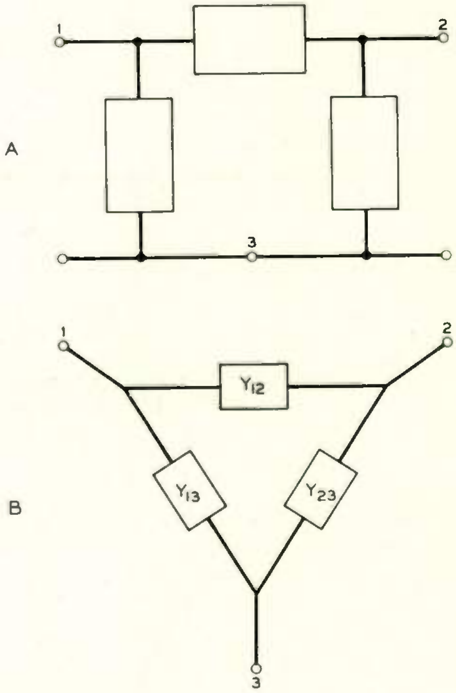


FIG. 3. EQUIVALENT $\pi$ OR $\Delta$

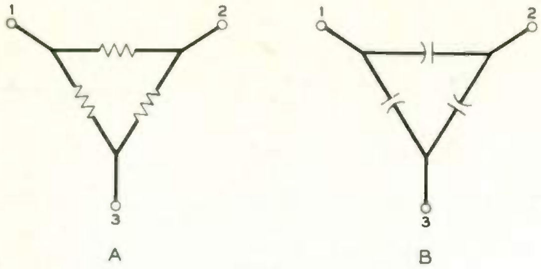


FIG. 4. REALIZATION OF BEHAVIOR AT S = 0 AND ∞

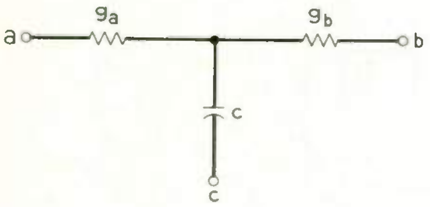


FIG. 5. A SIMPLE T NETWORK

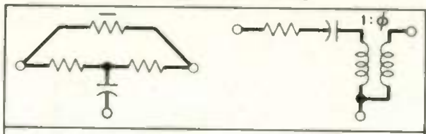
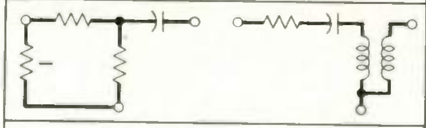
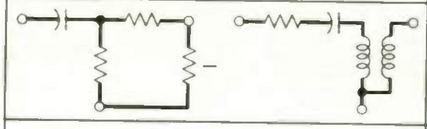
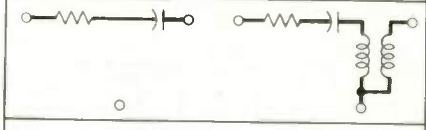
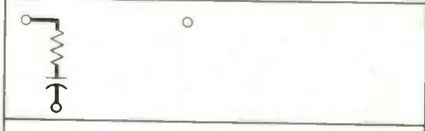
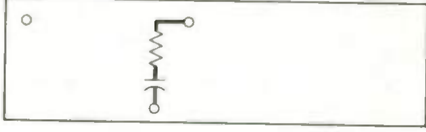
| EQUIVALENT NETWORKS | TRANSFORMER VOLTAGE RATIO |
|---|---|
|  | $\phi < 0$ |
|  | $0 < \phi < 1$ |
|  | $1 < \phi$ |
|  | $\phi = 1$ |
|  | $(\phi = \infty)$ |
|  | $(\phi = 0)$ |

FIG. 8. SOME NETWORK EQUIVALENCES

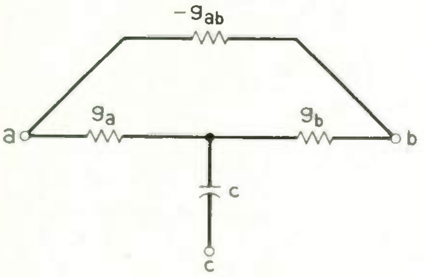


FIG. 6. A T PLUS ASSOCIATED NEGATIVE RESISTOR

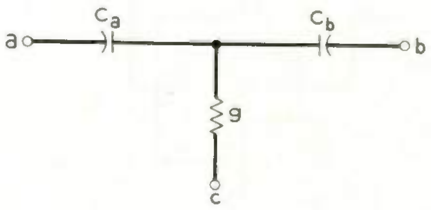


FIG. 9. AN ALTERNATIVE T NETWORK

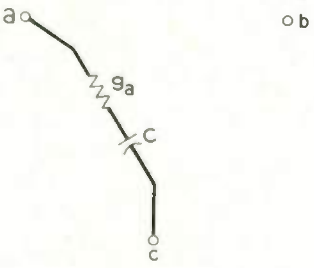


FIG. 7. A SINGLE-BRANCH DEGENERATE T

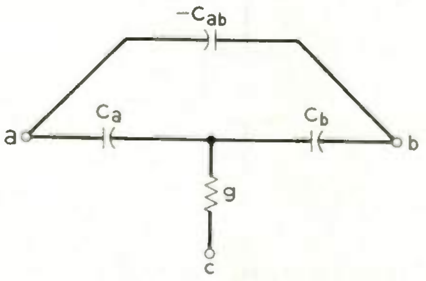


FIG. 10. A T PLUS ASSOCIATED NEGATIVE CAPACITOR

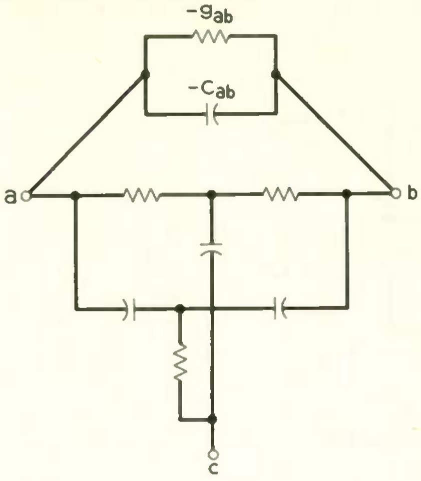


FIG. 11. A 2 T COMBINATION

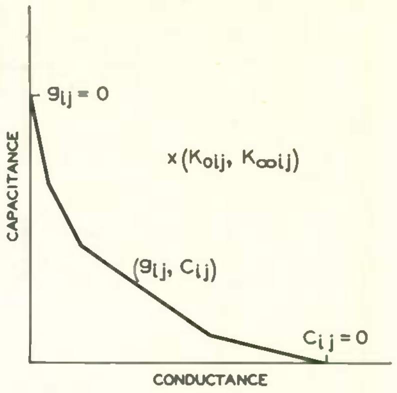


FIG. 13. THE BOUND ON $K_{0ij}$, $K_{\infty ij}$ FOR PARALLEL T SYNTHESIS

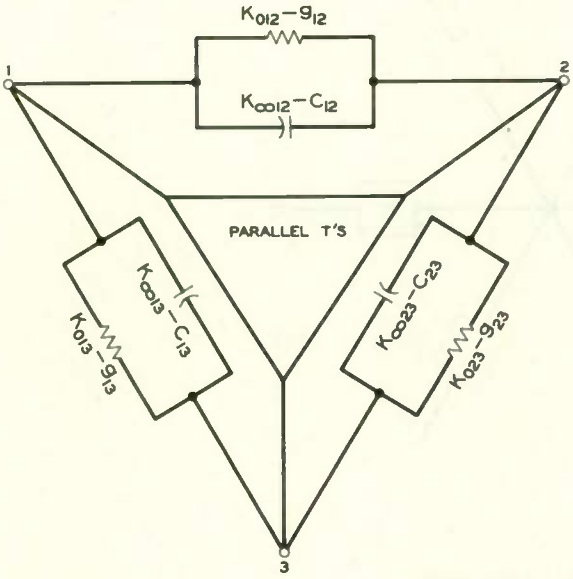
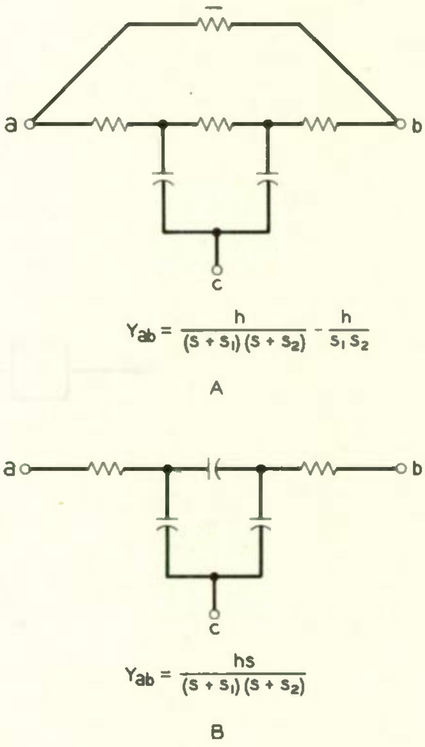


FIG. 12. NETWORK SYNTHESIS IN TERMS OF PARALLEL T SUBNETWORKS



$$Y_{ab} = \frac{h}{(s+s_1)(s+s_2)} - \frac{h}{s_1 s_2}$$

A

$$Y_{ab} = \frac{hs}{(s+s_1)(s+s_2)}$$

B

FIG. 14. SOME LADDER TYPE SUBNETWORKS

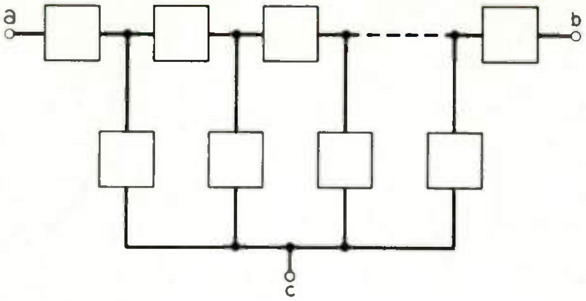


FIG. 15. A GENERAL LADDER SUBNETWORK

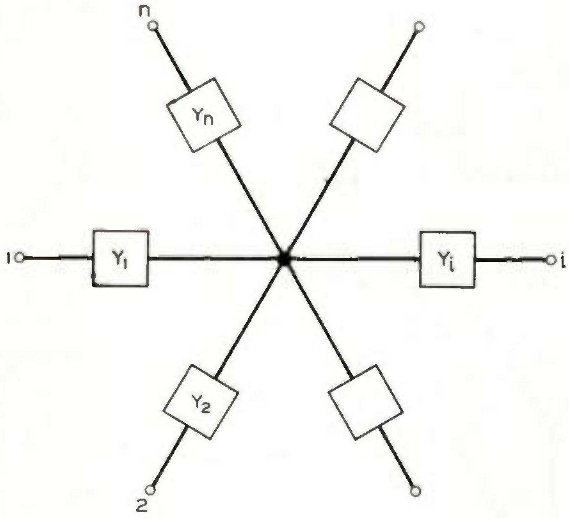


FIG. 16. A STAR NETWORK

iterative parameters positive real, equal and coincident with its conjugate matching parameters. For this to be possible, the invariant stability factor, s, of each stage, assumed equal, must be greater than (or equal to) unity.

At the frequency of maximum power gain, $f_o$, the chain network consists of m individual two-port networks with symmetry in self parameters ($p_{11} = p_{22}$) and with or without symmetry in transfer parameters ($p_{12} = p_{21}$ or $p_{12} \neq p_{21}$).

From the definition of transmission parameters

$$P_{o1} = P_{o2} = p_{o1} = p_{o2} = \rho_o \qquad (37)$$

$$P_{y1} = (p_{y1})^m \; ; \; P_{y2} = (p_{y2})^m \qquad (38)$$

In (37) and (38), upper case letters refer to the 'chain network' and lower case letters to the individual 'stage networks'.

From (38) and Table 1

$$\frac{P_{21}}{P_{12}} = \frac{P_{y1}}{P_{y2}} = \left(\frac{p_{y1}}{p_{y2}}\right)^m = \left(\frac{p_{21}}{p_{12}}\right)^m \qquad (39)$$

where $P_{21}$ and $P_{12}$ are the forward and reverse transfer parameters of the chain network. Also

$$P_{y1} P_{y2} = (p_{y1} p_{y2})^m \qquad (40)$$

$$P_{o1}P_{o2} = p_{o1}p_{o2} = P_{11}P_{22} - P_{12}P_{21} = p_{11}p_{22} - p_{12}p_{21} \qquad (41)$$

where $P_{11}$ and $P_{22}$ are the self parameters of the chain network.

From (24), (25) and (40)

$$g_{lc} = g_{lo} = p_{y1} \, p_{y2} \qquad (42)$$

$$p_{o2} = p_{Gc}; \; p_{o1} = p_{Lc}$$

$$G_{lc} = G_{lo} = P_{y1} \, P_{y2} = (g_{lc})^m = (g_{lo})^m \qquad (43)$$

$$p_{o2} = P_{o2} = P_{Gc} = p_{Gc}$$

$$p_{o1} = P_{o1} = P_{Lc} = p_{Lc}$$

(26) and (43) yields $S = s^m$ (44)

where s is the stability factor of the individual stages and S that of the chain network. From (26)

$$g_{max\ c} \; s = \left|\frac{p_{21}}{p_{12}}\right| \; ; \qquad G_{max\ c} \; S = \left|\frac{P_{21}}{P_{12}}\right| \qquad (45)$$

From (45) , (39), and (44) it follows that

$$G_{max\ c} = (g_{max\ c})^m \qquad (46)$$

In (45) and (46), $g_{max\ c}$ is the MAG of the individual stages and $G_{max\ c}$ that of the chain network.

(46) is capable of a physical interpretation. Each stage is correctly conjugate matched at its ports due to the 'port parameters' of its immediate neighbours, when the source and load ends of the chain network are terminated in conjugate matching parameters appropriate to the end stages (1st and m th) in isolation. As such the power gain of each stage equals its MAG in isolation and the power gain of the chain network the continued product of the individual MAG's.

Having grasped this principle, it may be used for the case of non-identical stages to follow in the next Section, in order to make the MAG of the chain network the maximum possible without recourse to external feedback.

If the interstage networks, considered part of the stages in Fig. 2 do not provide correct matching, but the end terminations are readjusted to retain conjugate matching of the chain network, the available power gain is reduced below the maximum possible; this reduction is accompanied by a corresponding increase in the value of the invariant stability factor of the chain network such that the product of the gain and stability factor is conserved. This conserved product equals the measure of non-reciprocity of the chain network, a factor unaffected by port terminations of the individual stages or chain network (Refer Appendix).

## V. Synchronously Tuned Non-Identical Stages

Consider a chain of m non-identical stages, synchronously tuned to $f_o$ and coupled to each other through lossless reactance two-port networks; let this coupling be such that for a source parameter of the first stage equal to its conjugate matching immittance in isolation, at the input port, and a load parameter of the last stage equal to its conjugate matching immittance in isolation, at the output port, these interstage networks provide conjugate matching in between. This is illustrated in Figs. 3 and 4. As mentioned in Section 2 the terminations ${}^d\rho_G$, ${}^d\rho_L$ are arbitrary and are useful in ensuring stability, controlling bandwidth, stabilising power gain and reducing turns ratio of transformer to a value less than ten.

The turns ratio of an 'ideal' interstage matching transformer between q th and (q + 1)th stages for h or g matrix environments is given by*

$${}_qT_{q+1} = \left\{ \frac{1}{{}_q\rho_{Lc} \;\; {}_{q+1}\rho_{Gc}} \right\}^{\frac{1}{2}} \qquad (47)$$

$$p = h \text{ or } g$$

*(47) and (48) are exact only for 'ideal' transformers; here reactances of windings are very great compared with load impedances. For practical near ideal transformers these equations are approximate. With non-ideal transformers these equations are inapplicable and it is better to check or adjust transformational properties experimentally under operating conditions.

while for z or y matrix environments it is given by

$$^{q}T_{q+1} = \left\{ \frac{^{q+1}\rho_{Gc}}{^{q}\rho_{Lc}} \right\}^{\frac{1}{2}} \qquad (48)$$

$$p = z \text{ or } y$$

where $^{q+1}\rho_{Gc}$ is the real part of the conjugate matched source parameter of the $(q + 1)$th stage in isolation and $^{q}\rho_{Lc}$ the real part of the conjugate matched load parameter of the qth stage in isolation. Fig. 4 shows the detailed arrangements for the qth stage with the device network embedded in h and y environments. By the principle of duality the corresponding arrangements for g and z environments may be obtained; for these cases the labelling of turns as primary and secondary must be interchanged.

For such an arrangement as in Fig. 3, the power gain of each stage in the chain equals its MAG in isolation and the MAG of the chain network equals the continued product of the individual MAG's. Hence

$$G_{\max c} = \prod_{q=1}^{q=m} {^{q}g_{\max c}} \qquad (49)$$

where $G_{\max c}$ is the MAG of the chain network and $^{q}g_{\max c}$ the MAG of the qth stage in isolation.

As shown in the Appendix, the 'measure of non-reciprocity',

$$\left| p_{21} / p_{12} \right| ,$$

of a chain network equals the continued product of the measures of non-reciprocity of the individual stages. Therefore

$$\left| P_{21} / P_{12} \right| = \prod_{q=1}^{q=m} \left| \frac{^{q}p_{21}}{^{q}p_{12}} \right| \qquad (50)$$

From (26), (49) and (50) it follows that the stability factor, S, of the chain network equals the continued product of the stability factors of the individual stages.

$$S = \prod_{q=1}^{q=m} {^{q}s} \qquad (51)$$

For unilateralised stage networks

$$\left| p_{21} / p_{12} \right|$$

and s are infinite but $g_{\max c}$ is finite.(49) is still applicable.

If interstage mismatch exists, but conjugate matching is retained at the ends of the chain network, its MAG is lowered below maximum value whereas S is increased above minimum value; but the product is conserved this being equal to

$$\left| P_{21} / P_{12} \right|$$

of the chain network, a factor unaffected by port terminations of individual stages or chain network.

## VI Alignability of Cascaded Stages

For the situation considered so far, there were resistances and reactances on both sides of each stage. If the frequency of MAG, $f_o$, is to be varied or in the presence of spreads in device parameters, the reactances to be used should be in part variable; it is then necessary to 'align' the amplifier, i.e. to tune the individual circuits as to obtain the maximum power gain for the amplifier at the central frequency $f_o$. If a systematic tuning of circuits or stages from one end to the other and back to the same end in the reverse order renders all the circuits or stages 'tuned' and the available power gain is practically the MAG (i.e. within a small fraction of a db), the amplifier may be said to possess a "good alignability". Good alignability is a basic requirement of amplifiers having two or more tuning circuits.

If the stability factor, s, of each stage in a chain of any number of stages is $\geqslant 10$, each stage is slightly affected by the tuning of its immediate 'neighbours' (one on either side for intermediate stages) but practically unaffected by more distant neighbours. This is because the interactions due to the 'internal loop gain' of two or more stages (i.e. with $S \geqslant 100$) is negligibly small as a consequence of the important theorem of Section III.

The requirement of good alignability imposes a lower limit on the stability factor of each stage ($s \geqslant 10$) and hence an upper limit on

$$g_{\max c} \Big/ \left| \frac{p_{21}}{p_{12}} \right| \quad \text{of } 0.10 .$$

This does not automatically mean a reduced power gain for each stage. An amplifier stage may yield a large power gain with $s \geqslant 10$, a simple example being a graded base transistor in the common emitter configuration at low frequencies ($f \ll f_\alpha$).

The above requirement for s, makes the internal loop gain modulus of each stage less than or equal to a tenth. This figure of 0.10 is a half of that observed by Holmes[9], Stanley and Philip-Jones[10] to avoid excessive skew in gain-frequency response for common emitter transistor amplifier stages with $y_{21}$ real and complex respectively.

## VII Bandwidth of Cascaded Stages

If the individual invariant stability factor of each stage is equal to or greater than ten, in the computation of half power bandwidth, the effect of the 'internal loop-gain' of each stage may be taken into account in its 'total port parameters' with conjugate matching terminations. Assume that the total port parameters are of the form

$$\begin{aligned} P_{p1} &= \rho_{p1}\left[1 + jQ_{p1}\,(f/f_o - f_o/f)\right] \\ P_{p2} &= \rho_{p2}\left[1 + jQ_{p2}\,(f/f_o - f_o/f)\right] \end{aligned} \qquad (52)$$

over the frequencies of interest, $f_o$ being the frequency of MAG. If the stages are coupled as in Fig. 4 (b) and (52) applicable in the appropriate parameter matrix z or y, the Q's of the tuned circuits from left to right of chain (refer Fig.3) are

$${}^{1}Q_{p1},\ ({}^{1}Q_{p2} + {}^{2}Q_{p1}),\ ({}^{2}Q_{p2} + {}^{3}Q_{p1})\ldots\ldots({}^{m-2}Q_{p2} + {}^{m-1}Q_{p1}),\ ({}^{m-1}Q_{p2} + {}^{m}Q_{p1}) \text{ and } {}^{m}Q_{p2} \qquad (53)$$

respectively, where the superscripts indicate the stage number. For the case of Fig. 4 (a), if (52) is applicable, the intermediate tuned circuits can have an appropriate Q only if $Q_{p2} >> Q_{p1}$ (or vice versa). For the case $Q_{p2} >> Q_{p1}$ the Q's of the tuned circuits from left to right are approximately

$${}^{1}Q_{p1},\ ({}^{1}Q_{p2} - {}^{2}Q_{p1}),({}^{2}Q_{p2} - {}^{3}Q_{p1})\ldots\ldots({}^{m-2}Q_{p2} - {}^{m-1}Q_{p1}),\ ({}^{m-1}Q_{p2} - {}^{m}Q_{p1}) \text{ and } {}^{m}Q_{p2} \qquad (54)$$

over the frequencies of interest, an example being a cascaded transistor amplifier in the common base configuration with impedance in series at emitter lead and admittance in shunt across collector base terminals of each stage.

The bandwidth equation for the chain network is given by[6]

$$\frac{\left|{}^{1}p_{21}\,{}^{2}p_{21}\ldots\ldots{}^{m-1}p_{21}\,{}^{m}p_{21}\right|^{2}_{f} \Big/ \left|{}^{1}p_{21}\,{}^{2}p_{21}\ldots\ldots{}^{m-1}p_{21}\,{}^{m}p_{21}\right|^{2}_{f_o}}{[1 + {}^{1}Q_{p1}^{2}\,x^{2}]\,[1 + ({}^{1}Q_{p2} + {}^{2}Q_{p1})^{2}x^{2}]\ldots[1 + ({}^{m-1}Q_{p2} + {}^{m}Q_{p1})^{2}x^{2}]\,[1 + {}^{m}Q_{p2}^{2}x^{2}]} \approx \tfrac{1}{2} \qquad (55)$$

where p = z or y and $x = (f/f_o - f_o/f)$ (56)

Superscripts indicate the stage number, subscript, f, a bandwidth frequency and subscript, $f_o$ the central frequency.

Similarly where $Q_{p2} >> Q_{p1}$ (or vice versa) in the h or g environment the bandwidth equation for the chain networks is

$$\frac{\left|{}^{1}p_{21}\,{}^{2}p_{21}\ldots\ldots{}^{m-1}p_{21}\,{}^{m}p_{21}\right|^{2}_{f} \Big/ \left|{}^{1}p_{21}\,{}^{2}p_{21}\ldots\ldots{}^{m-1}p_{21}\,{}^{m}p_{21}\right|^{2}_{f_o}}{[1 + {}^{1}Q_{p1}^{2}\,x^{2}]\,[1 + ({}^{1}Q_{p2} - {}^{2}Q_{p1})^{2}x^{2}]\ \ldots\ [1 + ({}^{m-1}Q_{p2} - {}^{m}Q_{p1})^{2}x^{2}]\,[1 + {}^{m}Q_{p2}^{2}x^{2}]} \approx \tfrac{1}{2} \qquad (57)$$

where p = h or g with $Q_{p2} >> Q_{p1}$ (or vice versa) and $x = (f/f_o - f_o/f)$. (58)

(55) or (57) as appropriate, reduces to a polynomial equation in $y = f/f_o$, which can be solved for bandwidth by Lin's method[11]. Where there are only two real roots for this equation bandwidth is unambiguously determined. This is so far a wide range of active two-port networks including valve and transistor amplifiers that are synchronously tuned. Only such networks are considered for bandwidth in this paper.

Bandwidth solution is further simplified when the numerator of (55) or (57) as appropriate containing ratios of magnitudes of the forward transfer parameters at a bandwidth and central frequencies, is close to unity. For vacuum device amplifiers, this is true for a wide range of frequencies; for junction transistor amplifiers such an approximation is valid when the bandwidth is narrow or in the case of common base and common emitter amplifiers with series elements at input port and shunt elements at output port (h-environment) when the bandwidth is a small fraction of their respective cut off frequencies ($f_\alpha$ and $f_\beta$ respectively). Assuming one of these cases (55) reduces to

$$\left[1 + {}^{1}Q_{p1}^{2}\, x^{2}\right]\left[1 + ({}^{1}Q_{p2} + {}^{2}Q_{p1})^{2} x^{2}\right]\cdots\left[1 + ({}^{m-1}Q_{p2} + {}^{m}Q_{p1})^{2} x^{2}\right]\left[1 + {}^{m}Q_{p2}^{2}\, x^{2}\right] \simeq 2 \quad (59)$$

where $p = z$ or $y$

Similarly (57) reduces to

$$\left[1 + {}^{1}Q_{p1}^{2}\, x^{2}\right]\left[1 + ({}^{1}Q_{p2} - {}^{2}Q_{p1})^{2} x^{2}\right]\cdots\left[1 + ({}^{m-1}Q_{p2} - {}^{m}Q_{p1})^{2} x^{2}\right]\left[1 + {}^{m}Q_{p2}^{2}\, x^{2}\right] \simeq 2 \quad (60)$$

where $p = h$ or $g$ and ${}^{q}Q_{p2} \gg {}^{q+1}Q_{p1}$

Usually the stages are identical or near identical, the variations being due to device parameter spreads about the average. Even for non-identical stages the input and output port Q's can be made $Q_{p1}$ and $Q_{p2}$ respectively throughout. For such situations (59) and (60) simplify to

$$\left[1 + Q_{p1}^{2} x^{2}\right]\left[1 + (Q_{p2} + Q_{p1})^{2} x^{2}\right]^{m-1}\left[1 + Q_{p2}^{2}\, x^{2}\right] \simeq 2$$

$p = z$ or $y$ (61)

$$\left[1 + Q_{p1}^{2} x^{2}\right]\left[1 + (Q_{p2} - Q_{p1})^{2} x^{2}\right]^{m-1}\left[1 + Q_{p2}^{2}\, x^{2}\right] \simeq 2$$

$p = h$ or $g$; $Q_{p2} \gg Q_{p1}$ (or vice versa) (62)

where the superscripts have been ommitted.

In order to design an individual stage and state the number of such stages required for a chain of near identical stages, given the overall specifications like

$$G_{max\ c},\ f_{o},\ B,\ \rho_{Gc},\ \text{and}\ \rho_{Lc} \quad (63)$$

it is first necessary to obtain an approximate explicit expression for the fractional bandwidth (FBW), $B/\omega_o$, in terms of the Q factors $Q_{p1}$, $Q_{p2}$ and the number of stages, m.

For (62) to be valid $Q_{p2} \gg Q_{p1}$ ( or vice versa). Therefore it can be closely approximated by

$$\left[1 + (Q_{p2} - Q_{p1})^{2} x^{2}\right]^{m} \simeq 2 \quad (64)$$

$p = h$ or $g$; $Q_{p2} \gg Q_{p1}$ (or vice versa)

and

$$\text{FBW} = \frac{B}{\omega_o} = x = \frac{1}{|Q_{p2} - Q_{p1}|}\left\{2^{1/m} - 1\right\}^{\frac{1}{2}} \quad (65)$$

where B is half power angular bandwidth.

If $Q_{p2} \gg Q_{p1}$ (or vice versa), (61) can also be closely approximated to give

$$\text{FBW} = \frac{B}{\omega_o} = \frac{1}{Q_{p2} + Q_{p1}}\left\{2^{1/m} - 1\right\}^{\frac{1}{2}} \quad (66)$$

$p = z$ or $y$

$Q_{p2} \gg Q_{p1}$ (or vice versa)

For the z or y environment use of (66) when $Q_{p2} = Q_{p1}$ gives maximum error; this error decreases rapidly with the number of stages, m, as hown in Fig. 5. If $Q_{p2} > Q_{p1}$ (or vice versa) the error is even further reduced.(65) and (66) are important. Each of them (as appropriate) leads to a simple design procedure for the build up of cascaded amplifiers on a stage by stage basis with even non-unilateral electron devices. For unilateral stage networks the total port parameters reduce to the total self parameters: here $Q_{p1}$ equals $Q_1$ and $Q_{p2}$ equals $Q_2$.

## VIII Conclusions

The MAG of an 'absolutely stable' amplifier stage equals the quotient of its 'measure of non-reciprocity' and 'invariant stability factor'. A stage that is 'pot entially unstable' may be stabilised by adding extra real parts to its self parameters and/or by unilateralising feedback; a well defined MAG may then be realised Extra real parts may be required in some cases to increase bandwidth and/or to reduce the turns ratio of interstage matching transformers when such stages are cascaded.

By suitable additions of passive linear R, L and C elements at the two-ports of a device, the conjugate matching terminations of the modified network may be made positive real and equal thus coinciding with its 'characteristic parameter'. It is then possible to cascade any number of such stages; the MAG of a chain of of m such identical stages equals the MAG of an individual stage raised to power m. So also for the measure of non-reciprocity and invariant stability factor. With mismatch in between, but with conjugate matching terminations at the source and load ends of such a chain network the MAG is lowered and the stability factor increased; this happens in such a manner as to conserve their product which equals the measure of non-reciprocity of the chain network, a factor unaffected by port terminations of individual stages or chain.

Where each lossless interstage network matches the conjugate matching load parameter of the preceding stage, in isolation, with the conjugate matching source parameter of the succeeding stage, in isolation, conjugate matching terminations at the source and load ports of the chain network realises the maximum power gain. The MAG of a chain of m such stages (identical or non-identical) equals the combined product of their individual MAG's; similarly for the measure of non-reciprocity and invariant stability factor. Here too, the product of MAG and stability factor equals the measure of non-reciprocity of the chain network.

If the stability factor of each stage in a chain of stages is equal to or greater than ten. that stage is slightly affected by the tuning of its immediate 'neighbours' but practically unaffected by its more distant neighbours. This restriction on stability factor values ensures that the internal loop gain of each stage be equal to or less than one tenth; the 'alignability' of the tuning circuits is greatly facilitated and the 'skew' in power gain frequency response due to internal feedback practically disappears.

When the magnitude of the square of the continued product of the forward transfer parameters of the stages is nearly constant, the individual invariant stability factors are equal to or greater than ten and the 'total port parameters' of these stages expressible in terms of Q factors $Q_{p1}$, $Q_{p2}$ (identical for stages) the bandwidth of the chain network (compared of m stages whose individual forward transfer parameters are equal or unequal) is simply obtained with good accuracy in terms of $Q_{p1}$, $Q_{p2}$ and m. For unilateralised stages each 'total port parameter' reduces to the corresponding 'total self parameter' viz $Q_{p1}$ equals $Q_1$ and $Q_{p2}$ equals $Q_2$.

The generality of the theory developed, coupled with its simplicity and close accuracy makes it useful in the design of synchronously tuned multistage cascaded amplifiers. This is treated elsewhere.[6]

## IX Acknowledgments

The author is deeply indebted to his research supervisor, Dr. A. R. Boothroyd of the Imperial College of Science and Technology, University of London for his guidance, encouragement and advice. He is also grateful to Mr. R. A. King of the same College for many helpful discussions and criticisms.

The personal support by the Ministry of Aviation, United Kingdom is gratefully acknowledged.

## X References

1. VENKATESWARAN. S., and BOOTHROYD, A.R., "Power Gain and Bandwidth of Tuned Transistor Amplifier Stages", 1959 IEE International Convention on Transistors and Associated Semiconductor Devices, Proc. I.E.E. Vol.106 B Suppl. 15, January, 1960; pp 518-529.

2. CRIPPS, L. J. "Cascaded Amplifiers with Uniform Stability Factors", Mullard Research Laboratories Report 313, August, 1959.

3. LIM MACROBIO., "Power Gain and Stability of Multistage Narrow Band Amplifiers Employing Non Unilateral Electron Devices". IRE Trans, Vol.CT-7 June, 1960; pp 158-166.

4. STERN, A.P., "Considerations on the Stability of Active Elements and Applications to Transistors" IRE National Convention Record, Part 2, 1956; pp 46-52.

5. VENKATESWARAN, S., "An Invariant Stability Factor and its Physical Significance", IEE Monograph 468E dated September, 1961; to be republished in Proc. IEE. Part C, March, 1962.

6. VENKATESWARAN, S., "Stability, Power Gain and Bandwidth of Linear Active Four-Pole Networks, with Particular Reference to Transistor Amplifiers at Higher Frequencies", London University Ph.D thesis; June, 1961

7. te WINKEL, J., "Transmission Line Analogue of a Drift Transistor", Philips Res. Rept., Vol 14, February, 1959; pp 52-64.

8. ROYAL SIGNALS., "Handbook of Line Communication - Vol I" - book, HMSO. London, 1947.

9. HOLMES. D.D., and STANLEY, T.O. "Stability Considerations in Transistor Intermediate Frequency Amplifiers", Transistors I - book RCA, New Jersey, March, 1956; pp 403-421.

10. PHYLIP-JONES, G., "Stability Conditions in Tuned Common-Emitter Transistor Amplifiers", 1959 IEE International Convention on Transistors and Associated Semiconductor Devices, Proc. IEE, Vol. 106 B. Suppl 15, January, 1960; pp 505-517.

11. MURPHY, G.J., "Basic Automatic Control Theory" - book, Van Nostrand, New York, 1957.

## Appendix

### Measure of Non Reciprocity of Chain Network in terms of the Measures of Non Reciprocity of the m Constituent Networks.

The 'A' matrix (associated matrix) of the chain network is the continued product of the 'a' matrices of the constituent networks. Thus for a chain of two networks

$$\left[A_2\right] = \begin{bmatrix} {}^{1}a_{11} & {}^{1}a_{12} \\ {}^{1}a_{21} & {}^{1}a_{22} \end{bmatrix} \qquad \begin{bmatrix} {}^{2}a_{11} & {}^{2}a_{12} \\ {}^{2}a_{21} & {}^{2}a_{22} \end{bmatrix}$$

$$= \begin{bmatrix} {}^{1}a_{11}\,{}^{2}a_{11} + {}^{1}a_{12}\,{}^{2}a_{21} & {}^{1}a_{11}\,{}^{2}a_{12} + {}^{1}a_{12}\,{}^{2}a_{22} \\ {}^{1}a_{21}\,{}^{2}a_{11} + {}^{1}a_{22}\,{}^{2}a_{21} & {}^{1}a_{21}\,{}^{2}a_{12} + {}^{1}a_{22}\,{}^{2}a_{22} \end{bmatrix} \qquad (67)$$

The determinant of this product matrix as obtainable from (67),

$$\Delta_{A_2} = ({}^{1}a_{11}\,{}^{1}a_{22} - {}^{1}a_{12}\,{}^{1}a_{21})\;({}^{2}a_{11}\,{}^{2}a_{22} - {}^{2}a_{12}\,{}^{2}a_{21})$$

$$= \Delta_{1_a}\,\Delta_{2_a} \qquad (68)$$

an. equals the product of the deterimants of the individual factor matrices.

Therefore, by induction

$$\Delta_{A_m} = \Delta_{1_a}\,\Delta_{2_a}\;\ldots\ldots\Delta_{m_a} \qquad (69)$$

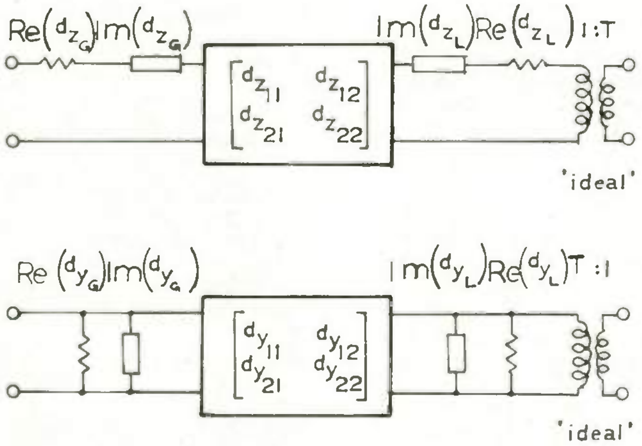
From matrix interrelations

$$\Delta_a = -\frac{h_{12}}{h_{21}} - \frac{z_{12}}{z_{21}} = \frac{y_{12}}{y_{21}} = -\frac{g_{12}}{g_{21}} \qquad (70)$$

Hence from (69) and (70) the ratio of the forward to reverse transfer parameters of a chain of m different stages equals the continued product of the ratios of the same factors of the individual networks.

$$\frac{P_{21}}{P_{12}} = \prod_{q=1}^{q=m} \frac{{}^{q}p_{21}}{{}^{q}p_{12}} \qquad (71)$$

(50) now follows from (71).

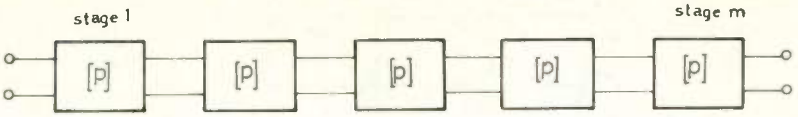


$${}^{d}p_G = {}^{d}\rho_G - j\,{}^{d}\sigma_{11} + j\,{}^{d}\nu\;T^2/2({}^{d}\rho_{11} + {}^{d}\rho_G)$$

$${}^{d}p_L = {}^{d}\rho_L - j\,{}^{d}\sigma_{22} + j\,{}^{d}\nu\;/2({}^{d}\rho_{22} + {}^{d}\rho_L)T^2$$

$$T = \left\{ ({}^{d}\rho_{11} + {}^{d}\rho_G)/({}^{d}\rho_{22} + {}^{d}\rho_L) \right\}^{\frac{1}{2}}$$

Fig. 1. Modified two-port network whose iterative parameters are positive real and equal.



$$P_{11} = \{(P_{01}+P_{02})/[1-(P_{T1}P_{T2})^{m}]\} - P_{02}$$

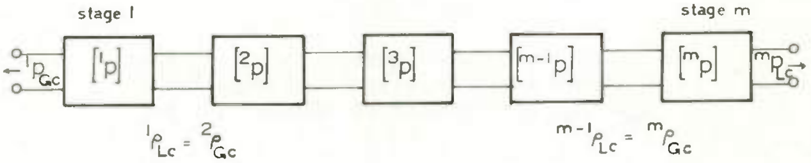
$$P_{22} = \{(P_{01}+P_{02})/[1-(P_{T1}P_{T2})^{m}]\} - P_{01}$$

$$P_{21} = \pm(P_{01}+P_{02})(P_{T1})^{m}/[1-(P_{T1}P_{T2})^{m}]^{\dagger}$$

$$P_{12} = \pm(P_{01}+P_{02})(P_{T2})^{m}/[1-(P_{T1}P_{T2})^{m}]^{\dagger}$$

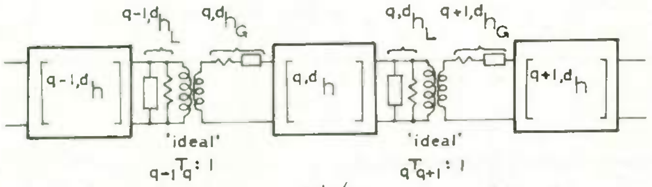
† + sign for z and − sign for y parameters

Fig. 2. A chain of m identical two-port networks; its four-pole parameters are given by above equations.



p = h, z, y or g

Fig. 3. A cascade of m non-identical two-port networks whose MAG equals the continued product of the MAGs of the m networks.

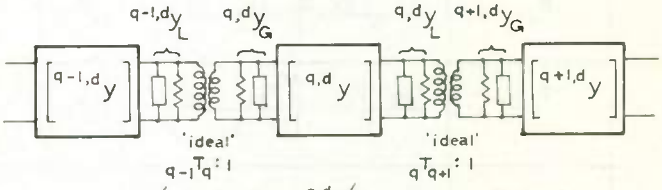


$$\mathrm{Im}({}^{q,d}h_G) = \frac{{}^{q,d}\nu/2}{\mathrm{Re}({}^{q,d}h_{22} + {}^{q,d}h_L)} - \mathrm{Im}({}^{q,d}h_{11})$$

$$\mathrm{Im}({}^{q,d}h_L) = \frac{{}^{q,d}\nu/2}{\mathrm{Re}({}^{q,d}h_{11} + {}^{q,d}h_G)} - \mathrm{Im}({}^{q,d}h_{22})$$

$${}_{q}T^{2}_{q+1} = \frac{1}{{}^{q}h_{Lc}\,{}^{q+1}h_{Gc}} = \frac{1}{\mathrm{Re}({}^{q}h_{Lc})\,\mathrm{Re}({}^{q+1}h_{Gc})}$$

Fig. 4. (a) Part of a chain of non-identical stages arranged to have conjugate match in between for conjugate matched terminations at the ends of chain network: h case. g case is obtained by the principle of duality (with interchange of turns of transformer).



$$\mathrm{Im}({}^{q,d}y_G) = \frac{{}^{q,d}\nu/2}{\mathrm{Re}({}^{q,d}y_{22} + {}^{q,d}y_L)} - \mathrm{Im}({}^{q,d}y_{11})$$

$$\mathrm{Im}({}^{q,d}y_L) = \frac{{}^{q,d}\nu/2}{\mathrm{Re}({}^{q,d}y_{11} + {}^{q,d}y_G)} - \mathrm{Im}({}^{q,d}y_{22})$$

$${}_{q}T^{2}_{q+1} = \frac{{}^{q+1}y_{Gc}}{{}^{q}y_{Lc}} = \frac{\mathrm{Re}({}^{q+1}y_{Gc})}{\mathrm{Re}({}^{q}y_{Lc})}$$

Fig. 4. (b) Part of a chain of non-identical stages arranged to have conjugate match in between for conjugate matched terminations at the ends of chain network: y case. z case is obtained by the principle of duality (with interchange of turns of transformer).

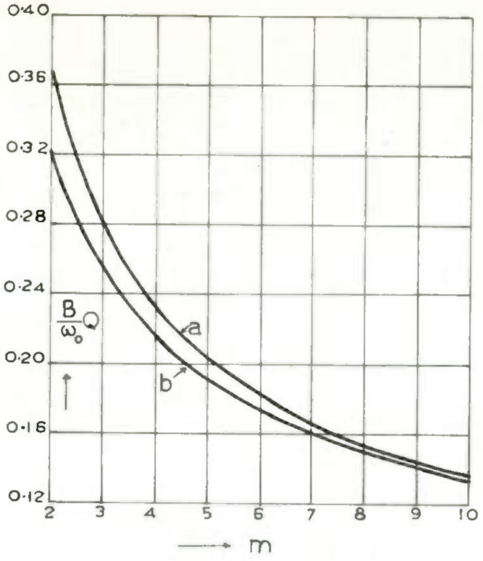


Fig. 5. (a) Exact and (b) approximate solutions for fractional bandwidths of cascaded stages for $Q_{p1} = Q_{p2}$ (maximum error case); p = z or y.

Table 1

Interrelations between four-pole immittances and transmission parameters.

| Transmission parameters | in terms of four-pole immittances | Four-pole immittances | in terms of transmission parameters |
|---|---|---|---|
| $p_{o1}$ | $\frac{p_{11}-p_{22}}{2}+\left\{\left(\frac{p_{11}+p_{22}}{2}\right)^2-p_{12}p_{21}\right\}^{\frac{1}{2}}$ | $p_{11}$ | $\frac{p_{o1}+p_{o2}}{1-p_{\gamma1}p_{\gamma2}}-p_{o2}$ |
| $p_{o2}$ | $\frac{p_{22}-p_{11}}{2}+\left\{\left(\frac{p_{11}+p_{22}}{2}\right)^2-p_{12}p_{21}\right\}^{\frac{1}{2}}$ | $p_{22}$ | $\frac{p_{o1}+p_{o2}}{1-p_{\gamma1}p_{\gamma2}}-p_{o1}$ |
| $p_{\gamma1}$ | $\pm\frac{p_{21}}{[(p_{11}+p_{22})/2]+\{[(p_{11}+p_{22})/2]^2-p_{12}p_{21}\}^{\frac{1}{2}}}{}^{\dagger}$ | $p_{21}$ | $\pm\left(\frac{p_{o1}+p_{o2}}{1-p_{\gamma1}p_{\gamma2}}\right)^{\dagger}p_{\gamma1}$ |
| $p_{\gamma2}$ | $\pm\frac{p_{12}}{[(p_{11}+p_{22})/2]+\{[(p_{11}+p_{22})/2]^2-p_{12}p_{21}\}^{\frac{1}{2}}}{}^{\dagger}$ | $p_{12}$ | $\pm\left(\frac{p_{o1}+p_{o2}}{1-p_{\gamma1}p_{\gamma2}}\right)^{\dagger}p_{\gamma2}$ |
| $p_{o1}-p_{o2}$ | $p_{11}-p_{22}$ | $p_{11}-p_{22}$ | $p_{o1}-p_{o2}$ |
| $p_{o1}p_{o2}$ | $p_{11}p_{22}-p_{12}p_{21}$ | $p_{11}p_{22}-p_{12}p_{21}$ | $p_{o1}p_{o2}$ |
| $p_{\gamma1}/p_{\gamma2}$ | $p_{21}/p_{12}$ | $p_{21}/p_{12}$ | $p_{\gamma1}/p_{\gamma2}$ |

† positive sign for z parameters and negative sign for y parameters

# COUPLED MODE THEORY, WITH APPLICATIONS TO DISTRIBUTED TRANSFORMERS

V. R. Saari
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey

## SUMMARY

The normal modes of uniformly distributed systems with two coupled modes are derived. Several sets of boundary conditions which seem of practical importance have been applied, resulting in network parameters and equivalent circuits. (Some considerations of applicability of the results to ferrite-core distributed transformers are made.) The treatment is extended to balanced, nonuniform systems. An interesting new family of solutions to the differential equation arising in the consideration of nonuniform systems is discussed, and application is made to a practical example (tapered lines of constant characteristic impedance).

## INTRODUCTION

The theory of coupled modes in coupled uniform transmission lines was developed in considerable generality before 1941 by persons working on cross-talk problems in communications systems.[1] This present paper concentrates particularly on two-mode systems, making the method clearer, it is hoped, for those dealing with practical engineering problems. The handling of a few sets of boundary conditions and the derivation of two-port network parameters and equivalent circuits constitute new contributions, as well as the discussion of applicability of these results to such partly distributed structures as bifilar transformer windings on ferrite cores.

Perhaps more generally interesting is the treatment of (balanced) nonuniform systems with the aid of some recently discovered solutions of the linear second-order differential equation with variable coefficients. (It may be mentioned that nonuniform systems can often be handled by a perturbation method developed by B. K. Kinariwala[7] for an analogous problem in time-varying networks.)

## GENERAL SOLUTIONS FOR TWO UNIFORMLY DISTRIBUTED COUPLED LINES

It has long ago been shown[2] that normal modes exist and can be found by straightforward means for any set of uniform coupled transmission lines. No attempt will be made here to apply boundary conditions to systems having more than two propagating modes.

Two coupled modes existing in a uniformly distributed one-dimensional system (Fig. 1) are each characterized by a voltage and a current (or their analogs) which are functions of time only.

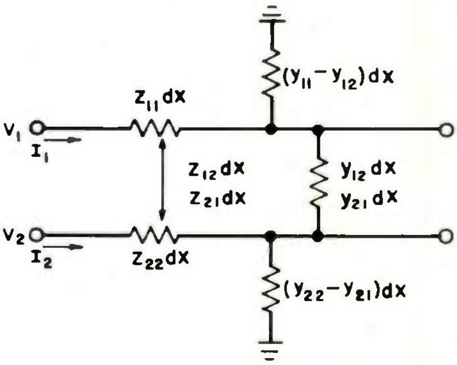


FIG.1 INFINITESIMAL SEGMENT OF TWO-MODE LINE

The differential relations in the frequency domain between these four quantities may be written as follows if there are no other significant interacting modes.

$$\frac{dV_1}{dx} = - z_{11}I_1 - z_{12}I_2 \tag{1}$$

$$\frac{dV_2}{dx} = - z_{21}I_1 - z_{22}I_2 \tag{2}$$

$$\frac{dI_1}{dx} = - y_{11}V_1 + y_{12}V_2 \tag{3}$$

$$\frac{dI_2}{dx} = y_{21}V_1 - y_{22}V_2 \tag{4}$$

Taking a linear combination of the variables, we write

$$\frac{d}{dx}(V_1+mV_2) = -\left(z_{11}+mz_{21}\right)\left(I_1 + \frac{z_{12}+mz_{22}}{z_{11}+mz_{21}} I_2\right) \tag{5}$$

$$\frac{d}{dx}(I_1+nI_2) = -\left(y_{11}-ny_{21}\right)\left(V_1 + \frac{-y_{12}+ny_{22}}{y_{11}-ny_{21}} V_2\right) \tag{6}$$

where it is assumed that m and n do not depend on x.

The new variables $V_m \equiv (V_1+mV_2)$ and $I_n \equiv (I_1+nI_2)$, can be separated when m and n satisfy the following relations:

$$m = \frac{-y_{12}+ny_{22}}{y_{11}-ny_{21}} \tag{7}$$

and

$$n = \frac{z_{12}+mz_{22}}{z_{11}-mz_{21}} \qquad (8)$$

Thus each pair of functions $V_m$ and $I_n$ corresponds to a normal mode of the system. There are two pair of values which satisfy (7) and (8), namely, $(m_1,n_1)$ and $(m_2,n_2)$; and these can be determined very easily.* There is now a new set of four differential equations (5)-(6) which is equivalent to the set (1)-(4). Wave equations in the normal variables are obtained by differentiating (5) and (6). These can easily be integrated, and the familiar telegrapher's equations in each of the four variables are obtained:

$$V_{m_i} = (V_{10}+m_iV_{20}) \cosh \gamma_i x$$

$$- Z_{o_i}(I_{10}+n_iI_{20}) \sinh \gamma_i x \qquad (9)$$

$$I_{n_i} = -\frac{V_{10}+m_iV_{20}}{Z_{o_i}} \sinh \gamma_i x$$

$$+ (I_{10}+n_iI_{20}) \cosh \gamma_i x \qquad (10)$$

where

$$\gamma \equiv \sqrt{(y_{11}-n_iy_{21})(z_{11}+m_iz_{21})}$$

$$= \text{propagation constant} \qquad (11)$$

$$Z_{o_i} = \sqrt{\frac{z_{11}+m_iz_{21}}{y_{11}-n_iy_{21}}}$$

$$= \text{characteristic impedance} \qquad (12)$$

and

$$i = 1,2 \qquad (13)$$

The "arbitrary" constants in Eqs. (9) and (10) have been expressed in terms of the input values $V_{10}$ and $I_{10}$ and the output values $V_{20}$ and $I_{20}$ of the total voltage and total current in the system. This facilitates the application of boundary conditions.

## SYMMETRICAL, NONUNIFORM SYSTEM - A NEW FAMILY OF EXACT SOLUTIONS FOR THE SECOND-ORDER D.E. WITH VARIABLE COEFFICIENTS

It can be shown that the four normal variables of a balanced two-mode system whose electrical parameters vary with distance satisfy second-order nonlinear differential equations of the type**

$$\frac{d^2\theta}{dx^2} - \frac{d}{dx}(\ln u)\frac{d\theta}{dx} - uv\theta = 0 \qquad (14)$$

where the variables and the corresponding functions u(x) and v(x) are listed in the following table:

| k | $\theta_k$ | $u_k$ | $v_k$ | |
|---|---|---|---|---|
| 1 | $V_1-V_2$ | $z_{11}-z_{12}$ | $y_{11}-y_{12}$ | $u_1 = v_3$ |
| 2 | $V_1+V_2$ | $z_{11}+z_{12}$ | $y_{11}+y_{12}$ | $u_2 = v_4$ |
| 3 | $I_1-I_2$ | $y_{11}-y_{12}$ | $z_{11}-z_{12}$ | $u_3 = v_1$ |
| 4 | $I_1+I_2$ | $y_{11}+y_{12}$ | $z_{11}+z_{12}$ | $u_4 = v_2$ |

The corresponding solutions are sometimes expressible in closed form. For example, when the relationship between u and v can be expressed as

$$v = \alpha_i\left(\frac{dg}{dx} + \alpha_i ug^2\right) \qquad (15)$$

where $\alpha_i$ is a constant and g is a function of x, then the general solution for $\theta$ is

$$\theta = C_1e^{\alpha_1\int ug\,dx} + C_2e^{\alpha_2\int ug\,dx} \qquad (16)$$

This can be shown by substituting (16) into (14).

For the case in which g is a constant, we have

$$v = \beta^2u \qquad (17)$$

and

$$\theta = C_1e^{\beta\int u\,dx} + C_2e^{-\beta\int u\,dx} \qquad (18)$$

This special case is that of a constant-impedance tapered line. Boundary conditions are applied below to such a coupled system.

*(It can be shown that for passive systems, $m_1n_2 = m_2n_1 = -1$; and for balanced systems, $m_1 = n_1 = -1$ and $m_2 = n_2 = 1$.)

**This equation can also be written as $\frac{d}{dx}\left(\frac{1}{u}\frac{d\theta}{dx}\right) = v\theta$, which suggest solutions of the form wherein $d\theta/dx$ contains u as a factor.

Another large class of solutions for $\theta$ is obtained by solving (15) for $\alpha_1$ and thus obtaining

$$\alpha_1 = k_1(-1 \pm \sqrt{1+k}) \qquad (19)$$

in which $k_1$ and $k$ are required to be constant with respect to $x$.* The following set of generating relations is thus obtained:

$$ug = \frac{1}{2k_1 g}\frac{dg}{dx} = \frac{1}{2k_1}\frac{d}{dx}(\ln g) \qquad (20)$$

$$v = \frac{k}{2k_1}\frac{dg}{dx} = kug^2 \qquad (21)$$

$$g = \sqrt{\frac{v}{ku}} \qquad (22)$$

The corresponding general solution is**

$$\theta = C_1\left(\frac{v}{u}\right)^{-\frac{1+\sqrt{1+k}}{4}} + C_2\left(\frac{v}{u}\right)^{-\frac{1-\sqrt{1+k}}{4}} \qquad (23)$$

It turns out that this solution is valid as long as u and v satisfy the condition

$$2v = \sqrt{k}\,\frac{d}{dx}\sqrt{\frac{v}{u}} \qquad (24)$$

When uv is constant, the lines are the familiar "exponential lines", in which phase velocity does not depend on the physical dimensions.

## APPLICATION OF BOUNDARY CONDITIONS - DERIVATION OF NETWORK PARAMETERS AND EQUIVALENT CIRCUITS

### 1. Uniformly Distributed Systems

Equations (9)-(10) imply that the normal variables at the remote terminus of a line of length $\ell$ are given by

$$V_{1\ell}+m_1V_{2\ell} = M_1(V_{10}+m_1V_{20}) - N_1Z_{o_1}(I_{10}+n_1I_{21}) \qquad (25)$$

$$I_{1\ell}+n_1I_{2\ell} = -\frac{N_1}{Z_{o_1}}(V_{10}+m_1V_{20}) + M_1(I_{10}+n_1I_{20}) \qquad (26)$$

where

$$M_1 \equiv \cosh \gamma_1\ell \qquad (27)$$

and

$$N_1 \equiv \sinh \gamma_1\ell \qquad (28)$$

It is sometimes more convenient to apply the boundary conditions to the following equivalent set of equations, which express the "initial" values in terms of the "final" values of the variables:

$$I_{10}+n_1I_{20} = M_1(I_{1\ell}+n_1I_{2\ell}) + \frac{N_1}{Z_{o_1}}(V_{1\ell}+m_1V_{2\ell}) \qquad (29)$$

$$V_{10}+m_1V_{20} = Z_{o_1}N_1(I_{1\ell}+n_1I_{2\ell}) + M_1(V_{1\ell}+m_1V_{2\ell}) \qquad (30)$$

It is considered that the "cascade" parameters (also known as "ABCD" parameters) are the most convenient form into which to render the results for each set of boundary conditions. (Multiplying a chain of cascade-parameter matrices corresponds to connecting the respective four-terminal networks in tandem to form a new four-terminal network.) For passive networks, an equivalent circuit based on the cascade parameters is shown in Fig. 2. The voltage transformation ratio $a_{11} \equiv R\,e^{j\theta}$ may be complex, with R representing the ratio of magnitudes and $\theta$ representing phase shift. (For active networks, the circuit is the same except in that the current transformation ratio $\Delta/a_{11}$ no longer equals $1/a_{11}$.)

The term "delay line", will be used to denote a system which incorporates a transmission line in such a way that the input voltage appears between the two terminals at one end of the line and the output voltage appears between the two terminals at the remote end (Fig. 4), as opposed to the "ordinary transformer" connection in which

*This class of solutions was given in a technical memorandum which was circulated by the author at the Bell Telephone Laboratories prior to the time when the same solutions appeared in an article by I. Sugai. The present paper shows these solutions and the corresponding relations between u and v in relatively simple form.

**This method can easily be revised to yield solutions for $\frac{d^2\theta}{dx^2} + s\frac{d\theta}{dx} + t = 0$, where s and t are functions of x. These solutions are

$$\theta = C_3 e^{-\frac{1+\sqrt{1-k}}{\sqrt{k}}\int\sqrt{t}\,dx} + C_4 e^{-\frac{1-\sqrt{1-k}}{\sqrt{k}}\int\sqrt{t}\,dx}$$

under the condition that

$$s = 2\sqrt{\frac{t}{k}} - \frac{d}{dx}\ln\sqrt{t}$$

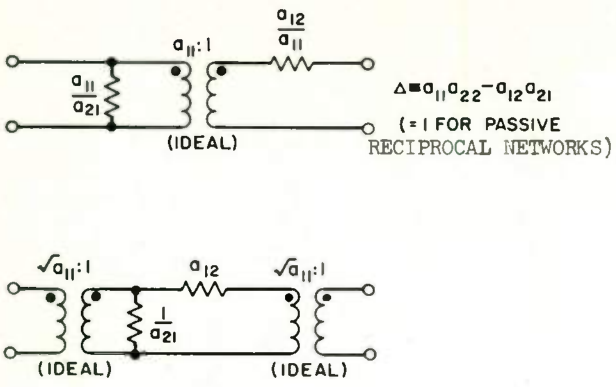


FIG. 2 EQUIVALENT CIRCUIT FOR PASSIVE NETWORKS

the input and output ports are not electrically remote from one another (Fig. 3). "Floating system", will denote a system wherein there is no significant connection to ground at either port (Fig. 4) although there may, in general, be distributed leakage to ground along the line.

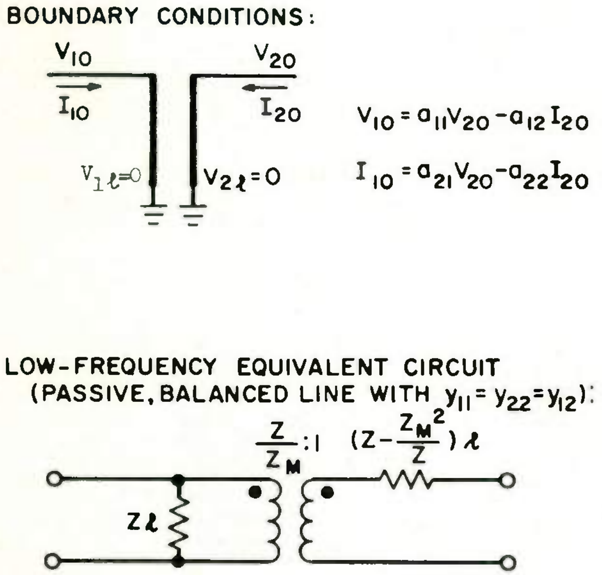

BOUNDARY CONDITIONS:
$V_{10}$
$V_{20}$
$I_{10}$
$I_{20}$
$V_{1\ell}=0$
$V_{2\ell}=0$
$V_{10} = a_{11}V_{20} - a_{12}I_{20}$
$I_{10} = a_{21}V_{20} - a_{22}I_{20}$
LOW-FREQUENCY EQUIVALENT CIRCUIT
(PASSIVE, BALANCED LINE WITH $y_{11} = y_{22} = y_{12}$):
$\frac{Z}{Z_M}:1$
$(Z - \frac{Z_M^2}{Z})\ell$
$Z\ell$

FIG. 3 GROUNDED "ORDINARY TRANSFORMER"

a. Grounded ordinary transformer. Applying the boundary conditions from Figs. 3 to Eqs. (25), we obtain two expressions relating $V_{10}$, $V_{20}$, $I_{10}$ and $I_{20}$. Eliminating $I_{10}$ or $V_{10}$, we obtain relations in which the coefficients are the $a_{ij}$'s. The results are

$$a_{11} = \frac{1}{\Delta}(m_2N_1M_2Z_{01} - m_1N_2M_1Z_{02})$$

$$\rightarrow 1 + \frac{2}{M_1\left(\frac{Z_{11}+Z_M}{N_1Z_{01}}\right)\ell - 1} \xrightarrow{f\rightarrow 0} \frac{Z}{Z_m} \qquad (31)$$

$$a_{12} = \frac{1}{\Delta}N_1N_2(n_2-n_1)Z_{01}Z_{02} \rightarrow \frac{2(Z_{11}+Z_M)N_1Z_{01}\ell}{M_1(Z_{11}+Z_M)\ell - N_1Z_{01}}$$

$$\xrightarrow{f\rightarrow 0} \left(\frac{Z^2}{Z_M} - Z_M\right)\ell \qquad (32)$$

$$a_{21} = \frac{1}{\Delta}M_1M_2(m_2-m_1) \rightarrow \frac{2M_1}{M_1(Z_{11}+Z_M)\ell - N_1Z_{01}}$$

$$\xrightarrow{f\rightarrow 0} \frac{1}{Z_M\ell} \qquad (33)$$

$$a_{22} = \frac{1}{\Delta}(n_2N_2M_1Z_{02} - n_1N_1M_2Z_{01}) \qquad (34)$$

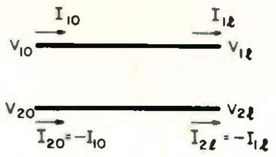
where

$$\Delta = M_1N_2Z_{02} - M_2N_1Z_{01} \rightarrow M_1(Z_{11}+Z_M)\ell - N_1Z_{01} \qquad (35)$$

The first arrow in each expression denotes the value approached when the system becomes balanced and when the distributed strays to ground vanish. The resulting equivalent circuit for the passive, balanced, low-frequency case* without strays to ground is shown in Fig. 3 A similar result for a floating delay line is shown in Fig. 4.

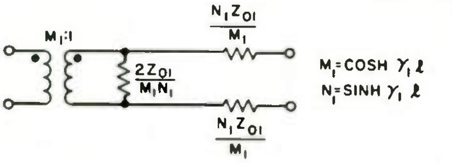
*This simplification is actually valid only if $\gamma\ell \rightarrow 0$. To justify its use when $f \simeq 0$, it is assumed that $\ell < \frac{\lambda}{4}$ and that the Q of the circuit is large.

BOUNDARY CONDITIONS:



$$V_{10}-V_{20}=a_{11}(V_{1\ell}-V_{2\ell})+a_{12}I_{1\ell}$$
$$I_{10}=a_{21}(V_{1\ell}-V_{2\ell})+a_{22}I_{1\ell}$$

EQUIVALENT CIRCUIT (FOR PASSIVE, BALANCED LINE WITH $y_{11}=y_{22}=y_{12}$):



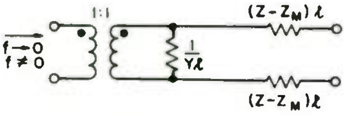


FIG. 4 FLOATING DELAY LINE

b. Two-wire autotransformer Applying the boundary conditions from Fig. 5 to Eqs. (26), we obtain a relationship between $I_{20}$, $I_3$, $V_{10}$, $I_{10}$ and $V_{20}$. Combining this with Eqs. (29), we obtain equations involving only $V_{10}$, $V_{20}$, $I_{10}$ and $I_3$; and $I_{10}$ can be eliminated from these to yield an equation of the form $V_{10} = a_{11}V_{20}+a_{12}I_3$. The network parameters $a_{11}$ and $a_{12}$ are thus obtained directly from this. (The expressions must be simplified using the trignometric identity $\cosh^2 x - \sinh^2 x = 1$ before reducing to the forms shown below.) If, instead of eliminating $I_{10}$, we eliminate $V_{10}$, we obtain the form $I_{10} = a_{21}V_{20}+a_{22}I_3$. The results can be shown to be

$$a_{11} = \frac{1}{A_2-A_1}(m_1A_1-m_2A_2-B_1+B_2) \qquad (36)$$

$$a_{12} = \frac{1}{A_2-A_1}(E_1N_2Z_{02}-E_2N_1Z_{01}) \qquad (37)$$

$$a_{21} = \frac{1}{A_2-A_1}\left[M_1F_2-M_2F_1 + \frac{n_1n_2}{n_2-n_1}(D_2+D_1-F_1-F_2) + \frac{N_1N_2}{n_2-n_1}\left(n_1^2\frac{Z_{01}}{Z_{02}} - n_2^2\frac{Z_{02}}{Z_{01}}\right)\right] \qquad (38)$$

$$a_{22} = \frac{1}{A_2-A_1}(M_2E_1-M_1E_2) \qquad (39)$$

where

$$A_1 \equiv N_2(M_1+n_2)Z_{02} \qquad A_2 \equiv N_1(M_2+n_1)Z_{01} \qquad (40)$$

$$B_1 \equiv (M_1n_2+1)N_2Z_{02} \qquad B_2 \equiv (M_2n_1+1)N_1Z_{01} \qquad (41)$$

$$D_1 \equiv M_2(M_1+m_2) \qquad D_2 \equiv M_1(M_2+m_1) \qquad (42)$$

$$F_1 \equiv M_1m_2+1 \qquad F_2 \equiv M_2m_1+1 \qquad (43)$$

$$E_1 \equiv n_1N_1Z_{01} \qquad E_2 \equiv n_2N_2Z_{02} \qquad (44)$$

If the system is passive and contains no distributed leakage to ground, it can be shown $(m \equiv m_2)$ that

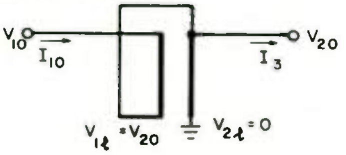
$$a_{11} = \frac{1}{\Delta}\left[N_1Z_{01}(m-1)^2 + 2mN_2Z_{02}(M_1+1)\right] \rightarrow 2 \qquad (45)$$

$$a_{12} = \frac{1}{\Delta}\left[N_1N_2(m+1)Z_{01}Z_{02}\right] \rightarrow \frac{2N_1Z_{01}}{M_1+1} \qquad (46)$$

$$a_{21} = -\frac{m}{\Delta}\left[M_1(1-m) - 2 - \frac{2(M_1-1) + mN_1\frac{N_2Z_{02}}{Z_{01}}}{1+m}\right]$$

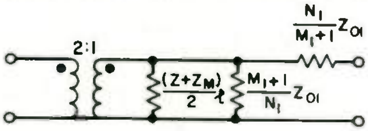
$$\rightarrow \frac{M_1+1 + \frac{N_1N_2Z_{02}}{2Z_{01}}}{(M_1+1)N_2Z_{02}} \qquad (47)$$

BOUNDARY CONDITIONS:



$$I_3 = I_{1\ell}-I_{20}$$
$$V_{10}=a_{11}V_{20}+a_{12}I_3$$
$$I_{10}=a_{21}V_{20}+a_{22}I_3$$

EQUIVALENT CIRCUIT (PASSIVE, BALANCED LINE WITH $y_{11}=y_{22}=y_{12}$):



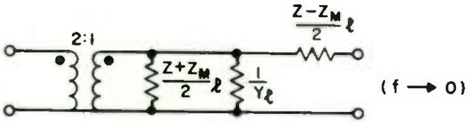


FIG. 5 TWO-WIRE AUTOTRANSFORMER

$$a_{22} = \frac{1}{\Delta}\left[N_1Z_{01}+mM_1N_2Z_{02}\right] \rightarrow \frac{N_1Z_{01}+M_1N_2Z_{02}}{(M_1+1)N_2Z_{02}} \qquad (48)$$

where

$$\Delta = -\,m(A_2-A_1) = mN_2Z_{02}(M_1+1) - N_1Z_{01}(m-1)$$

$$\rightarrow N_2Z_{02}(M_1+1) \qquad (49)$$

The arrows indicate the results when the system also becomes balanced ($m \rightarrow 1$). The elements of the equivalent circuit (Fig. 5) for this case are

$$\frac{a_{12}}{a_{11}} \rightarrow \frac{N_1Z_{01}}{M_1+1} \qquad (50)$$

and

$$\frac{a_{21}}{a_{11}} \rightarrow \frac{1}{2N_2Z_{02}} + \frac{N_1}{4(M_1+1)Z_0} = \frac{1}{2(Z+Z_M)\ell} + \frac{N_1}{4(M_1+1)Z_0} \qquad (51)$$

A similar equivalent circuit for a three-wire autotransformer is shown in Fig. 6.

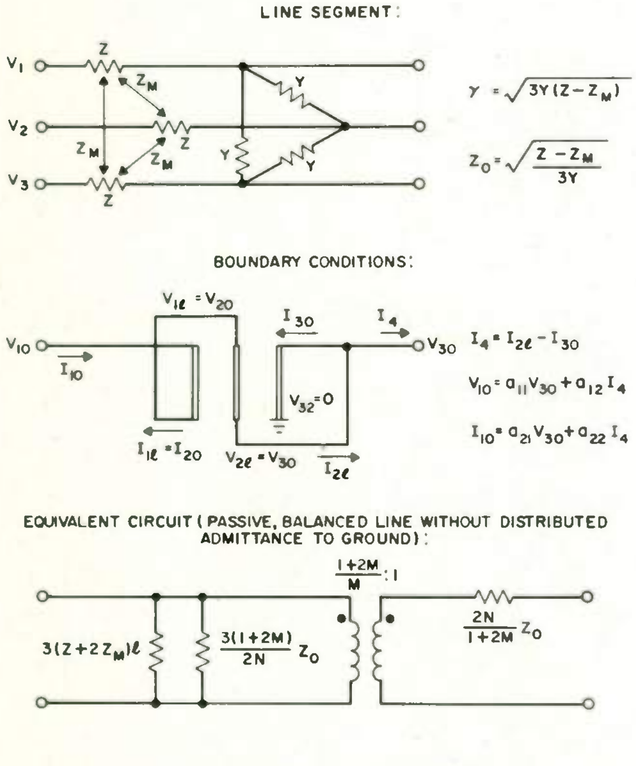


FIG. 6 THREE-WIRE AUTOTRANSFORMER

2. A Nonuniformly Distributed System

The normal variables on a nonlinear two-mode line such as has been discussed above may be expressed as functions of distance x as follows:

$$\theta_k \equiv C_{k1}e^{\alpha_{k1}\int_0^x u_k g_k dx} + C_{k2}e^{\alpha_{k2}\int_0^x u_k g_k dx} \qquad (52)$$

where

$$k = 1,2,3,4 \qquad (53)$$

and

$$\begin{aligned} \theta_1 &= V_1-V_2 \equiv \varphi_3 \\ \theta_2 &= V_1+V_2 \equiv \varphi_4 \\ \theta_3 &= I_1-I_2 \equiv \varphi_1 \\ \theta_4 &= I_1+I_2 \equiv \varphi_2 \end{aligned} \qquad (54)$$

The arbitrary constants may be written in terms of the terminal values of the variables (the added subscript $\ell$ always denoting the particular value occurring at the end of the line):

$$C_{k1} = \frac{\alpha_{k2}\theta_{k\ell} + \frac{1}{g_{k\ell}}\varphi_{k\ell}}{-\,M_{k1}(\alpha_{k2}-\alpha_{k1})} \qquad (55)$$

$$C_{k2} = \frac{\alpha_{k1}\theta_{k\ell} + \frac{1}{g_{k\ell}}\varphi_{k\ell}}{-\,M_{k2}(\alpha_{k2}-\alpha_{k1})} \qquad (56)$$

where

$$M_{kj} \equiv e^{\alpha_{kj}\int_0^\ell u_k g_k dx} \qquad (57)$$

In the special case of Eqs. (17)-(18), where $u_k$ and $v_k$ are proportional for all k, we have

$$v_k = \beta_{kj}^2 u_k \qquad (58)$$

It can now be shown that

$$\beta_{11} = \sqrt{\frac{v_1}{u_1}} = \sqrt{\frac{y_{11}-y_{12}}{z_{11}-z_{12}}} = \text{constant} \qquad (59)$$

$$\beta_{12} = \frac{1}{\beta_{32}} = -\beta_{11} = -\frac{1}{\beta_{31}} \qquad (60)$$

$$\beta_{21} = \sqrt{\frac{v_2}{u_2}} = \sqrt{\frac{y_{11}+y_{12}}{z_{11}+z_{12}}} = \text{constant} \qquad (61)$$

$$\beta_{22} = \frac{1}{\beta_{42}} = -\beta_{21} = -\frac{1}{\beta_{41}} \qquad (62)$$

$$V_1-V_2 = C_{11}e^{\beta_{11}\int_0^x (z_{11}-z_{12})dx} + C_{12}e^{-\beta_{11}\int_0^x (z_{11}-z_{12})dx} \qquad (63)$$

$$V_1+V_2 = C_{21}e^{\beta_{21}\int_0^x (z_{11}+z_{12})dx} + C_{22}e^{-\beta_{21}\int_0^x (z_{11}+z_{12})dx} \qquad (64)$$

$$I_1-I_2 = C_{31}e^{\beta_{11}\int_0^x (z_{11}-z_{12})dx} + C_{32}e^{-\beta_{11}\int_0^x (z_{11}-z_{12})dx} \qquad (65)$$

$$I_1+I_2 = C_{41}e^{\beta_{21}\int_0^x (z_{11}+z_{12})dx} + C_{42}e^{-\beta_{21}\int_0^x (z_{11}+z_{12})dx} \qquad (66)$$

The constants $C_{kj}$ can be obtained easily from Eqs. (55) and (56); for example,

$$C_{11} = \frac{\beta_{11}(V_{1\ell}-V_{2\ell}) - (I_{1\ell}-I_{2\ell})}{2\beta_{11}e^{\beta_{11}\int_0^\ell (z_{11}-z_{12})dx}} \qquad (67)$$

$$C_{12} = \frac{\beta_{11}(V_{1\ell}-V_{2\ell}) + (I_{1\ell}-I_{2\ell})}{2\beta_{11}e^{-\beta_{11}\int_0^\ell (z_{11}-z_{12})dx}} \qquad (68)$$

Let us calculate the network parameters for the case of a floating delay line (boundary conditions of Fig. 4). From Eq. (63), we obtain

$$V_{10}-V_{20} = C_{11}+C_{12} \qquad (69)$$

$$V_{10}-V_{20} = \frac{1}{2}\left(\frac{1}{M_{11}} + M_{11}\right)\left(V_{1\ell}-V_{2\ell}\right) + \frac{1}{2\beta_{11}}\left(M_{11} - \frac{1}{M_{11}}\right)\left(2I_{1\ell}\right) \qquad (70)$$

Hence, for any passive balanced nonlinear two-mode line on which the distributed $z_{ij}$'s and $y_{ij}$'s are proportional to one another, we have, for delay-line type external connection,

$$a_{11} = \frac{1}{2}\left(\frac{1}{M_{11}} + M_{11}\right) = a_{22} \qquad (71)$$

$$a_{12} = \frac{1}{\beta_{11}}\left(M_{11} - \frac{1}{M_{11}}\right) = \frac{a_{11}^2-1}{a_{21}} \qquad (72)$$

It remains, now, to consider a particular functional form for $u_1(x)$ in order that we may calculate a typical $M_{11}$. Let us consider that the taper is sinusoidal, with a minimum $|u_1|$ at $x = 0$ and a maximum at $x = \ell$; that is,

$$u_1 = ju_0(1 + a \sin bx) = \frac{1}{\beta_{11}^2} v_1 = Z_{11}-Z_{12} = \frac{Y_{11}-Y_{12}}{\beta_{11}^2} \qquad (73)$$

From (57),

$$M_{11} = e^{\beta_{11}\int_0^{\pi/2b} ju_0(1 + a \sin bx)dx} \qquad (74)$$

$$M_{11} = e^{j\frac{u_0\beta_{11}}{b}\left(a + \frac{\pi}{2}\right)} \qquad (75)$$

Thus, given $u_0$, a, b, and $\beta_{11}$, the network parameters are easily obtained.

## PRACTICAL CONSIDERATIONS - SEMI-DISTRIBUTED TRANSFORMERS

### 1. Self-Capacitance of Windings and Core Response

Transformers consisting of a parallel or twisted pairs of wires wound on ferrite cores[5] have recently become popular for use in high-frequency applications such as interstages in transistor IF amplifiers. Such a transformer is shown in Fig. 7 with a toroidal core. The frequency range over which the equivalent circuits given earlier are applicable can be extended upward by introducing the following consideration. There is a time delay (in the realization of full self and mutual inductance) which is associated with the time required for a wave to propagate

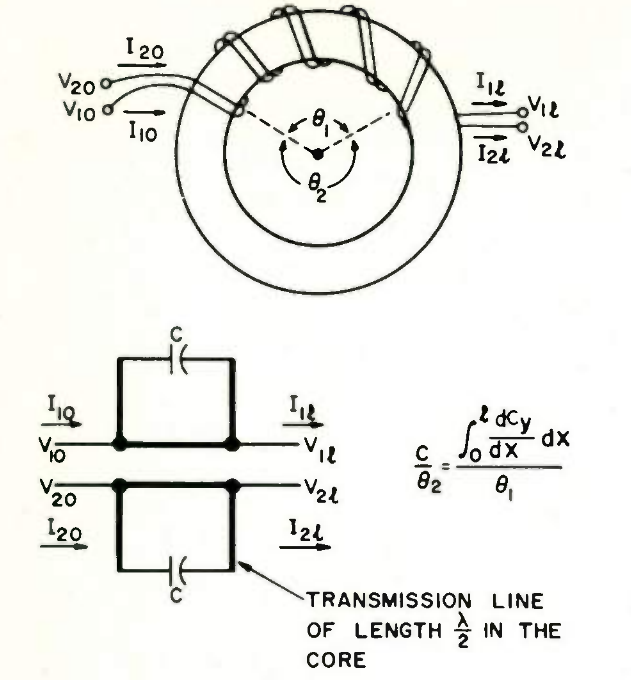


FIG.7 BIFILAR WINDINGS ON MAGNETIC CORE

around the core. This should suggest the inclusion of capacitance in shunt with the winding inductances. The capacitance should resonate with the corresponding inductance at a frequency for which the core length is a multiple of one wavelength. Actually, part of this capacitance is distributed along with the inductance; while the other part can be represented by a transmission line of length $\lambda/2$ at the first resonance* terminated in a capacitance. The interconnections are shown in Fig. 7.

The analysis of the circuit which consists of the windings and the portion of the core under them proceeds as has been shown above. The impedance parameters $Z_{ij}$ of the line must, of course, contain the effect of the distributed self capacitance. It can be shown that the effect of this capacitance (or of a more general shunt self-impedance element) is simply to multiply each element of the impedance matrix by a current-splitting factor q (Fig. 8). This factor is uniform only when just one of the two normal modes is excited in the system. The magnitude of the characteristic value of q for the difference (transverse) mode is less than unity; whereas the magnitude for the sum (longitudinal) mode is treater than unity. It may be shown that for a balanced system, these two characteristic values are, respectively,

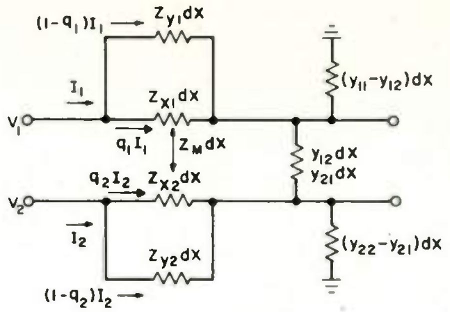


FIG.8 SYSTEM SEGMENT INCLUDING SHUNT SELF IMPEDANCE

$$q_d = \frac{Z_y}{Z_x+Z_y-Z_M} \quad \text{and} \quad q_s = \frac{Z_y}{Z_x+Z_y+Z_M} \tag{76}$$

Normally, the relation

$$Z_{x1}Z_{y2} \approx Z_{x2}Z_{y1} \tag{77}$$

is satisfied, so the values of q for an unbalanced system become

$$q_d \approx \frac{\sqrt{Z_{y1}Z_{y2}}}{\frac{Z_{x2}}{Z_{x1}}(Z_{x1}+Z_{y1}) - Z_M} \tag{78}$$

$$q_s \approx \frac{\sqrt{Z_{y1}Z_{y2}}}{\frac{Z_{x2}}{Z_{x1}}(Z_{x1}+Z_{y1}) + Z_M} \tag{79}$$

Equations (11)-(12) for the difference mode become

$$\gamma_1 \equiv \gamma_d = \sqrt{(y_{11}-n_1y_{21})(Z_{x1}+m_2Z_M)q_d} \tag{80}$$

and

$$Z_{01} \equiv Z_{0d} = \sqrt{\frac{(Z_{x1}+m_1Z_M)q_d}{y_{11}+n_1y_{21}}} \tag{81}$$

(The values of m and n are not changed by the factors q.) A similar set of relations hold for the sum mode.

### 2. Multiple-Loop Cores and Effective Turns Ratios

In order to achieve minimum leakage inductance and maximum uniformity of parameter distribution in a transformer it is desirable to use a toroidal core. However, it is simpler, from a manufacturing standpoint, to wind the coils on a

*Velocity of propagation in the core varies as $1/\sqrt{\mu\varepsilon}$; electrical length varies as $\sqrt{\mu\varepsilon}$; impedance varies as $\sqrt{\mu/\varepsilon}$.

bobbin and slip the bobbin on one leg of a rectangular core. A multiple-loop core is often chosen. It should be remembered that the manner in which the windings are connected to the external terminals will significantly affect the characteristics of the transformer even at low frequencies if the number of turns in the windings is small. For example, a two-wire autotransformer on a two-loop core will not have precisely a 2:1 ratio if each wire dces not thread both loops as many times as the other. This is true because each of the several legs branching away from the one which holds the bobbin carries only part of the total magnetic flux.

## CONCLUSION

The normal modes of uniformly distributed systems with two coupled modes were derived. Several sets of boundary conditions which seem of practical importance have been applied, resulting in network parameters and equivalent circuits. (Some considerations of applicability of the results to ferrite-core distributed transformers were made.) The treatment was extended to balanced, nonuniform systems. An interesting new family of solutions to the differential equation arising in the consideration of nonuniform systems was discussed, and application was made to a practical example (tapered lines of constant characteristic impedance).

## ACKNOWLEDGMENTS

Several colleagues have given the author helpful suggestions, among them being B. K. Kinariwala, I. Dostis and C. H. Tang.

## REFERENCES

1. S. O. Rice, "Steady-State Solution of Transmission Line Equations", BSTJ, Vol. 20, April, 1941, pp. 131-178.

2. J. R. Carson and R. S. Hoyt, "Propagation of Periodic Currents over a System of Parallel Wires", BSTJ, July, 1927, pp. 495-545.

3. H. G. Rudenberg, "The Distributed Transformer", Research Division, Raytheon Mfg. Co., Waltham, Massachusetts (April, 1952).

4. L. A. Pipes, "Computation of the Impedances of Nonuniform Lines by a Direct Method", AIEE Transactions, Vol. 75, Part I, November, 1956, pp. 551-554.

5. C. L. Ruthroff, "Some Broadband Transformers", Proc. IRE, Vol. 47, August, 1959, pp. 1337-1342.

6. W. H. Louisell, "Coupled-Mode and Parametric Electronics", Wiley (1960).

7. B. K. Kinariwala, "Analysis of Time-Varying Networks", 1961 IRE International Convention Record, Part 4, pp. 268-276.

8. C. H. Tang, "Optimization of Waveguide Tapers Capable of Multimode Propagation", IRE Trans. PGMTT, September, 1961, pp. 442-452.

9. I. Sugai, "A Generalized Hildebrand's Method for Nonuniform Transmission Lines", Proc. IRE, Vol. 49, December, 1961, p. 1944.

# TRANSIENT ANALYSIS OF A PARAMETRIC OSCILLATOR*

Edwin D. Banta
General Atronics Corporation
West Conshohocken, Pa.

## Summary

The parametric oscillator is investigated for use as a phase-synchronized oscillator; for such use the buildup rate, the phase errors occuring during buildup, and the final amplitude are all of interest. The paper begins with the general circuit equations of a three tank parametric device and shows a derivation of three first order, nonlinear differential equations specifying its performance. The only restriction is in the form assumed for the nonlinear capacity (this restriction could be removed, in principle, at the cost of more elaborate mathematics). By use of "coupled mode" theory these three equations are simplified into three nonlinear, first order differential equations which do not involve time explicitly.

These equations are applied first to the small signal case to study the effects of loading and detuning upon the signal buildup. A general expression is obtained for the buildup rate, $\alpha$, and several special cases receive detailed discussion. The principal result is that $\alpha$ is proportional to the square root of the idler frequency and inversely proportional to the geometric mean circuit capacity.

These small signal equations are next used in an analysis of the signal phase error accrued during buildup. Phase error is defined as the unpredictable portion of the oscillator phase relative to the phase of a synchronizing signal. It is found that by virtue of its extra degree of freedom, the idler tank, the phase error is less than occurs in a conventional oscillator operating under the same conditions.

The report continues with a consideration of the steady state configuration of the parametric device. This is possible since the differential equations are not limited to the small signal approximation. Explicit formulae for the steady state voltages in each tank are obtained resulting in the Manley-Rowe equations. These equations show that the idler frequency should be small to obtain maximum output at the signal frequency. It was stated previously that the idler frequency must be large to maximize the buildup rate and to reduce phase errors. Thus a design compromise is necessary.

In conclusion, an approximate buildup analysis showing saturation is presented. Interestingly, the form of buildup is the same as that for the Van Der Pol oscillator.

## Introduction

The analysis of parametric devices separates into two classes: in the first and more extensive class it is assumed that the parametric agent is a time varying element; this follows from the small signal approximation; i.e., the pump voltage dominates a nonlinear element which, in turn, exhibits a time varying characteristic determined by the pump voltage. In the second class the coupled, nonlinear circuit equations are kept; however, most of the analysis in this class concerns power relationships, not transient behavior.

The present paper analyzes the transient parametric oscillator by means of the "coupled mode" approach. This method has the advantage of retaining, in its first approximation, the nonlinear dependence needed for a meaningful discussion of saturation effects. Finally, while only the three tank parametric oscillator is described in detail, the method is applicable to parametric devices with any number of tuned circuits.

## General Circuit Equations

The simplest nondegenerate parametric device contains three separate tuned circuits, called the signal, idler, and pump. A typical circuit using a semiconducting diode as a nonlinear reactance is shown in Figure 1. At time t=0 the switch is closed and the signal source is disconnected from the signal tank, which is left with stored energy in its elements. A principal part of this analysis is devoted to relating these initial conditions in the signal tank, as well as any initial conditions in the idler tank, to the final state of the oscillator.

For mathematical convenience the diode resistance $R_D$ is assumed to be absorbed into the individual tank loadings; similarly, the fixed diode capacity is transferred into the $C_K$'s, leaving only the variable capacity for $C_D(V)$. Further, since the primary effort is to be a nonlinear analysis, it seems justifiable to

*This work was sponsored by Rome Air Development Center under Contract AF 30-(602)-2283.

choose the simplest $C_D(V)$ which gives parametric action. This form is

$$C_D(V) = \frac{C_o}{V_o} V \qquad (1)$$

Mode Equations. With the circuit of Figure 1 and the assumptions made above the circuit equations are

$$L_K\dot{I}_{K1} + R_K(I_{K1}-I_K) = V_p\delta_{K3} \qquad (2a)$$

$$\frac{d}{dt}[C_KV_K] = I_K-I \qquad (2b)$$

$$V_K = -L_K\dot{I}_{K1} + V_p\delta_{K3} \qquad (2c)$$

$$I = \frac{d}{dt}[VC_p(V)] = 2\frac{C_o}{V_o}V\dot{V} \qquad (2d)$$

where K ranges over 1-3 and $\delta_{K3}=1$ for K=3 and zero otherwise. These equations can be combined to eliminate I and $I_{K1}$ with the result

$$\dot{I}_K + G_K\dot{V}_K = -\frac{V_K}{L_K} + \frac{V_p\delta_{K3}}{L_K} \qquad (3a)$$

$$\dot{V}_K = \frac{I_K}{C_K} - 2\frac{C_o}{C_K}V\dot{V} \qquad (3b)$$

in which $G_K = 1/R_K$.

These two real, first order differential equations can be combined into one complex, first order differential equation as follows:[1] multiply Eq. (3b) by $\lambda_K$ and add it to Eq. (3a).

$$\frac{d}{dt}[I_K+G_KV_K+\lambda_KV_K] =$$

$$\frac{\lambda_K}{C_K}[I-\frac{C_K}{\lambda_K}\frac{V_K}{L_K}] - 2\frac{C_o\lambda_K}{C_K}V\dot{V} + \frac{V_P\delta_{K3}}{L_K} \qquad (3c)$$

Now choose $\lambda_K$ such that $G_K\lambda_K = -\frac{C_K}{\lambda_KL_K}$, i.e.

$$\frac{\lambda_K}{C_K} = i\omega_K\left\{\sqrt{1-\frac{1}{4Q_K^2}} + \frac{i}{2Q_K}\right\} = i\omega_K' \qquad (4)$$

in which $\omega_K$ and $Q_K$ have their usual meanings. Finally, let

$$\frac{a_K(t)}{\sqrt{L_K}} = I_K+(G_K+\lambda_K)V_K = I_K - \frac{C_K}{\lambda_K}\frac{V_K}{L_K} \qquad (5)$$

whence Eq. (3c) becomes the mode equation

$$a_K = i\omega_K'a_K - i\omega_K'\sqrt{L_K}\frac{C_o}{V_o}\frac{dV^2}{dt} + \frac{V_{Po}\delta_{K3}}{\sqrt{L_K}} \qquad (6)$$

In terms of this variable

$$V_K = \text{Im}\frac{a_K}{\sqrt{C_K(1-\frac{1}{4Q_K^2})}} \qquad (7a)$$

$$I_K = \text{Re}\frac{a_K}{\sqrt{L_K}} - V_K\frac{G_K}{2} =$$

$$\text{Re}\frac{a_K}{\sqrt{L_K}} - \text{Im}\frac{a_KG_K}{C_K(1-\frac{1}{4Q_K^2})} \qquad (7b)$$

and

$$V^2 = [\sum_K \frac{\omega_K\sqrt{L_K}}{\sqrt{1-\frac{1}{4Q_K^2}}}\text{Im}\ a_K]^2 \qquad (8)$$

"Coupled Mode" Method. To begin a solution of Eq. (6) let

$$a_K = A_Ke^{i\omega_{K1}t} \qquad (9)$$

where $\omega_{K1}$ is the actual radian carrier frequency of oscillation in the Kth tank; thus Eq. (6) becomes

$$\dot{A}_K+i(\omega_{K1}-\omega_K')A_K =$$

$$[-i\omega_K'\sqrt{L_K}\frac{C_o}{V_o}\frac{dV^2}{dt} - \frac{V_p\delta_{K3}}{\sqrt{L_K}}]e^{-i\omega_{K1}t} \qquad (10)$$

The "coupled mode" method argues that if $A_K$ varies slowly, the only significant terms on the right hand side of Eq. (10) are the slowly varying ones. Thus, by virtue of the term

$$e^{-i\omega_{K1}t},$$

the only significant terms in the bracket expression are those with a factor

$$e^{i\omega_{K1}t}.$$

In order to decide what terms these are, a relationship between the three frequencies $\omega_1$, $\omega_2$ and $\omega_3$ is needed. It turns out that for parametric oscillation to occur it is necessary that

$$\omega_1 + \omega_2 = \omega_3 + \Delta \qquad (11)$$

where $\Delta$ is small compared to any $\omega_K$ and it represents the error in circuit adjustment.

By use of the "coupled mode" method, Eq. (11) and an assumed pump voltage of the form

$$V_p = V_{po} \cos(\omega_3 t + \phi_p) \qquad (12)$$

it is found that the appropriate equations for the $A_K$'s are

$$\dot{A}_1 + [\alpha_1 - \frac{i\Delta}{2}]A_1 = (\omega_1 - \frac{\Delta}{2})\eta_1 A_3 A_2{}^* \qquad (13a)$$

$$\dot{A}_2 + [\alpha_2 - \frac{i\Delta}{2}]A_2 = (\omega_2 - \frac{\Delta}{2})\eta_2 A_3 A_1{}^* \qquad (13b)$$

$$\dot{A}_3 + [\alpha_3]A_3 = -(\omega_3)\eta_3 A_1 A_2 + \frac{V_{po} e^{i\phi_p}}{2\sqrt{L_3}} \qquad (13c)$$

where $\eta_K = \eta(1 + i\frac{\alpha_K}{\omega_K})$

$$\alpha_K = \frac{\omega_K}{2Q_K}$$

$$\eta = \frac{1}{\sqrt{C_1 C_2 C_3}} \cdot \frac{C_o}{2V_o}$$

(Certain small terms involving $\dot{A}_K$ have been deleted from the right hand side of these equations.) In addition, it has been assumed that Δ is due entirely to errors in the signal and idler frequencies, not in the pump frequency; this condition usually applies in practice.

## Small Signal Theory

Small signal theory implies that $A_3$ is a constant independent of $A_1$ and $A_2$. Thus

$$A_3 = Q_3 \sqrt{C_3}\, V_{po} e^{i\phi_p} \qquad (14)$$

While Eqs. (13a) and (13b) can be combined and written in operational form as

$$\Big\{(D+\alpha_1 - \frac{i\Delta}{2})(D+\alpha_2 + \frac{i\Delta}{2}) - \eta_1 \overset{*}{\eta}_2 \omega_1 \omega_2 (1- \frac{\Delta}{2\omega_1})(1- \frac{\Delta}{2\omega_2})|A_3|^2\Big\}A_1 = 0 \qquad (15)$$

This equation is solved by assuming a solution of the form $Ke^{\alpha t}$ whence it is found that

$$\alpha = - \frac{\alpha_1+\alpha_2}{2} \pm \sqrt{\alpha_o{}^2(1- \frac{\Delta}{2\omega_1})(1- \frac{\Delta}{2\omega_2}) \cdot (1+ \frac{i\alpha_1}{\omega_1})(1- \frac{i\alpha_2}{\omega_2}) + (\frac{\alpha_1 - \alpha_2 - i\Delta}{2})^2} \qquad (16)$$

where

$$\alpha_o = \frac{Q_3 V_{po} C_o}{2V_o} \sqrt{\frac{\omega_1 \omega_2}{C_1 C_2}} \qquad (17)$$

is the buildup constant for the ideal case of no loading and no detuning.

Ideal Case. When there is no detuning and no loading the buildup rate is given by $\alpha_o$, which shows the functional dependence of buildup upon the circuit parameters. It is important to note that for rapid buildup the signal and idler capacities should be as small as possible, and if all other parameters are fixed, the idler frequency should be as high as possible. A conflicting reason for limiting $\omega_2$ will be discovered in the discussion of power distribution between the signal and idler.

Detuning - No Losses. In this case $\alpha_1 = \alpha_2 = 0$ and Eq. (16) becomes

$$\alpha = \pm\sqrt{\alpha_o{}^2(1- \frac{\Delta}{2\omega_1})(1- \frac{\Delta}{2\omega_2}) - (\frac{\Delta}{2})^2} \qquad (18)$$

This quantity is clearly quadratic in Δ so that with sufficiently large detuning α becomes imaginary and the parametric oscillator degenerates into a frequency converter. Furthermore α achieves its maximum value for a detuning

$$\Delta = \frac{\alpha_o{}^2(\omega_1+\omega_2)}{\omega_1\omega_2 - \alpha_o{}^2} \qquad (19)$$

However since $A_K$ is assumed to vary slowly compared with $\omega_K$,

$$\alpha_o{}^2 << \omega_1\omega_2$$

and the amount of desirable detuning given by Eq. (19) is negligible. Thus, as a general rule, detuning is detrimental to the buildup rate.

Losses - No Detuning. In this important case Δ=0 and Eq. (16) yields

$$\alpha = - \frac{\alpha_1+\alpha_2}{2} \pm \sqrt{\alpha_o^2[1+ \frac{i\alpha_1}{\omega_1}][1- \frac{i\alpha_2}{\omega_2}] + \frac{(\alpha_1-\alpha_2)^2}{4}} \qquad (20)$$

while in most practical cases

$$\alpha_o{}^2 >> (\alpha_1+\alpha_2)^2$$

so that Eq. (20) is

$$\alpha \approx - \frac{\alpha_1+\alpha_2}{2} \pm \alpha_o[1 + \frac{i}{2}(\frac{\alpha_1}{\omega_1} - \frac{\alpha_2}{\omega_2})] \qquad (21)$$

This shows the buildup rate to be decreased by the mean damping of the two tanks. In addition, a frequency bifurcation occurs; its significance will be discussed in connection with its effect on phase uncertainty.

## Initial Conditions

To discuss phase characteristics during buildup it is necessary to define the constants in the solution equations

$$a_1 = e^{i(\omega_1 - \frac{\Delta}{2})t}[A_{II}e^{\alpha t} + A_{12}e^{-\alpha t}] \qquad (22a)$$

$$a_2 = e^{i(\omega_2 - \frac{\Delta}{2})t}[A_{21}e^{\alpha t} + A_{22}e^{-\alpha t}] \qquad (22b)$$

However, by the coupling conditions $A_{21}$ and $A_{22}$ are linearly dependent upon $A_{11}$ and $A_{12}$. This is proper since $A_{11}$ and $A_{12}$ are complex, thus requiring four constants: the two initial voltages and the two initial currents.

Ideal Case. In this case Eq. (13b) becomes simply

$$\dot{A}_2 = \omega_2 \eta A_3 A_1^*$$

and by equating coefficients of $e^{\alpha t}$ and $e^{-\alpha t}$

$$A_{21} = \sqrt{\frac{\omega_2}{\omega_1}}\, A_{11}^* e^{i\emptyset p}$$

$$A_{22} = -\sqrt{\frac{\omega_2}{\omega_1}}\, A_{21}^*\ e^{i\emptyset p} \qquad (23)$$

Substitution of these quantities into Eqs. (7) gives, after some arithmetic

$$2A_{11} = [\sqrt{L_1}I_1 + \sqrt{L_2}I_2\sqrt{\frac{\omega_1}{\omega_2}}\sec\emptyset_p] +$$

$$+\ i[\sqrt{C_1}V_1 - \sqrt{C_2}V_2\sqrt{\frac{\omega_1}{\omega_2}}\csc\emptyset_p] \qquad (24a)$$

$$2A_{12} = [\sqrt{L_1}I_1 - \sqrt{L_2}I_2\sqrt{\frac{\omega_1}{\omega_2}}\sec\emptyset_p] +$$

$$+\ i[\sqrt{C_1}V_1 + \sqrt{C_2}V_2\sqrt{\frac{\omega_1}{\omega_2}}\csc\emptyset_p] \qquad (24b)$$

At this point it is convenient to assume that sinusoidal currents are applied to the signal and idler tanks for $t \leq 0$. To avoid confusion with $I_K$ the coefficients will be denoted by $J_K$; thus at $t=0$

$$I_1 = J_1\cos\Theta_1 \qquad V_1 = \frac{J_1}{\omega_1 C_1}\sin\Theta_1 \qquad (25a)$$

$$I_2 = J_2\cos\Theta_2 \qquad V_2 = \frac{J_2}{\omega_2 C_2}\sin\Theta_2 \qquad (25b)$$

where $\Theta_K$ is the appropriate phase at $t=0$. Clearly then:

$$A_{11} = \frac{\sqrt{L_1}J_1}{2}\, e^{i\Theta_1} + \frac{\sqrt{L_2}J_2}{2}\sqrt{\frac{\omega_1}{\omega_2}}\, e^{-i\gamma} \qquad (26a)$$

$$A_{12} = \frac{\sqrt{L_1}J_1}{2}\, e^{i\Theta_1} - \frac{\sqrt{L_2}J_2}{2}\sqrt{\frac{\omega_1}{\omega_2}}\, e^{-i\gamma} \qquad (26b)$$

$$\tan\gamma = \frac{\tan\Theta_2}{\tan\Theta_p}$$

Loading-No Detuning-Quiescent Idler. In the case of finite loading but no detuning the algebra becomes more complex; however in the case of a quiescent idler ($J_2=0$) the results are still simple. In analogy with Eq. (26a) it is found that (approximately)

$$a_1 = \frac{J_1\sqrt{L_1}}{2}\ \overline{1 + \frac{\sin 2\Theta_1}{2Q_1}}\ \cdot$$

$$\cdot\ e^{i\gamma + i\omega_o t - \frac{\alpha_1 + \alpha_2}{2}t},$$

$$\cdot\ \{e^{\alpha_o(1+i\beta)t} + e^{-\alpha_o(1+i\beta)t}\} \qquad (27)$$

where

$$\beta = \tfrac{1}{2}(\frac{\alpha_1}{\omega_1} - \frac{\alpha_2}{\omega_2} = \frac{1}{4}(\frac{1}{Q_1} - \frac{1}{Q_2}) \quad \text{and}$$

$$\tan\gamma = \sqrt{1 - \frac{1}{4Q_1^2}}\ \frac{\sin\Theta_1}{\cos\Theta_1 + \frac{1}{2Q_1}\sin\Theta_1}$$

## Phase Error During Buildup

From Equs. (23) and (16) it is seen that as the time from start increases

$$a_1 \rightarrow [\frac{\sqrt{L_1}J_1}{2} + \frac{\sqrt{L_2}J_2}{2}\sqrt{\frac{\omega_1}{\omega_2}}\, e^{-i\gamma - i\Theta_1}]\ \cdot$$

$$\cdot\ e^{\alpha_o t + i\omega_1 t + i\Theta_1} \qquad (28)$$

Since the second term in the bracket may be complex, $a_1$ is not necessarily in phase with the initial excitation. Clearly the worst case occurs when the pump signal and idler phase are such as to make the second term pure imaginary; then the additional phase shift

$$\emptyset_\varepsilon = \tan^{-1}\sqrt{\frac{L_2\omega_1}{L_1\omega_2}}\,\frac{J_2}{J_1} \qquad (29)$$

This quantity is the amount of unpredictable phase shift, unpredictable since the phase relationship between the pump, signal and idler cannot be assumed known in general.

In principle it can be made to vanish by adequate damping of the idler tank prior to buildup, since $\emptyset_\varepsilon = 0$ if $J2=0$, although in practice, this is not usually possible.

In addition to the phase uncertainty due to initial conditions in the idler tank, Eq. (27) shows that even if $J_2=0$ a phase uncertainty can exist; in this case it is on the order of $\alpha_1/\omega_1$. By way of comparison a conventional oscillator introduces an unpredictable phase error of $\alpha_o/\omega_1$. The improved performance of the parametric oscillator is due to its extra degree of freedom, represented by the idler tank.

Eq. (27) also shows that $a_1$ is initially essentially a modulated carrier, due to the exponential factors $\pm i\beta t$, but that as t increases one sideband decays exponentially while the other builds up. The result is a net change in frequency of $\beta$, due to the finite Q's of the two tanks; it is not, however, a phase error since it is perfectly predictable. This change in frequency is the significance of the complex $\alpha$ referred to previously.

## Large Signal Analysis

In general as the signal and idler voltages build up the pump voltage begins to be influenced as shown by Equation (13c). The following section outlines the large signal theory for the ideal case, but the results will be qualitatively correct for most practical circuits.

Steady State Conditions. In the steady state $\dot{A}_K=0$ so that in the ideal case Eqs. (13) become

$$\omega_1\eta A_3A_2^* = \omega_2\eta A_3A_1^* = 0 \qquad (30)$$

$$\alpha_3A_3 = -\omega_3\eta_3A_1A_2 + \frac{V_{po}e^{i\emptyset_p}}{2\sqrt{L_3}} \qquad (31)$$

Clearly in the nontrivial case $A_1\neq0$. Eq. (30) requires $A_3=0$ so that Eq. (31) yields

$$|A_1A_2| = |A_1||A_2| = \frac{V_{po}}{2\sqrt{L_3}\omega_3\eta|1+i\frac{\alpha_3}{a_3}|} \approx$$

$$\approx \frac{C_3}{C_o} V_oV_{po}\sqrt{\frac{C_1C_2}{1+\frac{\alpha_3^2}{\omega_3^2}}} \qquad (32)$$

(The term $\alpha_3$ is retained since the pump must have a finite Q for Eq. (14) to be finite.) This equation gives the limit curve which relates $|A_1|$ and $|A_2|$ at saturation.

A second relationship is found from Eqs. (13a) and (13b). By forming their ratio:

$$\frac{\dot{A}_1}{A_2} = \frac{\omega_1}{\omega_2}\frac{A_2^*}{A_1^*}, \qquad (33)$$

from which it follows that

$$|A_1|^2 = \frac{\omega_1}{\omega_2}|A_2|^2 + \Omega \qquad (34)$$

where $\Omega$ is a constant. This equation holds for all time, and in particular, it is also valid at saturation. Thus Eqs. (32) and (34) allow computation of the saturated values of $|A_1|$ and $|A_2|$, as shown graphically in Figure 2.

Manley-Rowe Equation. By its definition $|A_K|^2$ is the total energy in the $K^{th}$ circuit so that the power

$$P_K = \frac{d}{dt}|A_K|^2$$

but from Eq. (34) this gives

$$\frac{P_1}{\omega_1} = \frac{P_2}{\omega_2} \qquad (35)$$

which is the familiar Manley-Rowe equation. It shows that if the available power is to go primarily to the signal tank, $\omega_1 \gg \omega_2$, but this is the converse of the condition necessary for fast buildup and low phase uncertainty due to residual idler current. This is roughly equivalent to the usual gain-bandwidth product conservation, i.e., fast buildup (wide bandwidth) results in low gain; the phase uncertainty forms an additional side constraint.

Approximate Nonlinear Buildup. It is possible to construct an approximate analysis of the buildup with the aid of a few reasonable approximations: first, if $|A_3|$ varies slowly, $|\dot{A}_3|\approx0$ and

$$\alpha_3\rho_3 \approx \frac{V_{po}}{2\sqrt{L_3}} - \omega_3\eta\rho_1\rho_2 \qquad (36)$$

where $\rho_K = |A_K|$. Second, from Eq. (34) it is clear that by the time $\rho_1$ becomes several times its initial value the constant $\Omega$ becomes negligible, so that

$$\rho_1 \approx \frac{\omega_1}{\omega_2}\rho_2 \qquad (37)$$

is valid for almost all time.

By use of the approximations contained in Eqs. (36) and (37) and the assumption that the idler is initially quiescent, Eq. (13a) becomes

$$\dot{\rho}_1 \approx \omega_1\eta\rho_2[\frac{V_{po}}{2\Omega_3} - \omega_3\eta\rho_1\rho_2]\frac{1}{\alpha_3}$$

$$= +\alpha_o[\rho_1-\lambda\rho_1^3] \qquad (38)$$

with

$$\lambda = 2\omega_2 Q_3 \eta^2/\alpha_o = \frac{C_o}{C_3 V_o V_{po}}\sqrt{\frac{\omega_2}{\omega_1 C_1 C_2}} \quad :$$

The solution of this equation is

$$\frac{\rho}{\rho_{10}} = \frac{e^{\alpha_o t}}{\sqrt{1+\lambda\rho_{10}{}^2[e^{2\alpha_o t}-1]}} \tag{39}$$

which shows the amplitude is limited as time becomes infinite to the saturated level given in Eq. (32).

It is of some interest that Eq. (39) is identical in form with that found by Van Der Pol as an approximate solution to his famous equation for the behavior of a triode oscillator.

## Acknowledgement

The author thanks his company and the Rome Air Development Center for permission to publish this report.

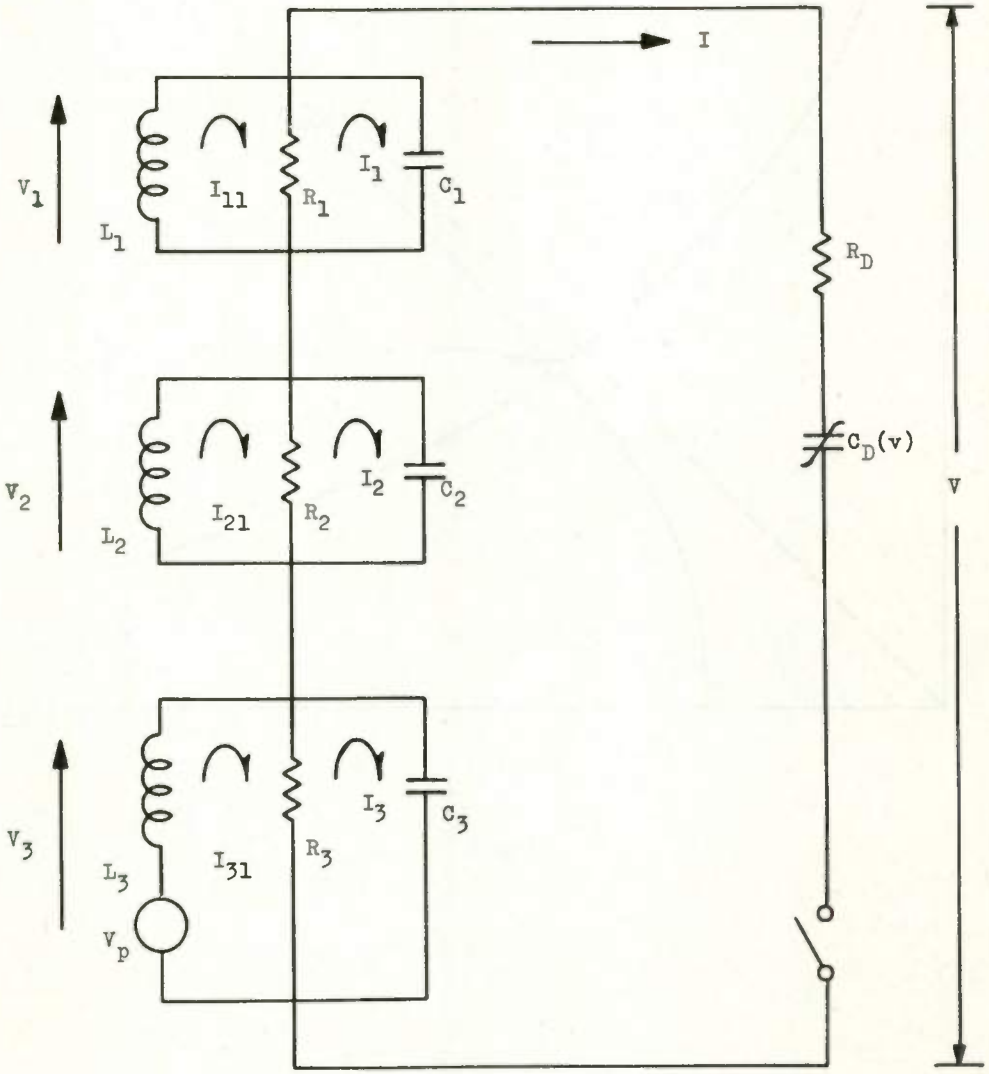


Fig. 1.

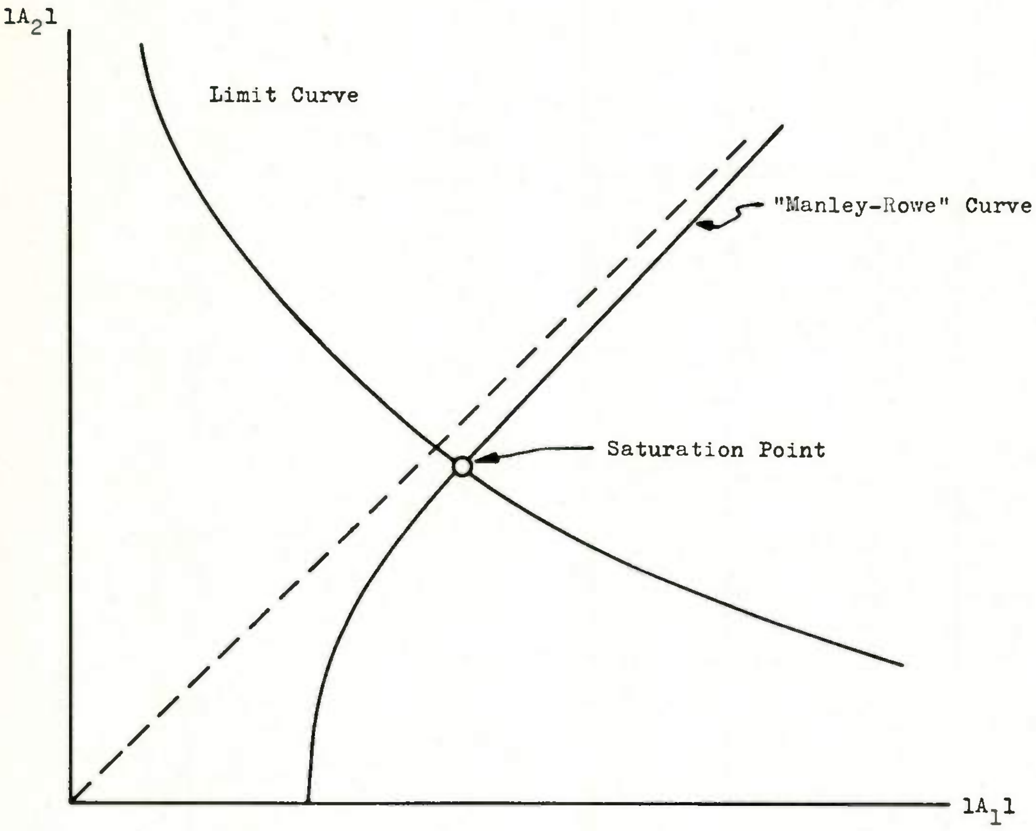


Fig. 2.

# MAXIMUM SAMPLING RATE FOR SUPERREGENERATIVE AMPLIFIERS*

Don N. Thomson
General Atronics Corporation
West Conshohocken, Penna.

## Summary

This paper presents the results of a study to determine the maximum sampling rates permissible for a superregenerative amplifier. Quantitative results, theoretical and empirical, are reported. For a VHF superregenerative amplifier a sampling rate of several megacycles can be achieved.

## Introduction

The superregenerative amplifier (SR) is very useful for some applications because of the extremely high gain that can be achieved in a single stage. Voltage gain-bandwidth products of $10^{10}$ cps are readily achieved. Because the SR samples the input signal the maximum bandwidth is determined by the permissible sampling rate.

One particularly useful application of the SR is as the amplifier in a recirculating loop data processor. It is convenient to operate with phase information, the SR being operated in a saturated mode. For this mode of use, two important parameters of the SR are phase distortion and dynamic range.

In subsequent sections the theory of the SR is reviewed and general relations are developed between sampling rate, dynamic range, and phase distortion. Then an equation is developed relating dynamic range to the sampling rate. Finally some experimental results are given for vacuum tube, transistor, and parametric SR's.

## Description of the Superregenerative Amplifier

The fundamental signal behavior of the SR was described in some detail by Bradley[1] in the late 1940's, but Bradley's work did not include the phase relations in the amplifier, nor the noise problem. A paper by George and Urkowitz[2] derived the noise relations for certain cases. Finally, under a contract with the Signal Corps, General Atronics extended the SR theory to include the general noise problem as well as the problem of phase distortion.[3,4]

a) Qualitative Theory. Basically, the SR design makes use of the extremely high regenerative gains which exist when an oscillator is building up. The SR consists of a parallel tank shunted by a conductance that can be varied from positive to negative as shown in idealized form in Figure 1(a). When the switch is in the $G_o$ position the tank is heavily loaded and very little voltage exists across it. If the switch is then changed to the $G=G_1$ position and if $G_1$ is small, the signal is being driven into a tank of very high Q. When $G_1=0$ the tank acts as a perfect integrator and the envelope of the voltage across the tank increases linearly with time. The conductance versus time is shown in Figure 1(b) and the voltage versus time is shown in Figure 1(c). The time interval during which the signal is integrated in the tank is called the "listening" time, $t_1$. At the end of the listening time the conductance is made negative. The envelope of the tank voltage then increases exponentially from the value it had at the end of the listening period. If the switch is returned to the positive conductance after a buildup time $t_2$, the energy in the tank is rapidly damped. If $t_2$ is sufficiently short, the active element that produces the negative conductance does not saturate. Such a mode of operation is termed the "linear" mode because the value of the envelope at the end of the buildup period is directly proportional to the drive current. If, on the other hand, the active element is allowed to saturate, the $g(t)$ and $e(t)$ are shown as in Figures 1(d) and 1(e). In this case $e(t)$ reaches a peak value which is limited by saturation effects; at the same time, $g(t)$ must become zero because G=0 is the only condition which will permit a steady envelope, $e(t)$. The solid curve of Figure 1(d) shows the $g(t)$ curve if saturation had not occurred; the dashed line shows the effect of saturation. The saturated mode is known as the logarithmic mode because the area of the envelope of $e(t)$ is proportional to the logarithm of the drive current.

Certain qualitative features of the SR performance can be deduced from the above description:

1) There should be as much decay in the $t_o$ period as there is gain in the $t_1$ and $t_2$ intervals if the buildup in

*This work was sponsored by Rome Air Development Center under Contract AF 30(602)-2283.

each cycle is to be independent of past history.

2) If a noise source is present it is desirous that $G_1$ be small and that $t_1$ be as long as possible, compatible with the signal duration. This is because the tank becomes a perfect integrator when $G_1=0$ and the signal voltage builds up coherently and linearly during $t_1$, whereas the noise voltage only increases as the square root of $t_1$.

3) If the signal is pulsed and $G_1 \simeq 0$, the best performance is obtained when the pulse is coincident with the listening interval.

4) If noise exists during both the listening and buildup periods, but is larger during buildup, it is desired that buildup occur as rapidly as possible.

5) The bandwidth can also be deduced in a qualitative manner. If $G_1$ is approximately zero and $|G_2|$ is much greater than $G_1$, then the final value of the envelope in the linear mode is determined almost entirely by the signal which enters during the listening interval, $t_1$. Then the SR can be represented by a linear amplifier preceded by a tank and preceded by a switch which connects the source to the tank for a time equal to $t_1$. And the bandwidth is determined by the bandwidth of a gated high-Q circuit.

An important factor which is not obvious from the above description is the fact that the SR is "phase transparent". By this is meant that the high level SR signal out of the SR has a phase which is determined by the phase of the low level signal fed into the SR.

b) Basic Equations. The following equations give the value of the signal envelope and of the noise envelope at a time $t=t_1+t_2$, measured from the start of the listening period. The assumed g(t) function is that of Figure 1(b). The equations ignore the quench period on the assumption that perfect quenching exists. Furthermore the equations assume that $\alpha_1 \ll \omega_o$, $|\alpha_2| \ll \omega_o$, and that the signal is at center frequency, $\omega_o$.

The peak envelope as determined by a pulsed signal which is properly aligned with the listening period is

$$E \simeq \frac{I_s}{G_1}(1-e^{-\alpha_1 t_1})e^{|\alpha_2| t_2} \qquad (1)$$

where $I_s$ = peak signal current
$G_1, G_2$ = conductance during listening and buildup periods
$\alpha_1, \alpha_2$ = $G_1/2C$, $G_2/2C$
$t_1, t_2$ = duration of listening and buildup periods

The mean square noise envelope is

$$\overline{E_n^2} = \frac{2\pi\alpha_1}{G_1^2}\overline{I_{N1}^2}(1-e^{-2\alpha_1 t_1})e^{2|\alpha_2| t_2} \qquad (2)$$

where $\overline{I_{N1}^2}$ = mean square noise current per radian bandwidth during the listening period.

It is assumed for the above equation that the circuit parameters are sufficiently ideal that the noise contributed during the buildup period is negligible. One other equation of interest is that which gives the value of the RF envelope at the end of the buildup period due to a current transient at the beginning of the listening period. For the simple case of a ramp of current having a rate of rise of $I/\delta$ the relation is

$$E_T \simeq \frac{2I}{\omega_o^2 C\delta} e^{-\alpha_1 t_1} e^{|\alpha_2| t_2} \qquad (3)$$

(If the transient occurs at the end of the listening period the

$$e^{-\alpha_1 t_1}$$

factor becomes unity.)

The equations for the signal and noise envelopes are derived in detail in Reference 3. The equation for the transient is a simple extension of the detailed theory given in Reference 3.

One additional comment must be made regarding the conductance-time waveform, g(t). The ideal waveform of Figure 1 does provide the best attainable S/N ratio for the SR, in theory, but it is highly impractical. The switches which produce the rapid changes in conductance would create intolerably high transient energy. In Reference 3, however, it is shown that a much softer g(t) waveform can be used with very little loss in S/N performance. In Figure 2 is shown a specific form of g(t) which causes negligible loss. The conductance is varied linearly during the listening period from the positive value $G_o$ to the negative value $G_2$.

## Limitations on SR Performance

For the saturated phase-sensitive SR two performance factors must be considered: dynamic range and phase distortion.

The dynamic range of the SR is defined as the ratio of the power output to the equivalent noise input. The input noise consists of thermal noise plus transient energy introduced by the switching action.

Phase distortion is an inherent property of the SR. By phase distortion is meant differences between output phase and input phase which are not invariant.

Phase distortion in the SR arises from two sources (in addition to noise). The first source is the result of a short listening time. At the end of $t_1$ there is both a steady state and a transient voltage term and the net phase is not identically the phase of the signal. The second source is caused by the rapid changes in the conductance, especially during the turn-on period.

The phase error of the first source is dependent upon the shape of the applied signal pulse. For a half-sinusoid

$$\Delta\phi_1 = \frac{\alpha_1^3}{\omega_o^3} \quad ; \qquad (4a)$$

for a rectangular pulse,

$$\Delta\phi_1 = \frac{1}{\omega_o t_1} \quad . \qquad (4b)$$

With typical numbers the sinusoidal pulse would cause little error, but the rectangular pulse could result in appreciable errors at high sampling rates.

The error due to the turn-on is a function of the way in which $g(t)$ changes. If $\alpha(t)$ rises from zero to $\alpha_2$ in a time $\Delta t$,

$$\Delta\phi_2 = \frac{|\alpha_2|}{2\omega_o^2 \Delta t} \quad ; \qquad (5a)$$

for a quadratic rise,

$$\Delta\phi_2 = \frac{|\alpha_2|}{2\omega_o^3 (\Delta t)^2} \qquad (5b)$$

The quadratic rise will cause negligible error but the linear rise can cause a significant error for very high sampling rates.

Using the equations presented above, the general relations can be developed between phase distortion and dynamic range versus sampling rate and center frequency. It is assumed that $\alpha_1 t_1 << 1$.

a) Sampling Rate. As the sampling rate, $f_s$, is increased $t_1$ must decrease. The voltage envelope due to thermal noise level is proportional to $\sqrt{t_1}$, whereas the envelope due to signal is proportional to $t_1$. The transient is little affected if the switching times are fixed in duration. Then both the signal-to-noise ratio and the signal-to-transient ratio decrease as $f_s$ increases. The phase distortion either increases or remains constant, depending on the shape of the input pulse; usually some increase will be expected though not as much as indicated by Equation (4a). The saturated output power is constant.

The net result is that as $f_s$ is increased the dynamic range must decrease and usually the phase distortion will increase.

b) Center Frequency. Phase distortion and transient energy both decrease as $\omega_o$ is increased. Thus it is desired that $\omega_o$ be as high as possible. However, many active devices suffer a loss in gain and an increase in noise figure at high frequencies, so that a compromise must be sought. The investigation reported herein was for VHF.

### Dynamic Range versus Sampling Rate

The previous section presented the pertinent factors and some qualitative relations. In this section will be developed a quantitative relation between sampling rate and dynamic range.

Consider the ideal conductance function of Figure 3a (and the corresponding RF envelope of Figure 3b). The buildup time $t_2$ is divided into $t_3$, the time for buildup, and $t_4$, the duration of the saturated pulse. Since $t_4$ and $t_1$ must be about equal (for a recirculating loop processor),

$$t_s = \frac{1}{f_s} = 2t_1 + t_o + t_3 \qquad (6)$$

To simplify the analysis consider some specific numbers. Let the required gain be 70 db and let the decay during $t_o$ be 120 db. (These are desirable values for a sweep integrator.) The gain occurs mostly during $t_3$, so

$$t_3 \approx \frac{70}{8.6|\alpha_2|} \qquad \text{(from Equation (1).}$$

The limitation on $t_3$ and $t_o$ must now be considered.

a) Buildup Speed. $t_3$ is minimized by maximizing $\alpha_2$. It can be shown[5] that for vacuum tube circuits, under the best conditions

$$|\alpha_2| \simeq \frac{gm}{4\sqrt{c_1 c_o}} \qquad (7)$$

where $c_1$ and $c_o$ are the input and output capacitances. It is interesting to note that $|\alpha_2|$ is proportional to the voltage gain-bandwidth product of the active device used. For the 6688 pentode the calculated value was approximately $5x10^8$. The measured value was $7x10^7$. Using the measured value, $t_3$=120 nanoseconds; if the calculated value could be achieved, $t_3$=18 nanoseconds. For the 2N769 transistor no calculation was made but the measured value of $\alpha_2$ was $6x10^7$.

b) Quenching. The conductance curve of the SR is generated by turning the active device on and off and by applying a quenching switch across the tank during the listening period. The best switch discovered was the 2N240 transistor. With optimum series resistance the desired 120 db of quenching could be achieved, in theory, in 30 nanoseconds. In practice it was found that close to 100 nanoseconds was required to provide the quenching. Some of this loss was due to limited rise times of the applied switching waveforms. Using Equation (6) and the calculated conditions of $t_3$ and $t_o$,

$$t_1 = \frac{t_s - 5x10^{-8}}{2} \qquad (8a)$$

If the measured values of $t_3$ and $t_o$ are used,

$$t_1 = \frac{t_s - 2x10^{-7}}{2} \qquad (8b)$$

Finally the dynamic range can be expressed as a function of $t_1$. At high sampling rates the sensitivity is limited by the transient energy which is independent of $t_1$. The signal energy integrated during the listening period is proportional to $t_1^2$. The dynamic range is defined as the ratio of the peak envelope (with input sufficient to provide saturation) to the signal power that produces a unity signal-to-interference ratio. Then DR = $P_o/P_s$, where $P_s$ is the signal power required for unity signal-to-transient. Since the transient power is essentially independent of $t_1$ but $P_s \alpha 1/t_1^2$,

$$DR = \frac{P_{out} t_1^2}{k} \qquad (9)$$

The saturated output power, $P_{out}$, is constant. Then by use of Eq. (8),

$$DR(db) = 20 \log k_1(t_s - k_2) \qquad (10)$$

where $k_2 = t_o + t_3$. $k_2$ has a value of $5x10^{-8}$ sec theoretical or $2x10^{-7}$ sec measured. The value of $k_1$ is obtained empirically. It is noted that Equation (10) is valid only above about 1 mc where transient interference predominates; below 1 mc DR varies with the first power of $t_1$. With a good vacuum tube circuit DR=117db[1] at 1 mc. Using this measured value in conjunction with Equation (10) the curves of Figure 4 can be derived.

The important fact regarding Equation (10) and Figure 4 is that there is a definite upper limit to the sampling rate: when the listening time becomes zero the dynamic range must be zero. The sampling rate which produces this condition is found to have a maximum theoretical value of 20 megacycles, and a probable practical value of about 5 megacycles.

## Experimental Results

Before describing the experimental results it is desirable to comment on two factors. The first is that the SR was being investigated for use with low impedance delay lines. For such use it is convenient to provide an input buffer amplifier and an output buffer amplifier. The input buffer amplifier provides gain and it does not increase quenching problems because it feeds directly into the SR tank. The output buffer is a wideband unity gain pentode "cable driver" and usually does not require quenching. Both buffers are properly considered as part of the SR. For the experiments the input buffer was employed for convenience. The output buffer was not included, however, because it is not usually a limiting factor and its use would not have simplified the measurements. With the use of an output buffer the source and load resistances are approximately equal and voltage-squared ratios can be used in place of power ratios.

The other factor to be noted is that the experiments could not be conducted in exactly the way implied by Equation (8). It is very important to insure that the quenching is adequate and this can on only be done by having the SR reach full saturated output in the absence of input signal. This has the effect of forcing the use of a larger value of $t_2$. The effect becomes appreciable as the limiting sampling rate is approached.

a) Vacuum Tube SR. The best vacuum tube circuit was found to be a balanced configuration as shown in Figure 5. The SR tubes are 6688. A 7721 is used as a buffer and the buffer is considered to be part of the SR. The balanced configuration provides a first order cancellation of the transient caused by turn-on of the tubes and by turn-on of the switches. The switches are 2N240 transistors. It was found necessary to quench both the plate and grid circuits.

Considerable care was needed to optimize the switching waveforms applied to the SR grids and to the bases of the switches. The limiting transient energy appeared to be that induced by the switches. The approximate modulation waveforms are shown in Figure 6.

The buildup speed, $\alpha_2$, was found to be about $7x10^7$ as opposed to the calculated value of $5x10^8$. Adequate quenching proved difficult and it was found necessary to provide about 0.1 μs for this action. The operating frequency was 90 mc.

The resulting dynamic range as a function of sampling frequency is shown in Figure 7. It is seen that the results follow the expected trend, but the dynamic range falls more rapidly than predicted. This can be explained in part by the measurement technique, as discussed at the beginning of this section.

b) Transistor SR. The transistor circuit also employed a balanced configuration as shown in Figure 8. A 3N25 tetrode was used as a buffer and 2N769 transistors were used for the SR. The switches were 2N240's. The carrier was 75 mc.

The measured $\alpha_2$ was $6x10^7$. The required switching time was about 0.1 μs. The curve of dynamic range versus sampling rate is shown in Figure 7. It is observed that the curve closely parallels that of the vacuum tube SR. The difference in the absolute values is probably due primarily to the lower output voltage attained with the transistor. However, it also appeared that the transient balance was not as perfect.

## Parametric SR

The parametric SR was investigated because it was originally believed that a form of switching might be implemented which produces less transient energy.

The paramp SR was investigated theoretically for buildup speed and phase distortion. Only the lower sideband up converter is useful. It was found that the paramp SR causes less buildup phase distortion than the conventional SR; the listening time phase distortion is of course the same. The buildup rate of the parametric SR is[5]

$$|\alpha_2| = \frac{\Delta c \sqrt{\omega_1 \omega_2}}{4\sqrt{c_1 c_2}} \qquad (11)$$

where $\Delta c$ = peak change in capacity at the pump frequency

$\omega_1, \omega_2$ = signal and idler frequencies

$c_1, c_2$ = equivalent total capacitances of signal and idler tanks.

The calculated buildup rate of a VHF parametric SR was found to be about equal to that of the pentode ($4x10^8$) if a good microwave diode was used with an optimum pumping frequency.

For the measurements a VHF diode was employed. The computed $\alpha_2$ was $4x10^7$ and the measured value was about $2x10^7$.

A measure of dynamic range versus sampling rate was not obtained for the paramp SR for two reasons. First, the design of the SR which produces the best value of $\alpha_2$ causes large values of idler and pump voltage to exist across the signal terminals. Since all these frequencies were at VHF it was not possible to separate them because the required narrowband filters would ruin the buildup performance. (This problem would be alleviated for a microwave paramp.) A second and more fundamental reason for terminating the parametric SR effort was that it had nothing to offer in the way of buildup speed and the problem of providing a balanced switching action is more complex.

It is concluded that for VHF operation the parametric SR is not as useful as a vacuum tube SR. For microwave frequencies the paramp SR may be practical.

## Conclusions

This paper has presented the outlines of a quantitative theory for the maximum sampling rate of SR's. It is shown that as the sampling rate is increased phase distortion must increase and dynamic range must decrease. Experimental results showed buildup times and quenching times longer than predicted. When the measured buildup and switching time were employed in the theory, the theoretical and measured curves of dynamic range versus sampling rate agreed in trend. The lack of detailed agreement is apparently due to the measurement technique.

The results of the investigation are:

1. If the theoretical buildup rate and quenching rates could be obtained the dynamic range would be usable at a sampling rate of almost 20 mc.
2. With practical (i.e., measured) values of buildup rate and quenching speed, the maximum theoretical sampling rate is 5 mc.
3. Measurement with a vacuum tube SR shows usable dynamic range at 4 mc.
4. The transistor SR follows the same trend as the vacuum tube SR but has a dynamic range poorer by 10 to 15 db.
5. At VHF the paramp SR offers no advantage in buildup rate and poses more complex switching problems.

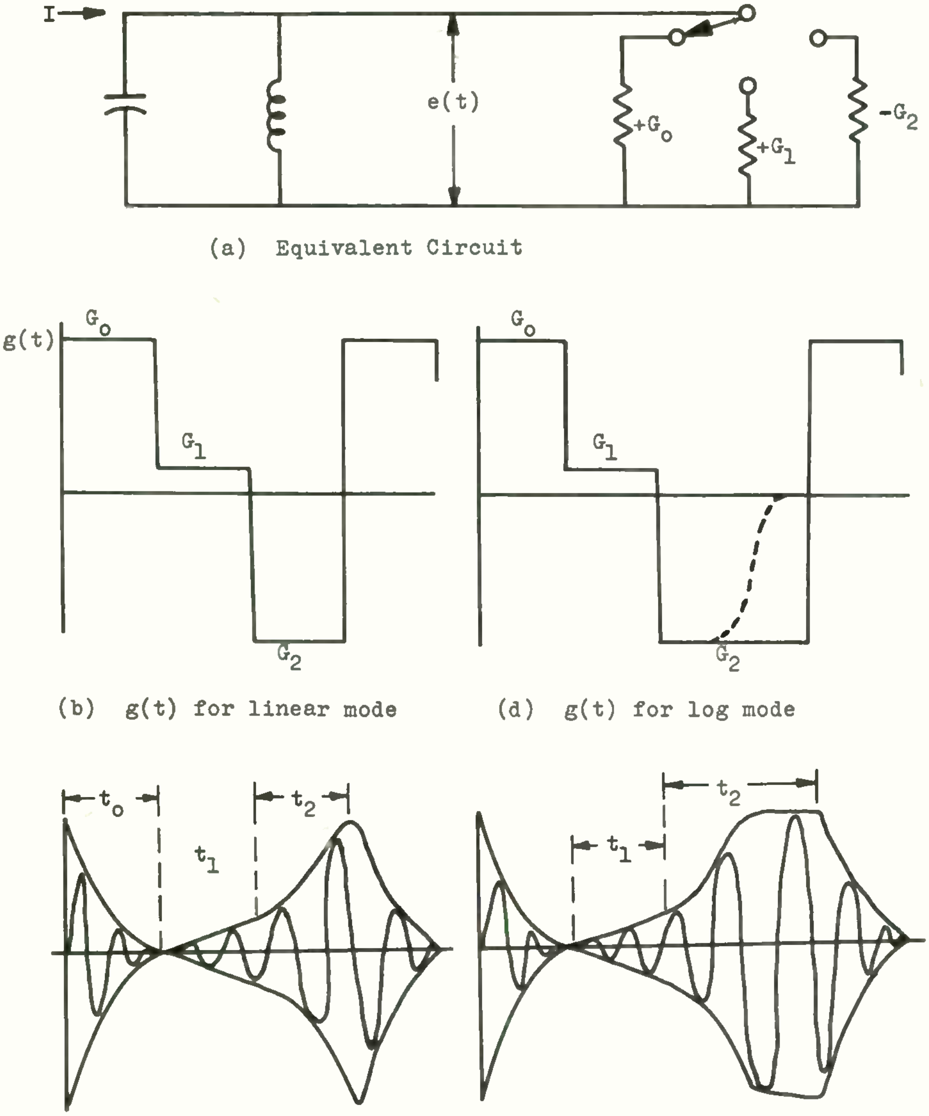
The very high gain-bandwidth product of the SR makes it very useful for certain applications. But the limiting sampling rate restricts the information bandwidth that it can handle.

## Acknowledgement

The valuable assistance of several members of the staff of General Atronics is gratefully acknowledged. Particular note is made of the help of Edwin D. Banta, Bernard D. Steinberg, and Mark S. Zimmerman.

## References

1. W.E. Bradley, "Superregenerative Detection Theory", Electronics Manual for Radio Engineers, McGraw-Hill, New York, p. 606.
2. T.S.George and H. Urkowitz, "Fluctuation in a Microwave Superregenerative Amplifier", Proc IRE, April 1953, p. 516.
3. "A Study of Microwave Superregenerative Amplifiers for Lightweight Radars" Atronics Report 761-196-6, 13 June 1960, ASTIA No. AD 318865.
4. D.N. Thomson, "Microwave Superregenerative Amplifiers", Proc NEC, 1960.
5. D.N. Thomson, "Investigation of Superregenerative Amplifier Techniques for Radar Data Processors", Atronics Report 879-222-12, 29 May 1961.
6. J.R. Whitehead, Super-Regenerative Receivers, Cambridge University Press, 1950



(a) Equivalent Circuit

(b) g(t) for linear mode

(d) g(t) for log mode

(c) e(t) for linear mode

(e) e(t) for log mode

FIGURE 1 - REPRESENTATION OF IDEAL SR

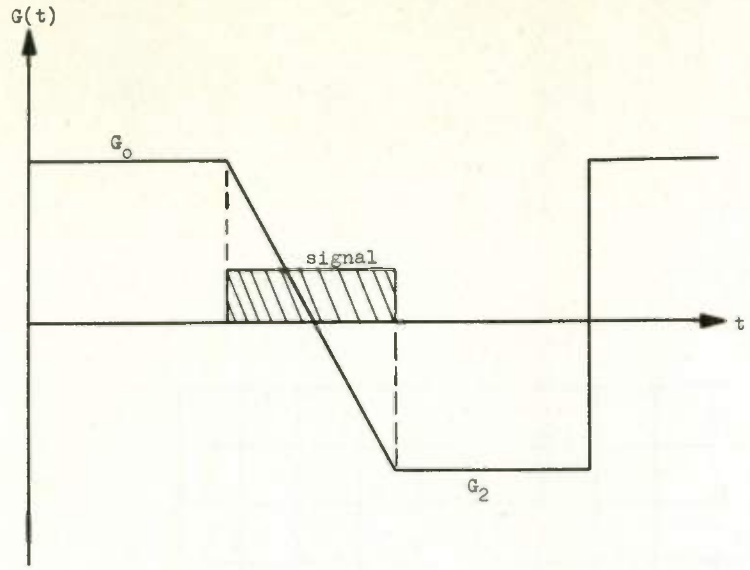
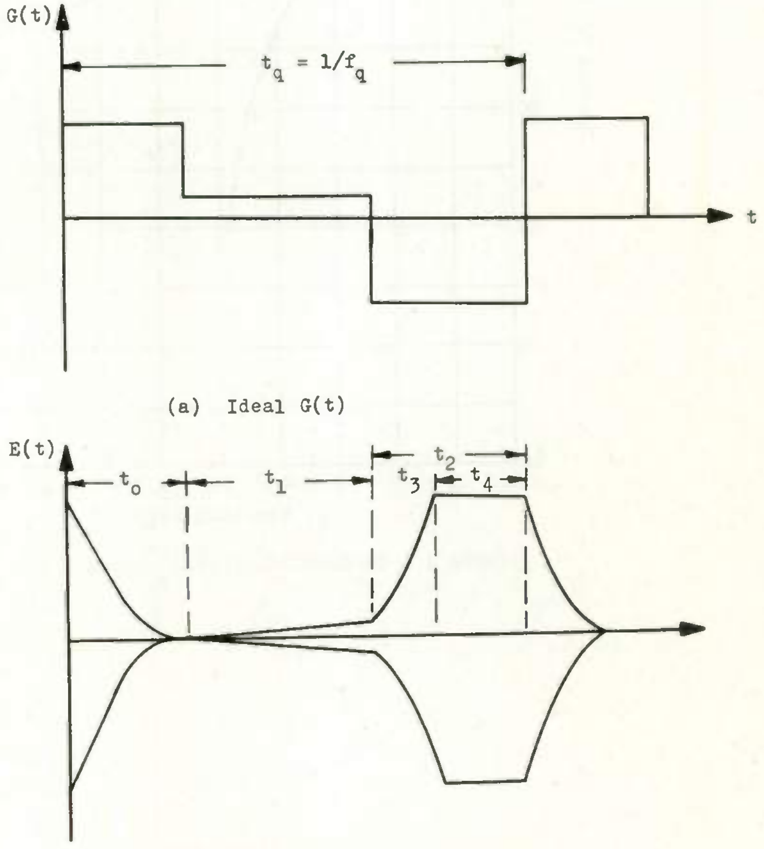


FIGURE 2 - MODIFIED G(t) WAVEFORM



(a) Ideal G(t)

(b) RF Envelope

FIGURE 3 - TIMING WAVEFORM

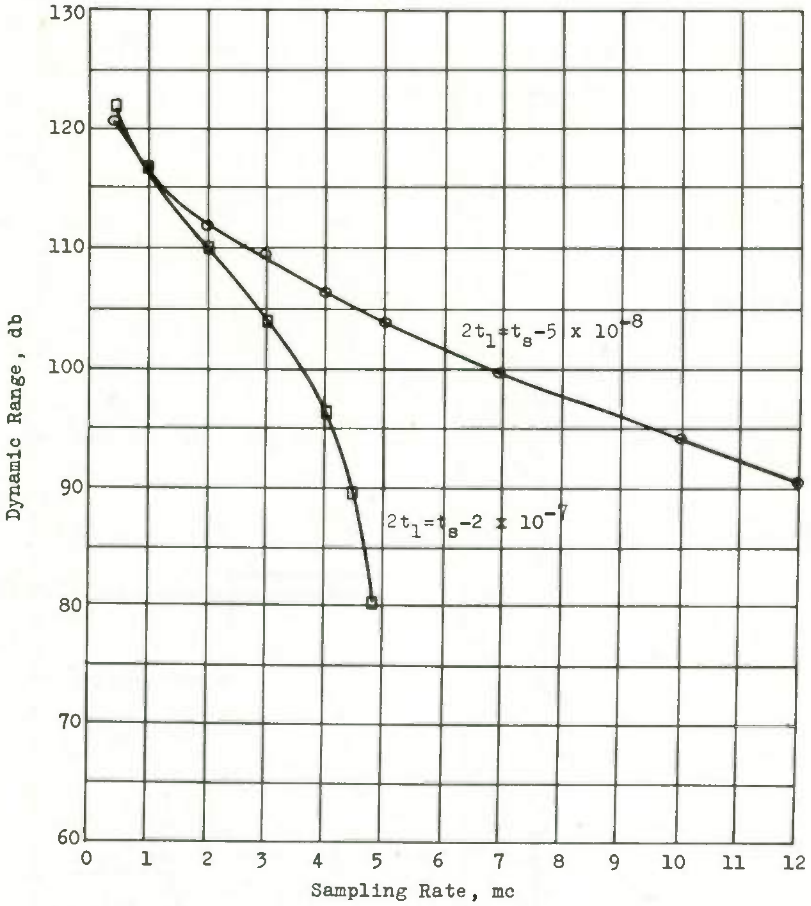
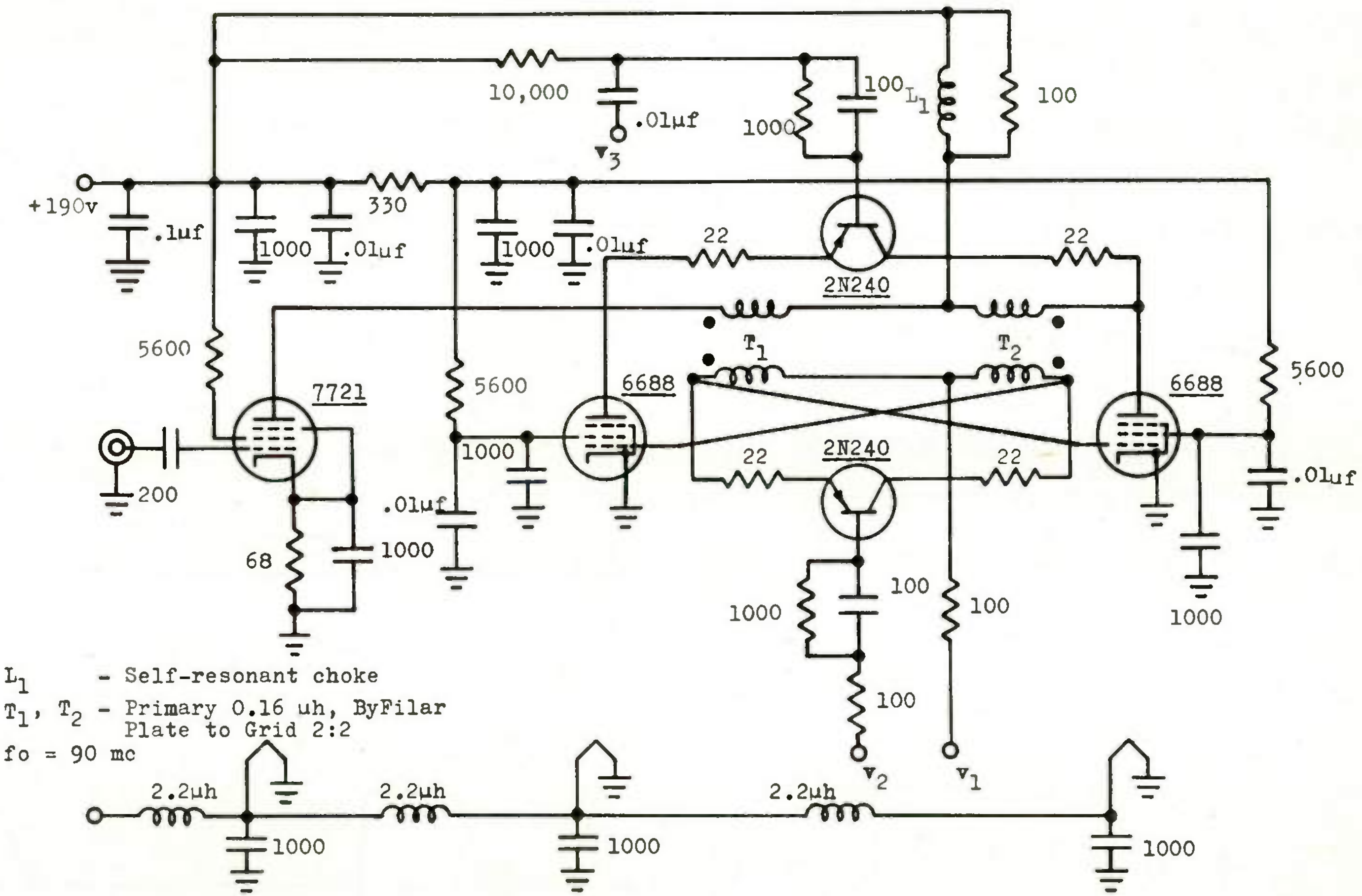


FIGURE 4 - THEORETICAL DYNAMIC RANGE VERSUS SAMPLING RATE

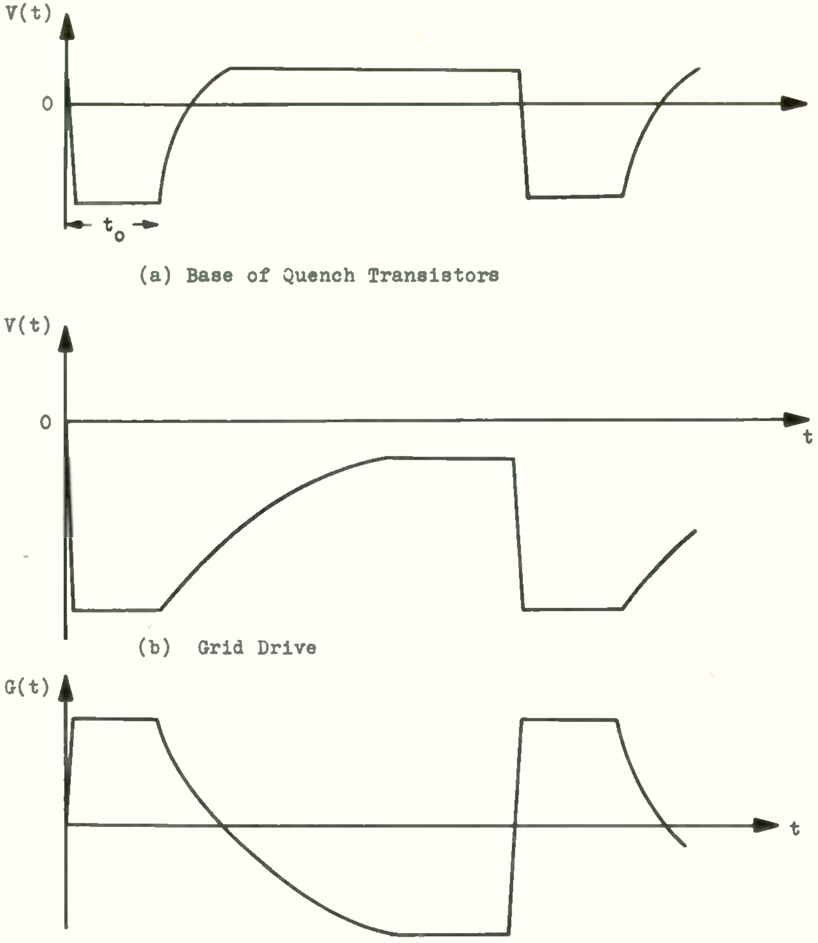


L1 - Self-resonant choke

T1, T2 - Primary 0.16 uh, ByFilar
Plate to Grid 2:2

fo = 90 mc

FIGURE 5 - SCHEMATIC FOR VACUUM TUBE PUSH-PULL SR



(a) Base of Quench Transistors

(b) Grid Drive

(c) Approximate G(t)

FIGURE 6 - MODIFIED TIMING WAVEFORMS FOR SR

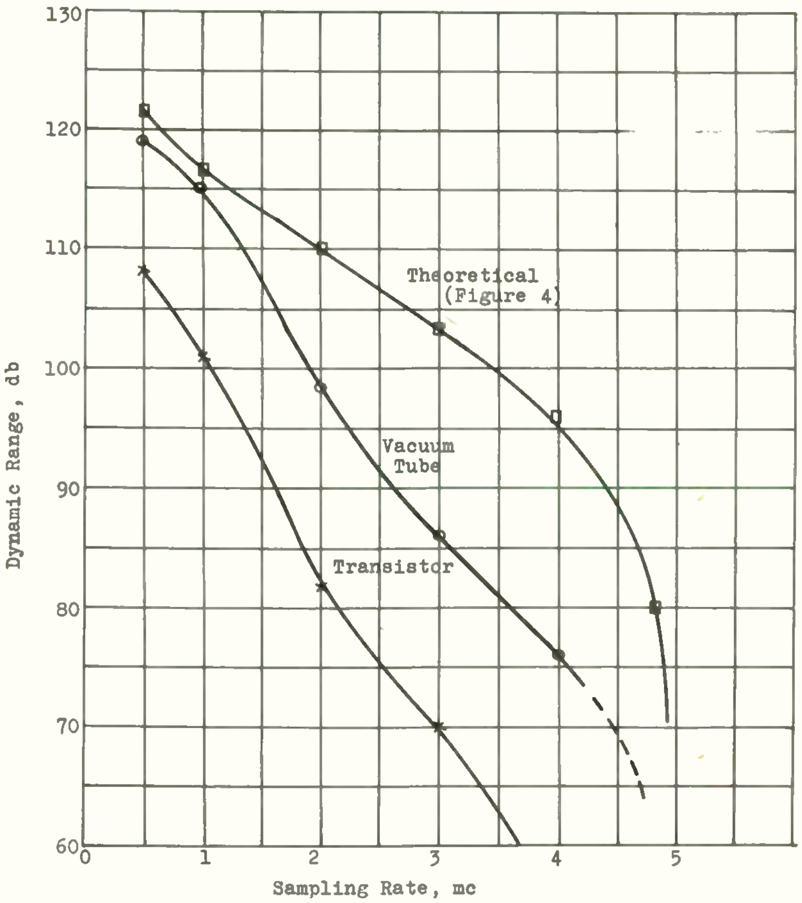
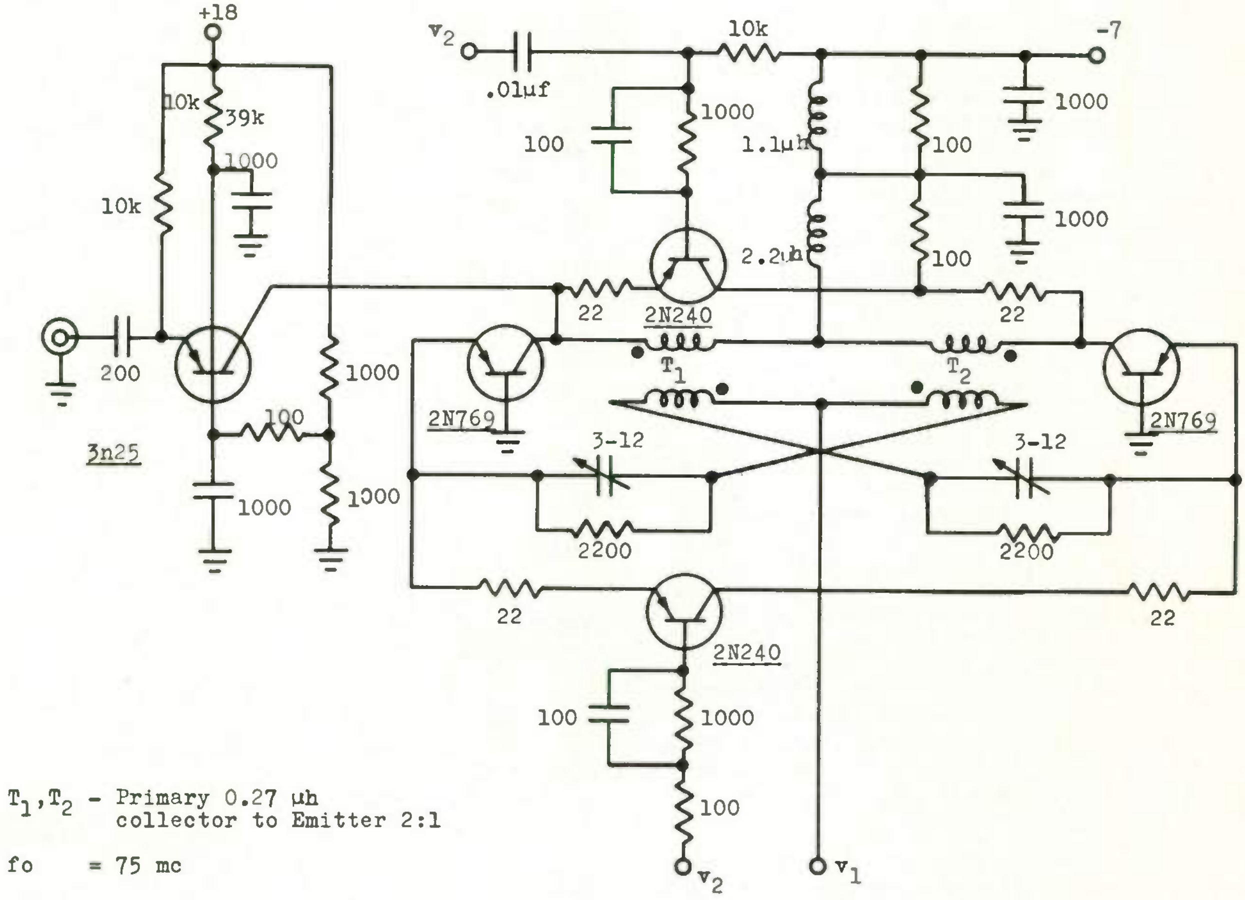


FIGURE 7 - EXPERIMENTAL DYNAMIC RANGE VERSUS SAMPLING RATE



$T_1, T_2$ - Primary 0.27 µh
collector to Emitter 2:1

fo = 75 mc

FIGURE 8 - TRANSISTOR PUSH-PULL SR

# OPTIMUM NONLINEAR FILTERS FOR RANDOM SIGNALS

Naresh K. Sinha
Department of Electrical Engineering
The University of Tennessee
Knoxville, Tennessee

## Introduction

Following the classical work of Wiener on optimum linear filters for random signals, there has been a lot of interest in optimization with the more general case of nonlinear filters. It has been shown that the Wiener filter is truly optimum for Gaussian probability distributions only, and in all other cases nonlinear filters may be found that will give much lower mean-square error between the desired output and the actual output. Important contributions have been made by Singleton,[1] Bose,[2] Rosenbrock[3] and Zadeh.[4] Singleton has shown that for optimizing a nonlinear filter one must determine higher order auto-correlation functions of the input along with higher order cross-correlation functions between the input and the desired output. However, unless the system is very simple, it is impracticable to calculate these correlation functions, even with the aid of computers. Rosenbrock's approach again involves a knowledge of higher order joint probability distributions, which are almost impracticable to calculate, as already pointed out. The classification procedure, as introduced by Zadeh, is indeed quite elegant, but one again encounters the problem of computing higher order probability distribution functions. In his work, Zadeh has restricted his attention mainly to canonically realizable forms which involve a knowledge of only the second order probability distributions for optimization in the mean-square sense. Thus, the practical utility of his work is, evidently, very limited for designing nonlinear filters. Bose's theory, for the experimental determination of optimum time invariant nonlinear systems, is quite simple mathematically, but its utility is limited due to the large number of gate circuits, averaging devices, etc., which are required for a reasonable accuracy.

In other words, it may be stated that the problem of optimization of nonlinear filters is mathematically intractable for continuous random inputs. A great headway can be made, however, by sampling the random signals. An important contribution is the development of "staircase techniques" by Prasad.[5] Prasad has defined "nonlinear correlation functions" which can be computed with the aid of a moderate size digital computer, and can be used for the optimization of nonlinear filters for a given random signal.

In this paper, staircase techniques have been used for the calculation of nonlinear correlation functions for a given non-Gaussian signal superimposed with Gaussian noise. These have been used for optimizing different kinds of nonlinear filters, and the mean-square error between the desired output and the actual output has been calculated in each case. These have been compared with the value of the mean-square error for the optimum linear filter, and it has been shown that a nonlinear filter with memory gives a considerably smaller error.

## Theory of Staircase Systems

In his work, Prasad has presented a comprehensive theory of the analysis and optimization of nonlinear systems subjected to random signals sampled by staircase functions. The theory will be discussed briefly, before going into the application.

The "staircase function" $\ulcorner x(t)$ is defined as below:

$$\ulcorner x(t) = \sum_{n=-\infty}^{\infty} x(nT)\, P(t-nT) \qquad (2.01)$$

where $x(t)$ is a given function of time,
$T$ is the sampling interval, selected in compliance with Shannon's Sampling Theorem,
$P(t-nT)$ is a rectangular pulse of duration $T$ and unit height, applied at $t = nT$,

and $n$ is an integer.

Consider a linear system having the weighting function $w(t)$. Its response to the pulse $P(t)$ is given by

$$u(t) = \int_{-\infty}^{\infty} w(\tau)\, P(t-\tau)\, d\tau \qquad (2.02)$$

The "staircase P-response" of the linear system is now defined by

$$\ulcorner u(t) = \sum_{n=0}^{\infty} u_n\, P(t-nT) \qquad (2.03)$$

where $u_n = u(nT)$

The continuous output of the system to a staircase input $\ulcorner x(t)$ is given by (Fig. 1).

$$y(t) = \sum_{n=-\infty}^{\infty} x(nT)\, u(t-nT) \qquad (2.04)$$

and the staircase output is given by

$$\ulcorner y(t) = \sum_{k=0}^{\infty} \sum_{r=0}^{k} u_r x_{k-r}\, P(t-kT) \qquad (2.05)$$

The output of a zero-memory (or instantaneous) nonlinearity $f[\;]$ to the staircase input $\ulcorner x(t)$ is

$$f[\ulcorner x(t)] = \sum_{n=-\infty}^{\infty} f[x_n]\, P(t-nT) \qquad (2.06)$$

A storage nonlinear system of "first order" is shown schematically in Fig. 2. In this case, the staircase output is given by

$$\mathcal{S}\, y(t) = \sum_{k} \sum_{r=o}^{k} u_r\, f\left[x_{k-r}\right] P(t-kT) \qquad (2.07)$$

It may be pointed out that the simple relationships given in equations (2.06) and (2.07) result from the fact that the operators P(t-nT) form an orthonormal set in the time-domain.

Equations (2.06) and (2.07) may be used for developing optimizing equations for filters subject to random inputs. Consider the linear system of Fig. 1. In this case, one has to design the linear filter in such a manner that the mean-square error between the desired output z(t), which is the noise-free signal, and the actual output y(t) is reduced to a minimum. It can be shown that the mean square error for a random input in which the noise and signals are uncorrelated, is given by

$$\overline{\epsilon^2} = \phi_{\mathcal{S}z\,\mathcal{S}z}(0) - 2\,\phi_{\mathcal{S}z\,\mathcal{S}z}(0) + \phi_{\mathcal{S}y\,\mathcal{S}y}(0) \qquad (2.08)$$

where $\phi_{\mathcal{S}z\,\mathcal{S}z}(mT)$ = staircase auto-correlation function of the desired output z(t)

$$= \frac{1}{2N+1-m} \sum_{n=-N}^{N+1-m} z(nT)\, z(\overline{n+m}\; T)$$

$\phi_{\mathcal{S}z\,\mathcal{S}z}(mT)$ = staircase cross-correlation function between the actual output and the desired output.

$$= \frac{1}{2N+1-m} \sum_{n=-N}^{N+1-m} z(nT)\, y(\overline{n+m}\; T)$$

and $\phi_{\mathcal{S}y\,\mathcal{S}y}(mT)$ = staircase auto-correlation function of the actual output y(t).

Prasad has shown in his work that the linear filter may be optimized by finding $u_r$, its P-response ordinates, by solving the following set of simultaneous equations:

$$\sum_{r=o}^{N} u_r\, \phi_{\mathcal{S}x\,\mathcal{S}x}(\overline{r-s}\; T) = \phi_{\mathcal{S}x\,\mathcal{S}z}(sT)$$

for s = 0, 1, 2, ......, N. (2.09)

With the aid of a digital computer these can be solved conveniently, and the mean-square error may be computed by using the following relationships for the linear filters:

$$\phi_{\mathcal{S}y\,\mathcal{S}y}(0) = \sum_{r=o}^{N} \sum_{s=o}^{N} u_r u_r\, \phi_{\mathcal{S}x\,\mathcal{S}x}(\overline{r-s}\; T) \qquad (2.10)$$

and

$$\phi_{\mathcal{S}z\,\mathcal{S}y}(0) = \sum_{r=o}^{N} u_r\, \phi_{\mathcal{S}x\,\mathcal{S}z}(-rT) \qquad (2.11)$$

A comparison of these equations with the Wiener filter was made by Prasad in some numerical examples, and the resulting mean-square errors were about the same. However, the computational advantages of the staircase techniques are obvious.

In the case of an instantaneous power series filter, for which

$$y(t) = \sum_{\mu=1}^{M} a_\mu x^\mu$$

the optimum $a_\mu$'s may be calculated from the following set of simultaneous equations:

$$\sum_{\mu=1}^{M} a_\mu \phi_{\mu r}(0) = \phi_{rz}(0) \qquad (2.12)$$

for r = 1, 2, ..., M,

where $\phi_{\mu r}$ and $\phi_{rz}$ are nonlinear correlation functions defined as below

$$\phi_{\mu r}(kT) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} (x_n)^\mu\, (x_{n+k})^r \qquad (2.13)$$

and

$$\phi_{rz}(kT) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} (x_n)^r\, z_{n+k} \qquad (2.14)$$

The mean square error in this case is given by

$$\overline{\epsilon^2}_{\min} = \phi_{\mathcal{S}z\,\mathcal{S}z}(0) - 2\sum_{\mu=o}^{M} a_\mu \phi_{rz}(0) + \sum_{\mu=o}^{M} \sum_{r=o}^{M} a_\mu a_r \phi_{\mu r}(0) \qquad (2.15)$$

In the general case of a no-memory multipith nonlinear filter shown in Fig. 3, the optimizing equations are

$$\sum_{\mu=o}^{M} a_\mu\, \phi_{\mathcal{S}f_\mu\,\mathcal{S}f_r}(0) = \phi_{\mathcal{S}f_r\,\mathcal{S}z}(0) \qquad (2.16)$$

where the nonlinear correlation functions are defined as

$$\phi_{\mathcal{S}f_\mu\,\mathcal{S}f_r}(kT) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} f_\mu(x_n)\, f_r(x_{n+k}) \qquad (2.17)$$

and

$$\phi_{\int f_r \int z}(kT) = \lim_{N} \frac{1}{2N+1} \sum_{n=-N}^{N} f_r(x_n)\, z_{n+k} \qquad (2.18)$$

The mean-square error is given by

$$\overline{\epsilon^2_{\min}} = \phi_{\int z \int z}(0) - 2 \sum_{\mu=0}^{M} a_\mu \phi_{\int f_r \int z}(0) + \sum_{\mu=0}^{M} \sum_{r=0}^{M} a_\mu a_r \phi_{\int f_\mu \int f_r} \qquad (2.19)$$

Finally, for the storage nonlinear filter shown in Fig. 2, the optimizing equations are given by

$$\sum_{r=0}^{N} u_r\, \phi_{\int f \int f}(|r-s|\, T) = \phi_{fz}(sT)$$

$$\text{for } s = 0, 1, 2, \ldots\ldots, N, \qquad (2.20)$$

where

$$\phi_{\int f \int f}(|r-s|\, T) = \frac{1}{N-s+1} \sum_{p=0}^{N-s} f[x_p]\, f[x_{p+s-r}] \quad \text{for } r < s$$

$$= \frac{1}{n-r+1} \sum_{p=0}^{N-r} f[x_p]\, f[x_{p+r-s}] \quad \text{for } r > s$$

$$= \frac{1}{n-r+1} \sum_{p=0}^{N-r} f[x_p]\, f[x_p] \quad \text{for } r = s \qquad (2.21)$$

## Comparison of Different Types of Filters for a given Random Input

Starting with 1000 samples of a given non-Gaussian signal, the random input x(t) was obtained by superimposing 1000 samples of a random Gaussian noise. The various linear and nonlinear correlation functions, defined in the previous section, were then calculated. These were used to calculate the following

(a) the optimum linear filter

(b) The optimum instantaneous power-series filter with the first five terms only.

(c) the optimum instantaneous power-series filter with the first ten terms.

(d) optimum nonlinear storage filter derived from the one in (b) followed by a suitable linear filter,

and (e) optimum nonlinear storage filter derived from the one in (c) followed by a suitable linear filter.

The mean-square was calculated for each of these optimum filters, and the values, normalized with respect to that for the optimum linear filter, are shown below:

| Serial Number | Type of filter | Mean-square error, normalized with respect to that for the linear filter |
|---|---|---|
| 1. | Linear | 1.00 |
| 2. | Instantaneous nonlinear power series with the first five terms. | 1.65 |
| 3. | Instantaneous nonlinear, power series with the first ten terms. | 1.12 |
| 4. | Storage nonlinear filter (#2 followed by a suitable linear filter) | 0.85 |
| 5. | Storage nonlinear filter (#3 followed by a suitable linear filter) | 0.68 |

## Conclusions

It will be seen from the comparisons made in the previous section that, for the given random signal, the optimum nonlinear filters with storage give a considerably smaller mean-square error. It may be emphasized that only the simplest kinds of nonlinear filters have been considered. The staircase techniques may be applied to more complicated nonlinear filters, with the possibility of a smaller mean-square error. An interesting possibility is the use of the multipath nonlinear filter with storage, the calculations for which require a fairly large computer.

## Acknowledgements

The author wishes to express his gratitude to Dean John R. Ragazzini of New York University for providing him with opportunities for carrying out this research. The author is also indebted to Professor S. S. L. Chang of New York University and Dr. T. Prasad of Bihar Institute of Technology, India, for constant help and encouragement.

## Bibliography

1. Singleton, H. E., "Theory of nonlinear transducers," Research Laboratory of Electronics, M.I.T., Technical Report Number 60, August 12, 1950.

2. Bose, A. G., "A theory nonlinear systems," Research Laboratory of Electronics, M.I.T., Technical Report Number 309, May 15, 1956.

3. Rosenbrock, H. H., "A class of nonlinear filters giving least r.m.s. error". Costain-John Brown Ltd., London, Report Number I.D.R.-2, January 11, 1956.

4. Zadeh, L. A., "A contribution to the theory of nonlinear systems," The Journal of the Franklin Institute, Vol. 255, Number 5, May 1953.

5. Prasad, T., "Analysis and Optimization of a class of Nonlinear Staircase Systems," Doctoral thesis, Department of Engineering, University of Cambridge, England, August 1960.

6. Solodnikov, V. V. "Introduction to the statistical Dynamics of Automatic Control Systems," State Publishing House for Theoretical-Technical Literature, Moscow-Leningrad 1952, (translated into English by Zadeh).

7. Lanning, J. H. and Battin, R. H., "Random Processes in Automatic control" McGraw Hill Book Company, Inc. 1956.

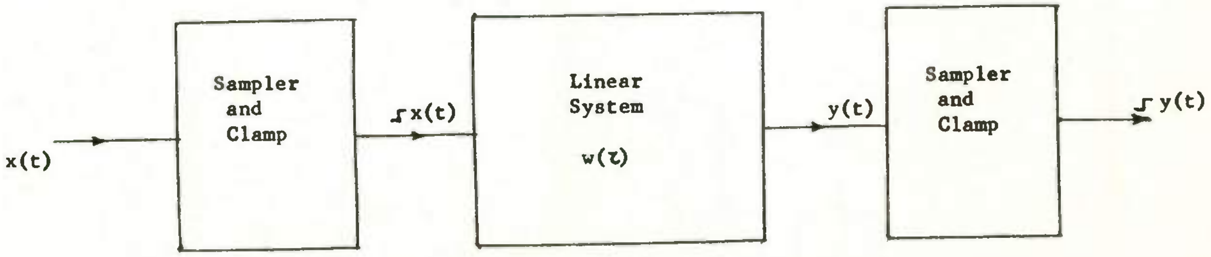


Fig. 1 Staircase Signals Applied To A Linear System

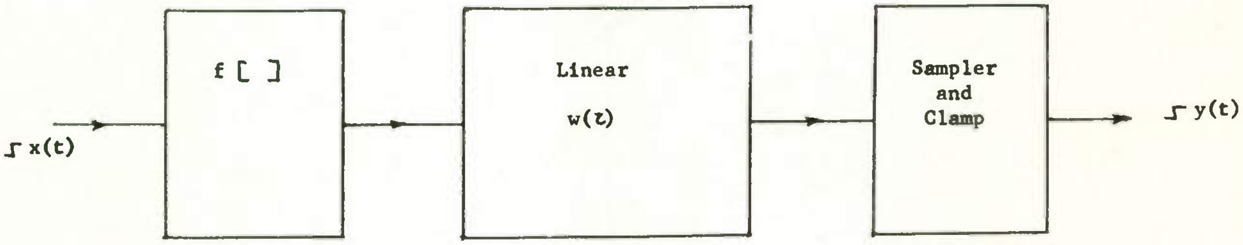


Fig. 2 Storage Nonlinear Staircase System

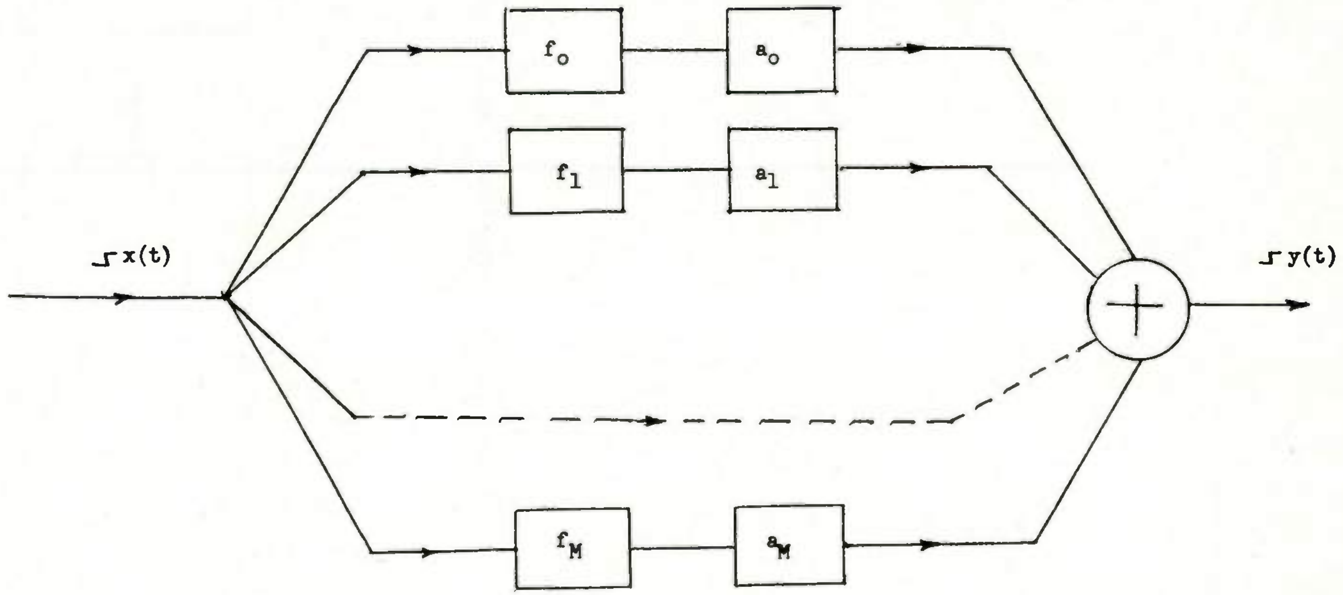


Fig. 3. General No-storage Multipath System

# DYNAMIC AND NOISE ERRORS IN LINEAR SERVOS

Stanley W. Gery

Airborne Instruments Laboratory
A Division of Cutler-Hammer, Inc.
Deer Park, Long Island, New York

## Abstract

The errors caused by noise and dynamics of an input variable are formulated for linear servo systems in terms of the servo noise bandwidth relative to the input channel bandwidth, the input signal-to-noise ratio, and the error constants of the servo. The input variable dynamics are assumed to be defined by a finite time-power series. The order of the servo is then chosen to make the steady-state dynamic error finite. The optimum servo bandwidth is formulated for minimum combined error, which is given in terms of the signal-to-noise ratio and the appropriate derivative of the input variable.

These results have particular application to radar tracking, AFC, phase-locked loops, etc., where the dynamic range of the error is bounded. This dynamic range, the signal-to-noise ratio, and the accelerations of the input are then mutually constrained and impose broad performance limitations on linear time-invariant servo systems.

## Introduction

In many communication and radar system applications, the servo input variable is remote from the servo and is conveyed by its modulation of a parameter of an electromagnetic wave. Demodulation reproduces the input variable with the thermal noise of the system causing apparent noisiness of the input variable.

Since the modulation bandwidth is usually much greater than the actual spectrum of the input variable, the servo is capable of reducing the apparent noise by averaging in time. As the servo bandwidth is reduced to affect this smoothing, the response to the input variable is impaired. Thus, conflicting requirements on the servo bandwidth exist. It is evident that the combined error due to motion of the input variable and to the thermal noise cannot be made arbitrarily small, and that some minimum combined error exists.

Most error detectors in such systems have limited dynamic range in the sense that they are linear and of proper slope only for small errors. For the servo to properly follow the input variable, the combined error must be less than the error dynamic range. Thus the need frequently arises in preliminary system considerations for a simple technique to evaluate the minimum combined error. The work presented here is intended to meet this need with relatively general and appropriate restrictions on the input variable and the shape of the servo transfer function.

## Steady-State Analysis

The systems under consideration have the form shown in Figure 1. They comprise an input variable, a carrier that is modulated by it, a communication link, an error detector, demodulator, and a servo forward path G(s) (Figure 2). By restraining the generality of these elements to a few parameters, a relatively simple method of evaluating combined noise and dynamic errors can be developed.

For the steady-state analysis, it is assumed that the input variable is in the form of a finite time power series (equations 1 and 2).

$$X(t) = \sum_{r=0}^{R} \frac{a_r t^r}{r!} \qquad (1)$$

$$X(s) = \sum_{r=0}^{R} \frac{a_r}{s^{r+1}} \qquad (2)$$

where the $a_r$'s are the maximum expected.

The servo model has the simplest transfer function that will result in a finite steady-state error in response to the input variable. The transfer function 3

$$H(s) = 1 - \left(\frac{s}{s + \omega_o}\right)^R \qquad (3)$$

is descriptive of an $R^{th}$ order servo.

The lag error is found from equations 2 and 3 by means of the final value theorem:

$$e_{lag} = \lim_{s \to 0} sX(s)[1 - H(s)] \qquad (4)$$

$$= \frac{a_R}{\omega_o^R} \qquad (5)$$

As would be expected, the lag error is proportional to the highest order acceleration and inversely proportional to the $R^{th}$ power of the servo bandwidth $\omega_o$.

The rms fluctuation of the output variable is formulated by first relating the error demodulator output thermal noise voltage to an apparent noise on the input variable. Then, the output noise is formed from the spectral character of the apparent input noise and the power frequency-response of the closed loop transfer function.

The error demodulator in Figure 1 usually produces a voltage in the form:

$$v_{dem} = \eta V_{sig}(x - y) + v_{noise} \qquad (6)$$

where

$\eta$ has the units of volts per unit error per volt of signal,

$(x - y)$ = error,

$v_{noise}$ = system noise voltage at the demodulator output,

$V_{sig}$ = signal voltage.

The use of some form of AGC permits the extraction of a voltage proportional to the error. The apparent error detected by the servo is

$$(x - y)_{apparent} = \frac{v_{dem}}{\eta V_{sig}} \qquad (7)$$

$$= (x - y) + \frac{v_{noise}}{\eta V_{sig}} \qquad (8)$$

The last term in equation 8 is interpreted as the apparent noise added to the input variable due to the thermal noise of the communication link. The rms value of this apparent noise is then

$$\sigma_x = \frac{1}{\eta\rho} \qquad (9)$$

$$\rho = \frac{V_{sig}}{\text{rms } v_{noise}} \qquad (10)$$

where $\rho$ is the signal-to-noise voltage ratio.

The parameter $\eta$ depends on the error detector design. It usually cannot be adjusted without causing the signal-to-noise ratio to also change, and there usually exists an optimum design which maximizes $\eta\rho$. This design procedure can, for example, take the form of selecting the width and shape of the "early" and "late" gates in the time discriminator of a radar range tracking servo or of prescribing the "squint" angle and radiation pattern of a monopulse radar angle tracking antenna.

The rms output noise is found from

$$\sigma_y = \sqrt{\int_{-\infty}^{\infty} F(\omega)\; H(j\omega)\; H^*(j\omega)\; dw} \qquad (11)$$

where

$\sigma_y$ = rms output noise,

$F(\omega)$ = power spectral density of $v_{noise}$.

Assuming that the noise spectrum is white out to a band limit frequency $B_1$, which is very much greater than the servo bandwidth $\omega_o$, equation 11 can be approximated by

$$\sigma_y = \frac{1}{\eta\rho}\sqrt{\frac{B_R}{B_1}} \qquad (12)$$

where $\rho$ = S/N, the signal-to-noise voltage ratio, and

$$B_R = \int_{-\infty}^{\infty} H(j\omega)\; H^*(j\omega)\; df \qquad (13)$$

For the transfer function given by equation 3, the servo noise bandwidth, $B_R$, is proportional to $\omega_o$.

$$B_R = \int_{-\infty}^{\infty} df \left[ 1 - 2 \left( \frac{\omega^2}{\omega^2 + \omega_o^2} \right)^{\frac{R}{2}} \times \cos R \left( \frac{\pi}{2} - \tan^{-1} \frac{\omega}{\omega_o} \right) + \left( \frac{\omega^2}{\omega^2 + \omega_o^2} \right)^{R} \right] \tag{14}$$

$$= \omega_o \int_{-\pi/2}^{\pi/2} \frac{1}{2\pi} \left[ 1 - 2 \sin^R x \cos R \left( \frac{\pi}{2} - x \right) + \sin^{2R} x \right] \sec^2 x \, dx \tag{15}$$

$$= K_R \cdot \omega_o \qquad (16)*$$

where $K_R$ is the integral in equation 15. The rms output noise is then

$$\sigma_y = \frac{1}{\eta\rho} \sqrt{\frac{K_R \omega_o}{B_1}} \tag{17}$$

Since the servo bandwidth is very small compared with the input bandwidth, $B_1$, the probability distribution for the output noise is very nearly gaussian, with the mean value given by the lag error, and with the standard deviation given by equation 17.

The probability that the magnitude of the instantaneous error will, in the steady state, exceed a limit L is then

$$P = 1 - \frac{\eta\rho \sqrt{B_1}}{\sqrt{2\pi K_R \omega_o}} \int_{-L}^{L} dx \exp - \left[ \frac{\eta^2 \rho^2 B_1}{2 K_R \omega_o} \left( x - \frac{a_R}{\omega_o^R} \right)^2 \right] \tag{18}$$

Minimizing this probability that the error will exceed a preset limit, by means of adjusting the bandwidth $\omega_o$, is approximately effected by minimizing the sum of the standard deviation and the lag error which is called the combined error, $e_T$,

$$e_T = \frac{a_R}{\omega_o R} + \frac{1}{\eta\rho} \sqrt{\frac{\omega_o K_R}{B_1}} \tag{19}$$

Setting the derivative of $e_T$ with respect to $\omega_o$ equal to zero yields the optimum bandwidth under these conditions as

$$\omega_{opt} = \left( \frac{4R^2 B_1 \eta^2 \rho^2 a_R^2}{K_R} \right)^{\frac{1}{2R+1}} \tag{20}$$

for R = 1,

$$\omega_{opt_1} = \left( 8 B_1 \eta^2 \rho^2 a_1^2 \right)^{1/3} \tag{21}$$

for R = 2,

$$\omega_{opt_2} = \left( \frac{64}{5} B_1 \eta^2 \rho^2 a_2^2 \right)^{1/5} \tag{22}$$

The minimum combined error is then

$$e_{T_{min}} = (2R + 1) \left( \frac{K_R a_R^{\frac{1}{R}}}{4R^2 B_1 \eta^2 \rho^2} \right)^{\frac{R}{2R+1}} \tag{23}$$

for R = 1,

$$e_{T1_{min}} = 1.5 \left( \frac{a_1}{B_1 \eta^2 \rho^2} \right)^{1/3} \tag{24}$$

for R = 2,

$$e_{T2_{min}} = 1.8 \left( \frac{\sqrt{a_2}}{B_1 \eta^2 \rho^2} \right)^{2/5} \tag{25}$$

These results for the minimum combined error are plotted in Figure 3 for the cases of constant velocity input to a first order servo, and of a constant acceleration input to a second order servo. If the displacement dimension of the acceleration term is normalized to units of the error demodulator dynamic

* $K_1 = \frac{1}{2}$, $K_2 = \frac{5}{4}$, $K_3 = \frac{41}{32}$.

range, then the minimum combined error is automatically compared with the dynamic range.

For example, consider a type one radar angle tracking servo with a target having a constant velocity of one beamwidth per second, with a signal-to-noise ratio of 10 db, with $\eta$ equal to one volt per beamwidth per volt of signal, and with a PRF of 200 pulses per second making B, equal to 100 cps. Then the abscissa in Figure 3 is 0.001 and the minimum combined error is 0.15 beamwidth. If the signal-to-noise ratio were 30 db lower--that is, were it minus 20 db--the minimum combined error would be 1.5 beamwidths for which tracking would not be possible. The tradeoffs in transmitter power, antenna gain, noise figure, etc. are facilitated by use of these results. For instance, in the example above, it can quickly be seen that the minimum combined error in beamwidths varies inversely as the sixth root of the antenna gain and inversely as the cube root of the transmitter peak power, assuming a pencil beam.

These results assume that the lag error and the signal-to-noise ratio are stationary, or vary only very slowly compared with the servo time constant. In the latter case, it may be desirable to automatically tune the servo bandwidth to optimum in accordance with predictions or measurements of the accelerations and/or the signal-to-noise ratio.

In some instances, the initial values of the derivatives of the input variable are predictable to some accuracy. Then, a noise-free prediction signal

$$y(t) = \sum_{r=0}^{R} \frac{a_r(1 - p_r)t^r}{r!} \qquad (26)$$

(where $p_r$ is the error in the prediction of $a_r$) can be made available and subtracted, leaving as the net input variable

$$x_p(t) = \sum_{r=0}^{R} \frac{p_r a_r t^r}{r!} \qquad (27)$$

Substituting $p_R a_R$ instead of $a_R$ into the preceding results extends them to include this use of prediction.

## Transient Case

By making the order of the servo one unit higher than the highest order nonzero derivative of the input variable, the steady-state lag error is reduced to zero. The instantaneous dynamic error is

$$e(t) = \sum_{r=0}^{R-1} \frac{a^r}{(R - 1)!} \left\{ \frac{d^{R-r-1}}{dt^{R-r-1}} \left[ t^{R-1} \epsilon^{-\omega_o t} \right] \right\} \qquad (28)$$

Assuming that $a_r = 0$ for $r \neq R - 1$ results in the error given by equation 29

$$e(t) = \frac{a_{R-1} \epsilon^{-\omega_o t} t^{R-1}}{(R - 1)!} \qquad (29)$$

which has a maximum given by equation 30.

$$e_{max} = \frac{a_{R-1}}{(R - 1)!} \left( \frac{R - 1}{\epsilon \omega_o} \right)^{R-1} \qquad (30)$$

Thus, when the input variable is a pure (R - 1)th power parabola, the peak transient error of the assumed Rth order servo varies inversely as the (R - 1)th power of the servo bandwidth $\omega_o$.

Taking the combined error as the sum of the peak error and the standard deviation of the output noise permits the determination of an optimum bandwidth with respect to the acquisition error. Thus, with

$$\hat{e}_T = \frac{a_{R-1}}{(R - 1)!} \left( \frac{R - 1}{\epsilon \omega_o} \right)^{R-1} + \frac{1}{\eta \rho} \sqrt{\frac{K_R \omega_o}{B_1}} \qquad (31)$$

then

$$\hat{\omega}_{opt} = \left[ \frac{2\eta\rho a_{R-1}(R - 1)^R}{\epsilon^{R-1}(R - 1)!} \sqrt{\frac{B_1}{K_R}} \right]^{\frac{2}{2R-1}} \qquad (32)$$

and

$$\hat{e}_{T_{min}} = \frac{(2R - 1)(R - 1)^{R-1}}{(R - 1)! \, \epsilon^{R-1}} \times \left\{ \frac{K_R \epsilon^{2R-2} \left[ (R - 1)! \right]^2 a_{R-1}^{\frac{1}{R-1}}}{4(R - 1)^{2R} B_1 \eta^2 \rho^2} \right\}^{\frac{R-1}{2R-1}} \qquad (33)$$

and the residual steady-state rms noise output is

$$\sigma_y = \left[\frac{2a_{R-1}(R - 1)^R \; K_R^{R-1}}{\epsilon^{R-1}(R - 1)! \; B_1^{R-1} \; (\eta\rho)^{2R-2}}\right]^{\frac{1}{2R-1}} \qquad (34)$$

These results are useful in evaluating the assumed type 2 servo in acquiring a constant velocity input with zero initial position error. The results for R = 2 are

$$\hat{\omega}_{opt_2} = \left(\frac{4}{\epsilon} \sqrt{\frac{B_1}{5}} \; \eta\rho a_1\right)^{2/3} \qquad (35)$$

$$\hat{e}_{T2_{min}} = 1.46 \left(\frac{a_1}{B_1\eta^2\rho^2}\right)^{1/3} \qquad (36)$$

$$\sigma_{y_2} = 0.975 \left(\frac{a_1}{B_1\eta^2\rho^2}\right)^{1/3} \qquad (37)$$

The errors are plotted in Figure 4.

In some applications, it would be undesirable to have a steady-state noise error of about half the dynamic range, as indicated by the ratio of (37) to (36). Reduction of the servo bandwidth at a rate and to an amount consistent with stability and practical considerations is therefore indicated after acquisition.

## Conclusions

Although these results lack generality, they have the advantage of simplicity and rapid preliminary estimation of system requirements and performance capabilities.

The extension of the method of minimizing the sum of the dynamic and rms noise errors to other transfer functions and other classes of input variables could follow the same procedure used above. The considerations given here might also be useful as criteria for adaptive servo design.

It has been tacitly assumed that the principal sources of error in the system were thermal noise and dynamic errors. If other sources of error exist, or if it is desirable to smooth actual noise on the input variable such as the presence of glint in radar angle tracking, somewhat different considerations apply (reference 1).

## Bibliography

1. James, Nichols, and Phillips, "Theory of Servomechanisms," McGraw-Hill, New York, 1947.
2. W. R. Bennett, "Methods of Solving Noise Problems," Proc IRE, Vol 44, p 609-638, May 1956.
3. R. S. Raven, "Techniques for Signal and Noise Analysis," Chapter 5 in Airborne Radar, G. Merril, Ed, Van Nostrand, 1961.
4. R. P. Cheetham and W. A. Mulle, "Enhanced Real Time Data Accuracy for Instrumentation Radars by Use of Digital-Hydraulic Servos," IRE Wescon Convention Record, Part 4, 1958.
5. M. F. Gardner and J. L. Barnes, "Transients in Linear Systems," Vol I, Wiley, 1942.
6. G. D. Young, "Radar Velocity Tracking Systems Analysis," in Ballistic Missile and Space Technology, Vol III, D. P. LeGalley, Ed, Academic Press, 1960.
7. W. J. Gruen, "Theory of AFC Synchronization," Proc IRE, Vol 41, p 1043-1049, August 1953.

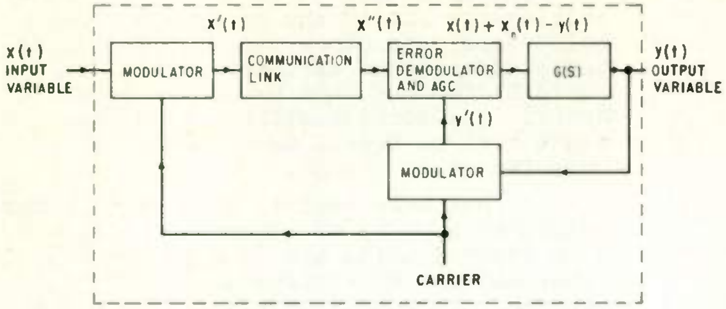


Fig. 1. Servo system model.

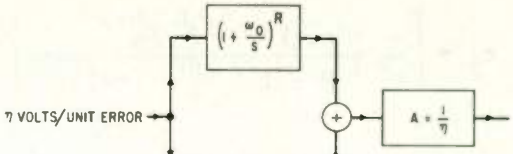


Fig. 2. Possible configuration for G(S) to obtain the desired transfer function.

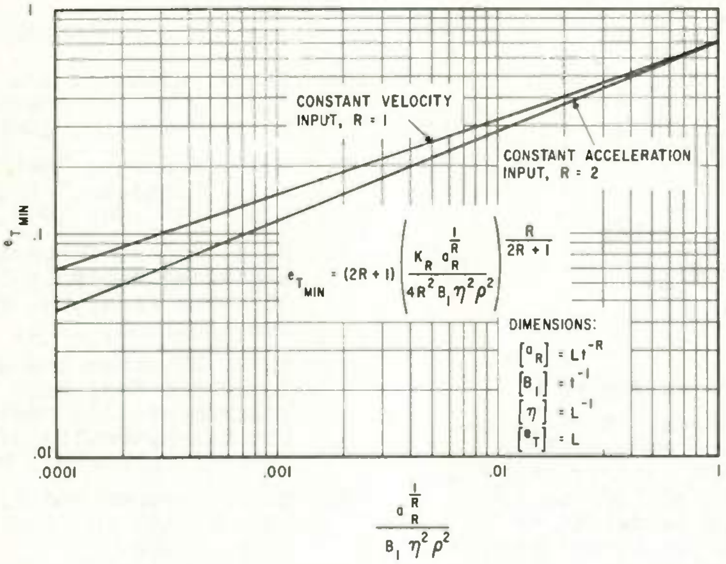


Fig. 3. Steady-state minimum combined error.

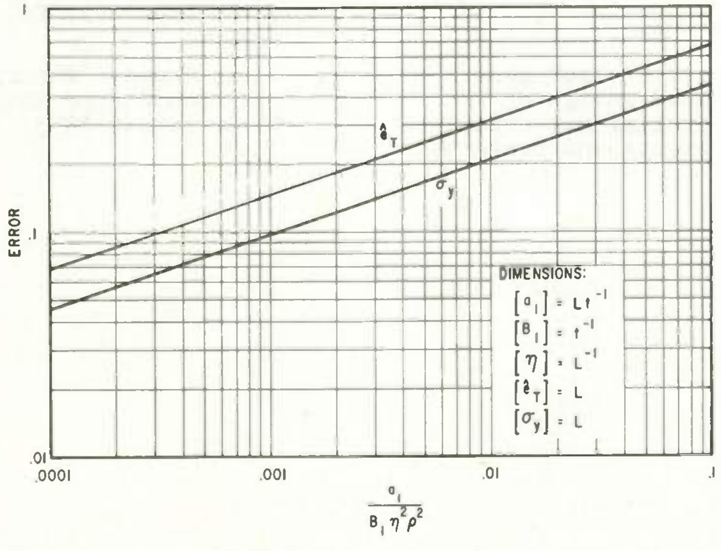


Fig. 4. Minimum combined peak error, $\hat{e}_T$ and residual output-noise, $\sigma_y$.

# DESIGN OF RELAY TYPE SAMPLED-DATA CONTROL SYSTEMS USING DISCRETE DESCRIBING FUNCTION

Benjamin C. Kuo
University of Illinois
Urbana, Illinois

## Summary

The discrete describing function and its use in the analysis and design of nonlinear sampled-data systems are presented in this paper. For a nonlinear sampled-data system, the input and output signals of the nonlinear element may be in the form of pulse trains. Therefore, it is natural to define a "discrete describing function" N(z) which is equal to the ratio of the z-transform of the output to the z-transform of the sinusoidal pulse-modulated input of the nonlinear element. In this paper, the discrete describing function for a relay with dead zone is derived; although, using the same method, other types of amplitude-dependent nonlinearities can be treated in similar fashions. The discrete describing function N(z) is used to derive the critical regions of -1/N(z) which correspond to the critical point (-1,jo) for linear continuous-data systems. Stability study of the nonlinear sampled-data system is made by investigating the relative positions of the critical regions and the linear transfer locus $G_1(z)$ of the system. The effects of varying the gain and the sampling period on system stability is readily observed. Reshaping the linear transfer loci by digital or continuous-data controllers may be done in the usual manner using z-transforms.

The discrete describing function has at least the following advantages:

1. The discrete describing function is natural for sampled-data systems; it is analogous to the use of the conventional describing function for continuous-data systems.

2. The method can be applied to sampled systems with or without hold devices.

3. Systems with more than one sampler can be studied.

4. The actual output of the nonlinear element, rather than the fundamental component of the Fourier series representation, is used.

5. Compensation with digital controllers as well as with continuous-data networks can be designed in a straight forward manner.

## Introduction

The study of relay-type sampled-data systems by means of describing function technique has been made previously by Chow[1] and Russell[2]. In these early investigations, the nonlinear element, for which the describing function is derived, is considered to include the sampling switch, the zero-order hold and the relay. For a sinusoidal input whose period is an integral multiple of the sampling period T. the output of the nonlinear element is a periodic rectangular wave. Thus, the conventional describing function technique ordinarily used for continuous-data systems can be applied directly, and the nonlinear sampled-data system is essentially treated as a nonlinear continuous-data system. These early studies are subjected to the limitations that a zero-order hold device must be present and there is only one error-sampling switch in the system.

The nonlinear linear system under investigation is shown in Fig. 1. The sampler is assumed to be ideal, which means that the output of the sampler, e*(t), is an impulse train. The operational characteristic of the relay is shown in Fig. 2.

If the error signal e(t) is sinusoidal with period Tc = nT, where n = 2, 3, ....., and T is the sampling time in seconds, e*(t) must also be a periodic function; although it may not have the same frequency as e(t). Since e*(t) can have values only at the sampling instants, the period of e*(t) must also equal nT. Thus, based on the characteristics of the zero-order hold H, the relay R, and G(s), the time functions h(t), m(t), c(t) must have the same period as e*(t). This is a very important feature of this nonlinear system. Under the condition of self-sustained oscillation r(t)=o; hence

$$e(t) = r(t) - c(t) = -c(t) \qquad (1)$$

which means that e(t) must also have the same period as c(t).

## Modification of System Configuration

The z-transform of the output of the system shown in Fig. 1 is written as:

$$C(z) = E(z) \left[ NG(z) \right] \qquad (2)$$

where NG(z) denotes the z-transform of N(s) G(s). The overall transfer function of the system is:

$$\frac{C(z)}{R(z)} = \frac{NG(z)}{1+NG(z)} \qquad (3)$$

It is clear that the describing function technique is not very useful here since NG(z) cannot be separated into two functions of N and G. A modified block diagram of the system is suggested in Fig. 3, in which, the zero-order hold and the relay are transposed. From the analytical point of view, the system behavior is not altered by this modification. In this case, the nonlinear element N is considered to include only the relay R whose output v*(t) is a train of impulses having constant amplitudes. The input to G(s) is not affected by the change. In terms of the z-transform, the analytic description of the modified system in Fig. 3 takes the following form:

$$\frac{C(z)}{R(z)} = \frac{N(z)G_1(z)}{1+N(z)G_1(z)} \qquad (4)$$

where N(z) is defined as the discrete describing function of the relay, and $G_1(z)$ is the pulse transfer function of the zero-order hold and the linear system G connected in cascade. Now, the study of the stability of the nonlinear sampled-data system involves the investigation of the equation

$$1+N(z)G_1(z) = 0 \qquad (5)$$

or

$$G_1(z) = - \frac{1}{N(z)} \qquad (6)$$

## The Discrete Describing Function and the Critical Regions[3,4]

The derivation of the discrete describing function N(z) is based on the assumption that the input signal to the sampler is a sinusoid. Consequently, the input to the relay in Fig. 3 is a sinusoidally modulated impulse train. N(z) is defined as the ratio of the z-transform of the output v*(t) to the z-transform of the sinusoidally modulated input of the relay, e*(t); that is

$$N(z) = \frac{V(z)}{E(z)} \qquad (7)$$

Suppose that the input to the sampler is given by

$$e(t)=E\cos(at+\phi)=E(\cos\phi\cos at - \sin at \cos\phi) \qquad (8)$$

The z-transform of e(t) is

$$E(z)= \frac{Ez}{z\ -{}^{2}z\cos aT+1} \times \left[(z-\cos aT)\cos\phi - \sin aT \sin\phi\right] \qquad (9)$$

Since the period of the self-sustained oscillation is an integral multiple of T, only these periods are considered in deriving N(z). This N(z), when applied to Eq. (6), will define the conditions for self-sustained oscillations to occur in the system.

Referring to the block diagram of Fig. 3, the output of the zero-order hold, m(t), is either constant or zero between any two successive sampling instants. For a given period Tc of c(t), m(t) may have many possible forms. Figure 4 illustrates some possible forms of n(t) for Tc=4T. However, since the error signal e(t) is assumed to be sinusoidal having period Tc=4T, only a few of the waveforms of n(t) illustrated in Fig. 4 may occur in the system. In fact, when the phase shift of e(t) is varied from 0° to 360° and Tc =4T, m(t) can have only the waveforms of Figs. 4(a), (b) and those of (a) and (b) shifted by nT (n= 1. 2. ...). It is important to note that the function of Fig. 4(b) has one positive and one negative relay correction during each period Tc; the waveform of Fig. 4(c) has two successive positive and two negative relay corrections during each period. These have to be considered separately, even though they have the same period Tc. Similarly, it can be shown that for all Tc=nT, n=even integers ,the number of positive relay correction signals is equal to that of the negative relay corrections during one period Tc. This number of correction is designated as Δ . For Tc=2T, it is apparent that Δ can only be unity. In general, for Tc=nT, n=even integers, the values of Δ can be 1, 2, 3, ... (n-1).

For a given period of Tc=nT, n=even integers, the loci of -1/N(z) form regions in the decibel versus phase shift plot, each for one possible value of Δ . These regions are defined as the critical regions for the specific

$T_c$ and $\Delta$. The symbols

$$-\frac{1}{N(z)}\bigg|_{\max} \text{ and } -\frac{1}{N(z)}\bigg|_{\min}$$

are used to indicate the boundaries of the regions. It can be shown that all the periodic functions generated by shifting the waveform of Fig. 4(a) have the same critical region. This means that the study of the conditions of self-sustained oscillation is independent of the phase shift of m(t) once $T_c$ and $\Delta$ are given.

However, for $T_c = nT$, n = odd integers, the number of positive and negative relay corrections may be different. In this case, the designation of the type of oscillation by $\Delta = 1, 2, 3, \ldots,$ is inadequate. It is necessary to designate this type of oscillation by $\Delta_{ij}$, where i is the number of positive relay corrections and j is the number of negative relay corrections during each period.

The typical procedure for the derivation of $-1/N(z)$, $-1/N(z)|$ max, and $1/N(z)|$ min is given in Appendix I for Tc = 4T. The expressions for $-1/N(z)$, $-1/N(z)|$ max, and $-1/N(z)|$ min for the relay type nonlinearity are derived for Tc = 2T, 3T, ... 6T, and are tabulated in Table I.

The corresponding critical regions are plotted in decibels versus phase shift in Fig.5 through Fig. 9. By use of the same principle and procedure, the critical regions of $-1/N(z)$ for Tc = 7T, 8T, ... can be obtained if necessary.

Examination of the expression derived for $-1/N(z)$ and the conditions on e(t) to give each value of $\Delta$, leads to the following conclusions:

1. The discrete describing function N(z) of the relay under consideration is a function of the frequency a, amplitude E, and phase shift $\phi$, of the input sinusoid e(t), and a function of the relay dead zone D.

2. The regions bounded by the loci of $-1/N(z)|_{\max}$ and $-1/N(z)|_{\min}$ in the amplitude (decibel) versus phase shift plot for $T_c = nT$ (n = 2, 3, ...) are symmetrical about the $-180^{\circ}$ axis. The maximum widths for the critical regions are $(2\pi/T_c)T$ for $T_c$ = 4T, 6T, 8T,... nT, n = even integers, and $(2\pi/T_c)(T/2)$ for $T_c = nT$, n = odd integers. When n is very large, the critical regions become very narrow, and finally approaching a straight line along the $-180^{\circ}$ axis as n approaches infinity.

3. For certain values of E and $\Delta$, $-1/N(z)$ is infinite, and the critical regions extend to infinity (open region). The asymptote of the $-1/N(z)|_{\max}$ boundaries is always the -180 deg. axis, while the asymptotes of the $-1/N(z)|_{\min}$ boundaries are $-\pi \pm (2\pi/T_c)(T/2|_{T_c=nT}$ for n-even integer > 2, and $-\pi \pm (2\pi/T_c) \times (T/4)|_{T_c=nT}$ for n=even integers greater than 1.

## Stability Study and Limit Cycles Using Gain-Phase Plot and Discrete Describing Function

The condition of self-sustained oscillation in the System of Fig. 3 is defined as when

$$G_1(z) = -1/N(z) \qquad (10)$$

Since the left-hand side of Eq. (10) is the z-transfer function of the linear plant, and the right-hand side consists of the negative inverse of the discrete describing function of the nonlinear element, the stability analysis of the nonlinear system with sampled-data follows the well-known procedure of the describing function and gain-phase plane studies of a nonlinear system with continuous-data. In terms of the gain-phase plot of $G_1(z)$, the graphical stability study of the sampled-data system consists of the following steps:

1. Plot $G_1(z)$ on gain-phase coordinates for $T_c$ = 2T, 3T, 4T, .... using T as a parameter on the loci.

2. Superpose on the gain-phase coordinates the family of critical regions of $-1/N(z)$ for $T_c$ = 2T, 3T, 4T, ....

The following conclusions can be reached from the inspection of the relative position of the $G_1(z)$ loci and the critical regions:

1. If a portion of the $G_1(z)$ locus for some $T_c$ falls within the critical region of this $T_c$ for a certain relay dead zone D, (when D varies, the critical regions of $-1/N(z)$ simply shift up or down along the = -180 degree axis) then there exists a set of E, $\phi$, $\Delta$ and $T_c$, such that

$$G_1(z) = -1/N(z)$$

Consequently, for any sampling period T, the portion of the $G_1(z)$ locus that lies in the corresponding critical region will produce a self-sustained oscillation at $T_c$, $\phi$ and E, characterized by $\Delta$, and the system is unstable.

2. If the same $G_1(z)$ locus for more than one value of $T_c$ falls within their respective critical regions, for the same value of D, the system can have more than one mode of oscillation.

3. The T's along the portion of $G_1(z)$ locus outside the corresponding critical region provide a stable system.

One special condition is when the relay dead zone D becomes Zero, and the relay is considered to be ideal. From the physical viewpoint, the relay will provide a corrective output whenever there is a signal at its input, no matter how small this signal may be. Under this condition, some types of oscillation disappear in the nonlinear system, and some can only appear for several discrete values of $\Delta$. For instance, when D=0, the sustained oscillation of the type, $T_c = 4T$ and $\Delta = 1$ can only occur in the system when the phase angle $\phi$ of e(t) is a multiple of $\pm \pi/2$. This is observed by referring to Fig. 12-39 in which the critical regions for $\Delta = 1$ and 2 are shifted to $-\infty$ along the -180 deg. axis when D becomes zero. Evidently, the critical region for $\Delta = 2$ degenerates into a region which is bounded by two vertical straight lines with the left and the right hand boundaries located at -225 deg. and -135 deg., respectively. The critical region for $\Delta = 1$ is degenerated into a single vertical line along the -180 deg. axis.

For $T_c = 6T$ and D = 0, self-sustained oscillation of the type, $\Delta = 1$ can no longer exist. This isdue to the period of 6T of the sinusoidal e(t). This e(t) will produce more than two relay corrective signals in either direction during one period, while $\Delta = 1$ specifies that there can be only one positive and one negative relay corrections during each period $T_c$. Referring to Fig. 12-41, the critical region for $\Delta = 1$ is a closed region; when this closed region is shifted to $-\infty$ along the -180 deg. axis, it is no longer expected to enclose any portion of the $G_1(z)$ locus for T = 6T. Similarly, the critical region for $\Delta = 3$ becomes a vertical strip centered along the -180 deg. axis, with its left and right hand boundaries located at the -210 deg. and -150 deg. lines, respectively. The critical region for $\Delta = 2$ is narrowed down to a vertical line along the -180 deg. axis. For $T_c = 8T$, the sustained oscillations are characterized only by $\Delta = 3$ and $\Delta = 4$, and those of $\Delta = 1$ and 2 are eliminated.

The following numerical example illustrates the application of the critical regions in the sinusoidal study of a nonlinear sampled-data control system.

## Illustrative Example

Consider that the open-loop transfer function of the relay type sampled-data system of Fig. 3 is given by

$$G(s) = \frac{1}{s(s+1)} \qquad (11)$$

The sampling period is one second, and the relay has a dead zone of 0.1. With reference to the system configuration of Fig. 3, we have

$$G_1(s) = \frac{1}{s(s+1)} \times \frac{1 - e^{-Ts}}{s} \qquad (12)$$

The z-transform of $G_1(s)$ is given by

$$G_1(z) = \frac{z(T - 1 + e^{-T}) + 1 - e^{-T}(T+1)}{(z-1)(z-e^{-T})} \qquad (13)$$

The loci of $G_1(z)$ with T as a variable parameter are plotted in Fig. 10 for $T_c$ = 2T, 3T, 4T, ... 8T. For the relay dead zone specified (D=0.1), the normalized critical regions of Figs. 5 through 9 are shifted down by 20 db. Then, superposition of these shifted critical regions on the $G_1(z)$ loci of Fig. 10 shows that four modes of oscillations are possible in this system when T=1 sec: The results are tabulated in Table II.

These results are obtained by observing that when D=0.1, the T=1 sec. points on the $G_1(z)$ loci fall inside their respective critical regions only forthe four above listed modes of oscillations. When D=1, the critical regions of Figs. 5 through 9 are not shifted; it is seen that all the T=1 sec. points on the $G_1(z)$ loci fall outside their respective critical regions, and the system is always stable (without sustained oscillation).

The amplitude of oscillation of c(t) can be computed directly from the $G_1(z)$ loci. For r(t)=0, the amplitude of c(t) is equal to the amplitude of the error e(t). Therefore,

$$E = |e(t)| = |c(t)| \qquad (14)$$

Also, self-sustained oscillations occur when

$$G_1(z) = -1/N(z) \qquad (15)$$

For certain $T_c$ and T, z is a point on the unit circle in the z-plane, say, $z_1$, we can write

$$|G_1(z)| = |-1/N(z_1)| \qquad (16)$$

The values of $|-1/N(z)|$ tabulated in Table I suggest a simple method of predicting the amplitude of oscillation of c(t) from the values of $G_1(z_1)$. From Table I, it is observed that $|-1/N(z)| = kE$, where k is a constant for any $T_c$. From Eqs. (14) and (16) we have

$$E = \frac{1}{k} G_1(z_1) \tag{17}$$

Equation (17) implies that whenever the $G_1(z)$ locus for a certain $T_c$ is given, the locus of the amplitude of oscillation of c(t) is related to the $G_1(z)$ locus by only a constant factor k. However, it should be kept in mind that Eq. (17) is valid only for the portion of $G_1(z)$ which lies inside the critical regions for the same $T_c$. For example, for $T_c = 4T$, and $\Delta = 1$, from Table I,

$$|-1/N(z)| = E \tag{18}$$

Therefore, Eq. (17) gives

$$E = |G_1(z)|_{T_c = 4T} \tag{19}$$

which means that when the system has a self-sustained oscillation which is characterized by $T_c = 4T$ and $\Delta = 1$, every point on the $G_1(z)$ locus inside the critical region in the gain-phase plot represents the amplitude of c(t) for the corresponding sampling period T. In the present example, for T=1 sec., the following value of E is obtained from the $G_1(z)$ locus in Fig. 7:

$$E = |G_1(z)|_{T_c = 4 \text{ sec}} = 0.31 \quad (-10 \text{ db}) \tag{20}$$

Similarly, for $T_c = 4T$ and $\Delta = 2$, Table I gives

$$|-1/N(z)| = 0.707E = |G_1(z)|_{T_c = 4T} \tag{21}$$

Therefore, for T=1 sec., E=0.31/0.707=0.438

It is interesting to compare the results which are obtained in this example by use of the sinusoidal analysis with describing function to those of the same system which are computed by the difference equation method 5. While the difference equation approach gives only the solution to a particular set of initial conditions, it is extremely difficult to determine what are the possible modes of oscillations under various initial conditions. However the describing function method introduced in this paper does point out that for the system under study, with T=1 sec. and D=0.1, only the four modes of oscillations listed in Table II are possible under any initial condition. The results in Table II show that the amplitudes of self-sustained oscillation which are predicted by the two different methods are quite close, except for mode 2 when $T_c = 4$ sec. and $\Delta = 2$. It should be pointed out that both the difference equation and the discrete describing function methods deal with the system response at the sampling instants only. Furthermore, the describing function method assumes that the input to the sampler is Sinusoidal and only the fundamental component of the relay output is considered to be significant.

In reality, the oscillations in a nonlinear system are seldom sinusoidal. Therefore, the accuracy of the sinusoidal analysis depends entirely on how much the output c(t) differs from a sine wave. In fact, the amplitudes of oscillation listed in Table II which are determined by the discrete describing function are extremely close to the results obtained by Chow[1] using the conventional describing function method.

## Nonlinear Systems with More Than One Synchronized Sampler

One distinct advantage of the discrete describing function method is that it can be applied to nonlinear system with more than one synchronized sampler. Figure 11 shows the block diagram of a nonlinear sampled-data system with two samplers which are synchronized to open and close at the same time. The closed-loop transfer function of this multi-sampler system is

$$\frac{C(z)}{R(z)} = \frac{N(z)G_1(z)}{1 + G_1(z)H(z)N(z)} \tag{22}$$

The stability study of the system now involves the investigation of the following condition:

$$G_1(z)H(z) = -1/N(z) \tag{23}$$

For the relay-type nonlinearity with dead zone, all the critical regions shown in Figs. 5 to 9 are still valid for this system.

## Nonlinear Systems Without Zero-Order Hold -- Finite Pulse Width Considerations

The nonlinear system studied in the preceding sections is considered to have a zero-order hold following the sampler. If the zero-order hold is absent, and the sampling duration of the sampler can be assumed to be infinitesimal, the describing function technique presented here can still be applied. Instead of plotting $G_1(z)$, which includes the transfer function of the hold, it is only necessary to plot the locus of G(z) in the gain-phase plot and the same critical regions are used.

However, in some cases, with the absence of a hold device, a finite pulse width has to be considered. In this case, an approximation can be made if the narrow pulses with width p are approximated by flat-topped pulses. The sampler can then be represented by an ideal sampler followed by a fictitious hold device which holds the sampled signal for p second only, and then drops to zero instantaneously until the next impulse comes along. The transfer function of $G_1(s)$ is written

$$G_1(s) = \frac{1 - e^{ps}}{s} G(s) \qquad (24)$$

The z-transform of $G_1(s)$ is

$$G_1(z) = (1 - z^{-p/T}) \mathcal{Z}\left[\frac{G(s)}{s}\right] (p << T) \quad (25)$$

## Conclusion

A discrete describing function utilizing the z-transformation has been introduced for a relay-type nonlinearity for sampled-data systems. The critical regions which are defined by the describing function are used to study the condition of self-sustained oscillations by investigating the relative position of the $G_1(z)$ loci and the critical regions. The effect of varying the sampling time T on the system stability is shown clearly on the $G_1(z)$ loci. It is shown that the method can be applied to multisampler system as well as systems without hold devices. It might be pointed out that the discrete describing function method can also be applied to other types of amplitude dependent nonlinearities, such as saturation with dead zone.

## Appendix I

### Derivation of the Z-Transform-Describing Function N(z)

Case (I) Tc = 4T. $\Delta = 1$:

If the output of the zero-order hold is assumed to be of the form whown in Fig. 4 (b), the Laplace transform of the relay output v*(t) can be written as

$$V^*(s) = 1 - e^{-2Ts} + e^{-4Ts} + \dots$$

$$= \frac{1}{1 + e^{-2Ts}} \qquad (26)$$

The z-transform of which is

$$V(z) = \mathcal{Z}\left[V^*(s)\right] = \frac{z^2}{z^2 + 1} \qquad (27)$$

Thus

$$-\frac{1}{N(z)} = -\frac{E(z)}{V(z)}$$

$$= \frac{-Ez(z\cos\phi - \sin\phi)}{z^2 + 1} \times \frac{z^2 + 1}{z^2}$$

$$= -\frac{E(z\cos\phi - \sin\phi)}{z} \qquad (28)$$

When Tc = 4T, $z = \underline{1/90^\circ} = j$, Eq. (28) becomes

$$-\frac{1}{N(z)} = \frac{E(\sin\phi - j\cos\phi)}{J}$$

$$= E\underline{/-180^\circ + \phi} \qquad (29)$$

Referring to Fig. 4(b), the limitations on the magnitude of e(t) are:

$E\cos\phi > D$ and $E\,|\sin\phi| < D$ for

$-45^\circ < \phi < 45^\circ$

or

$E_{min} = D/\cos\phi$ and $E_{max} = D/|\sin\phi|$

Hence

$$-\frac{1}{N}\Big|_{max} = E_{max}\underline{/-180^\circ + \phi}$$

$$= \frac{D}{|\sin\phi|}\underline{/-180^\circ + \phi} \qquad (30)$$

and

$$-\frac{1}{N}\Big|_{min} = E_{min}\underline{/-180^\circ + \phi}$$

$$= \frac{D}{\cos\phi}\underline{/-180^\circ + \phi} \qquad (31)$$

Let

$E'_{max} = E_{max}/D$ and $E'_{min} = E_{min}/D$,

then

$$-\frac{1}{N}\Big|'_{max} = -\frac{1}{N}\Big|_{E = E'_{max}}$$

$$= \frac{1}{|\sin\phi|}\underline{/-180^\circ + \phi} \qquad (32)$$

$$-\frac{1}{N}\Big|'_{min} = -\frac{1}{N}\Big|_{E = E'_{min}}$$

$$= \frac{1}{\cos\phi}\underline{/-180^\circ + \phi} \qquad (33)$$

The critical region for the condition of self-sustained oscillation characterized by Tc = 4T and $\Delta = 1$ is the region enclosed by the loci of

$$-\frac{1}{N}\Big|'_{max} \quad \text{and} \quad -\frac{1}{N}\Big|'_{min}$$

given above, and is shown in Fig. 7.

Case (II) $Tc = 4T$, $\Delta = 2$:

It is shown in Fig. 4(c) that when a system has a sustained oscillation of period $T_c = 4T$, the output of the hold circuit can also have the form of $\Delta = 2$. The Laplace transform of v(t) corresponding to the waveform shown in Fig. 4(c) is

$$V^*(s) = 1 + e^{-sT} - e^{-2sT} + e^{-3sT} + e^{-4sT} + e^{-5sT} \ldots = \frac{1 + e^{-sT}}{1 + e^{-2sT}} \tag{34}$$

Thus,

$$V(z) = \frac{z^2 + z}{z^2 + 1} \tag{35}$$

$$\text{Thus } -\frac{1}{N(z)} = -\frac{E(z\cos\phi - \sin\phi)}{z + 1} \tag{36}$$

$$\text{and for } z = j,\ -\frac{1}{N(z)} = 0.707\ E\underline{/-135^\circ + \phi} \tag{37}$$

The range of $\phi$ as indicated in Fig. 4(c) is between 0° and -90°. The limitations on E are:

$$E_{max} = \infty \quad E_{min} = \frac{D}{|\sin\phi|} \quad \text{for } -45^\circ < \phi < 0^\circ$$

$$E_{max} = \infty \quad E_{min} = \frac{D}{\cos\phi} \quad \text{for } -90^\circ < \phi < -45^\circ$$

Thus

$$-\frac{1}{N}\Big|'_{max} = \infty \quad \text{for } 0^\circ > \phi > -90^\circ \tag{38}$$

$$-\frac{1}{N}\Big|'_{min} = \frac{1}{|\sin\phi|}\ \underline{/135^\circ + \phi}$$

$$\text{for } -45^\circ < \phi < 0^\circ \tag{39}$$

$$-\frac{1}{N}\Big|'_{min} = \frac{1}{\cos\phi}\ \underline{/-135 + \phi}$$

$$\text{for } -90^\circ < \phi < -45^\circ \tag{40}$$

The critical region of $Tc = 4T$ and $\Delta = 2$ is plotted in Fig. 7.

## Bibliography

1. C. K. Chow, "Contactor Servomechanisms Employing Sampled Data", Trans. AIEE, vol. 73, pp. 51-64; 1954.

2. F. A. Russell, "Design Criterion for Stability of Sampled-Data On-Off Servomechanisms" Ph.D. thesis, Columbia University, 1953.

3. B. C. Kuo, "A Z-Transformer-Describing Function for On-Off Type Sampled-Data Systems, Proc. I.R.E., Vol. 48, No. 5, pp. 941-942, May, 1960.

4. B. C. Kuo, Nonlinear Feedback Control Systems With Sampled-Data, Ph.D. Thesis University of Illinois, 1958.

5. R. E. Kalman, "Nonlinear Aspects of Sam Sampled-Data Control Systems", Proc. of the Symposium on Nonlinear Circuit Analysis, vol. VI; Polytechnic Institute of Brooklyn, New York, 1956.

Table I

-1/N(z) for various modes of oscillation
(relay with dead zone)

| $T_c$ | $\Delta$ | $-1/N(z)$ | $-1/N(z)\vert_{max}$ | $-1/N(z)\vert_{min}$ | Arg.(-1/N) | Range of $\phi$ |
|---|---|---|---|---|---|---|
| 2T | 1 | $E \cos \phi$ | $\infty$ | D | $-180^\circ$ | $-90^\circ < \phi < 90^\circ$ |
| 3T | 1. | 0.866E | $\frac{0.866D}{\cos(60^\circ - \vert\phi\vert)}$ | $\frac{0.866D}{\cos(60^\circ - \vert\phi\vert)}$ | $-150^\circ + \phi$ | $-30^\circ < \phi < 0^\circ$ |
| | | | $\frac{0.866D}{\cos \phi}$ | $\frac{0.866D}{\cos(120^\circ - \vert\phi\vert)}$ | $-150^\circ + \phi$ | $-60^\circ < \phi < -30^\circ$ |
| | $\Delta_{12}$ | 0.75E | $\infty$ | $\frac{0.75D}{\sin(30^\circ - \vert\phi\vert)}$ | $-180^\circ + \phi$ | $-30^\circ < \phi < 0^\circ$ |
| 4T | 1 | E | $\frac{D}{\cos(90^\circ - \vert\phi\vert)}$ | $\frac{D}{\cos \phi}$ | $-180^\circ + \phi$ | $-60^\circ < \phi < 0^\circ$ |
| | 2 | 0.707E | oo | $\frac{0.707D}{\cos(90^\circ - \vert\phi\vert)}$ | $-135^\circ + \phi$ | $-45^\circ < \phi < 0^\circ$ |
| | | | oo | $\frac{D}{\cos \phi}$ | $-135^\circ - \phi$ | $-90^\circ < \phi < -45^\circ$ |
| 5T | 1 | 1.31E | $\frac{1.31D}{\cos \phi}$ | $\frac{1.31D}{\cos(72^\circ - \vert\phi\vert)}$ | $-126^\circ + \phi$ | $-72^\circ < \phi < -36^\circ$ |
| | $\Delta_{12}$ | 0.954E | $\frac{0.954D}{\sin(18^\circ + \vert\phi\vert)}$ | $\frac{0.954D}{\sin(54^\circ - \vert\phi\vert)}$ | $-180^\circ + \phi$ | $-18^\circ < \phi < 0^\circ$ |
| | 2 | 0.81E | $\frac{0.81D}{\cos(36^\circ + \vert\phi\vert)}$ | $-\frac{0.81D}{\cos \phi}$ | $-126^\circ + \phi$ | $-72^\circ < \phi < -36^\circ$ |
| | $\Delta_{23}$ | 0.772E | oo | $\frac{0.772D}{\sin(18^\circ - \vert\phi\vert)}$ | $-180^\circ + \phi$ | $-18^\circ < \phi < 0^\circ$ |
| 6T | 1 | 1.5E | $\frac{1.5D}{\cos(60^\circ - \vert\phi\vert)}$ | $\frac{1.5D}{\cos \phi}$ | $-180^\circ + \phi$ | $-30^\circ < \phi < 30^\circ$ |
| | 2 | 0.866E | $\frac{0.866D}{\cos \phi}$ | $\frac{0.866D}{\cos(120^\circ + \phi)}$ | $-90^\circ + \phi$ | $-90^\circ < \phi < -60^\circ$ |
| | | | $\frac{0.866D}{\cos \phi}$ | $\frac{0.866D}{\cos(60^\circ + \phi)}$ | $-90^\circ + \phi$ | $-120^\circ < \phi < -90^\circ$ |
| | 3 | 0.75E | oo | $\frac{0.75D}{\cos \phi}$ | $-120^\circ + \phi$ | $-90^\circ < \phi < -60^\circ$ |

Table II

| Mode of Oscillation | Tc(sec) | $\Delta$ | Amplitude of Oscillation | |
|---|---|---|---|---|
| | | | Describing Function | Difference Equation |
| 1 | 4 | 1 | 0.31 | 0.30 |
| 2 | 4 | 2 | 0.438 | 0.352 |
| 3 | 5 | 2 | 0.572 | 0.525 |
| 4 | 6 | 3 | 0.838 | 0.80 |

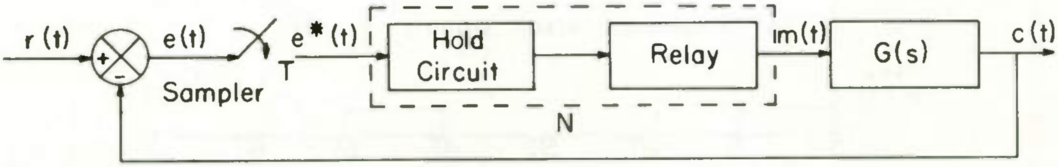


Fig. 1. Relay type sampled-data control system.

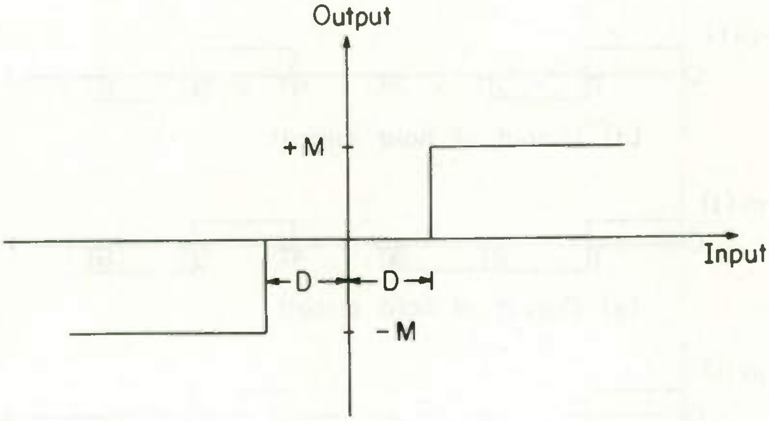


Fig. 2. Relay characteristics.

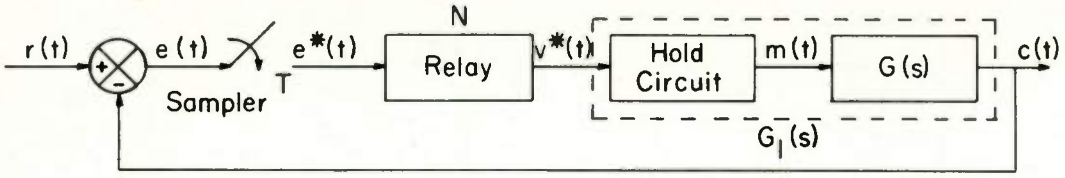
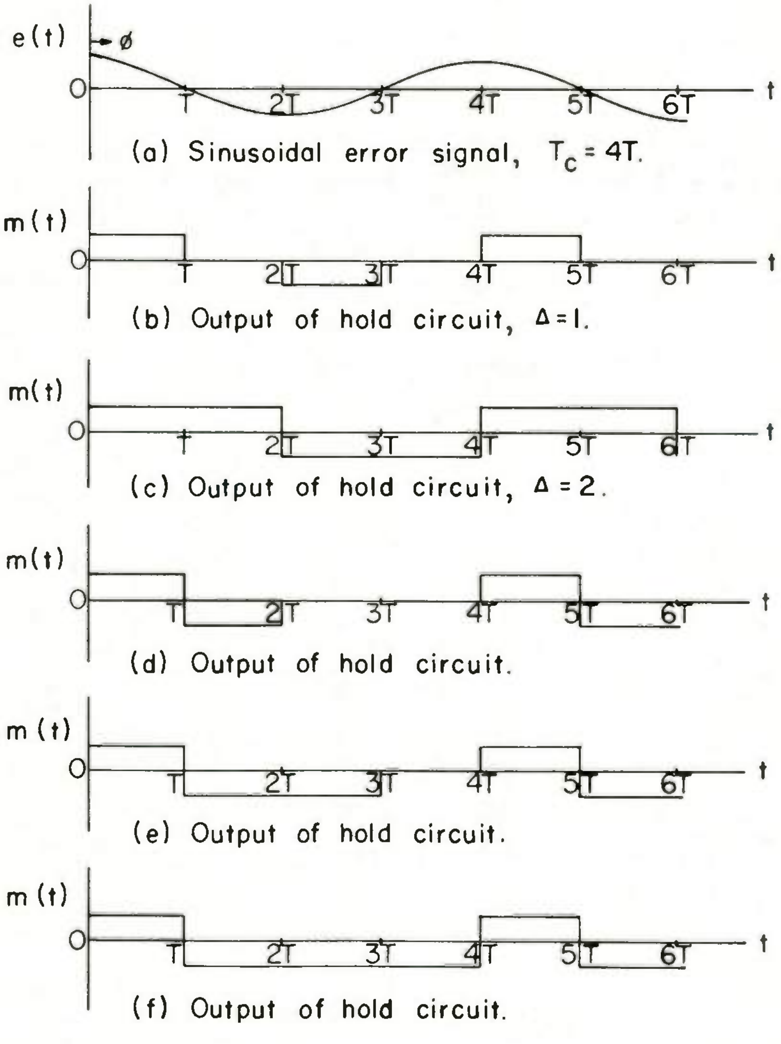


Fig. 3. Modified system.



(a) Sinusoidal error signal, $T_c = 4T$.

(b) Output of hold circuit, $\Delta = 1$.

(c) Output of hold circuit, $\Delta = 2$.

(d) Output of hold circuit.

(e) Output of hold circuit.

(f) Output of hold circuit.

Fig. 4. Possible configurations of m(t) for $T_c = 4T$.

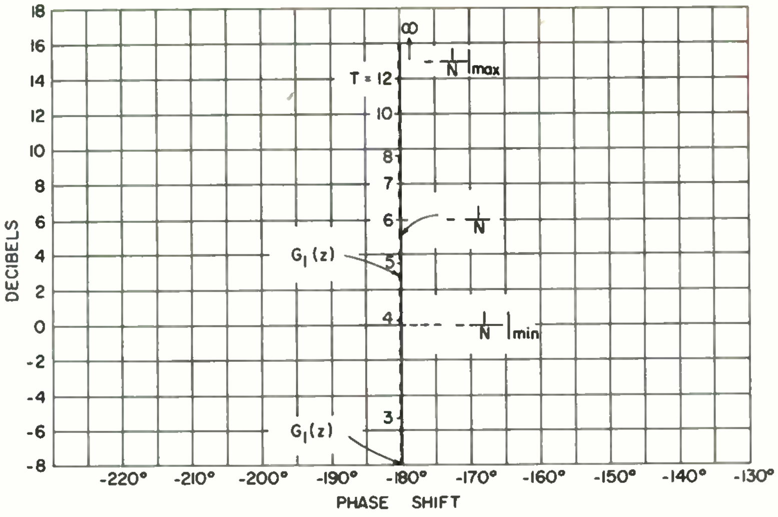


Fig. 5. Critical region for relay-type nonlinearity, $T_c = 2T$, and $G_1(z)$ plot, $G(s) = \frac{1}{s(s+1)}$.

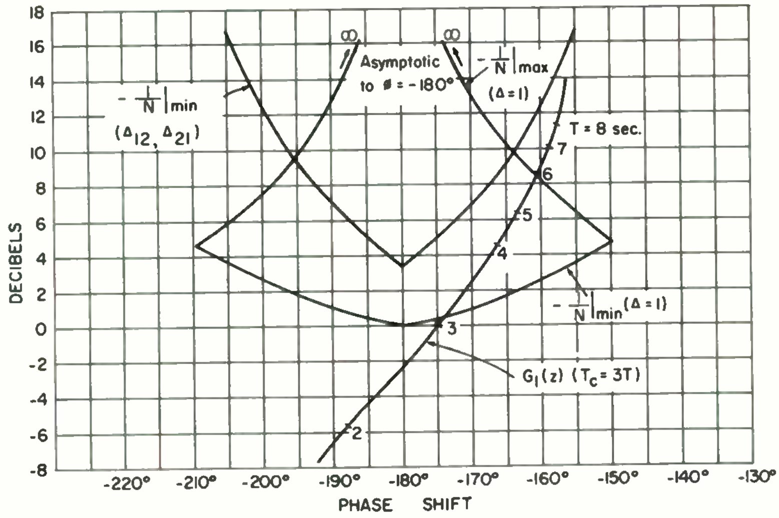


Fig. 6. Critical region for relay-type nonlinearity, $T_c = 3_T$, and $G_1(z)$ plot, $G(s) = \frac{1}{s(s+1)}$.

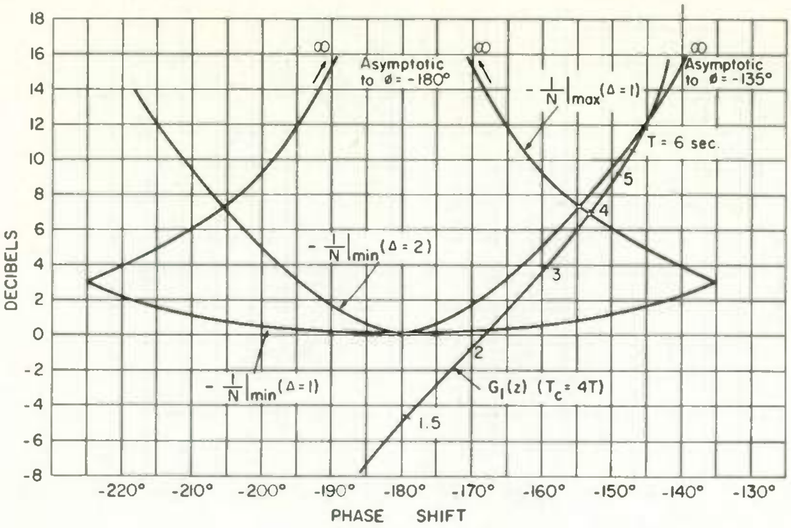


Fig. 7. Critical regions for relay-type nonlinearity, $T_c$ = 4T, and $G_1$(z) plot, $G(s) = \frac{1}{s(s+1)}$.

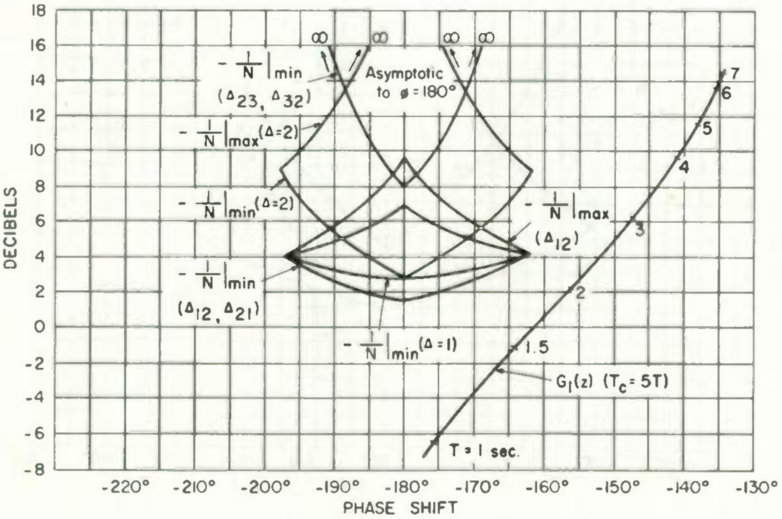


Fig. 8. Critical regions for relay-type nonlinearity, $T_c$ = 5T, and $G_1$(z) plot, $G(s) = \frac{1}{s(s+1)}$.

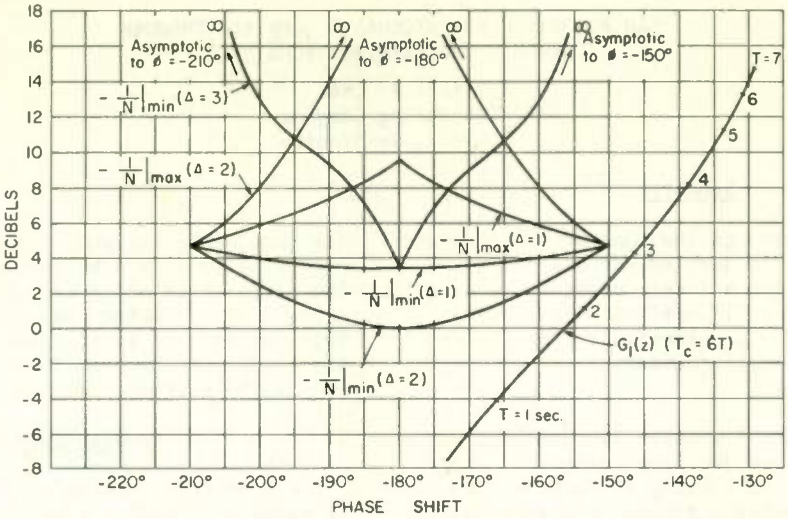


Fig. 9. Critical regions for relay-type nonlinearity, $T_c = 6_T$, and $G_1(z)$ plot, $G(s) = \dfrac{1}{s(s+1)}$.

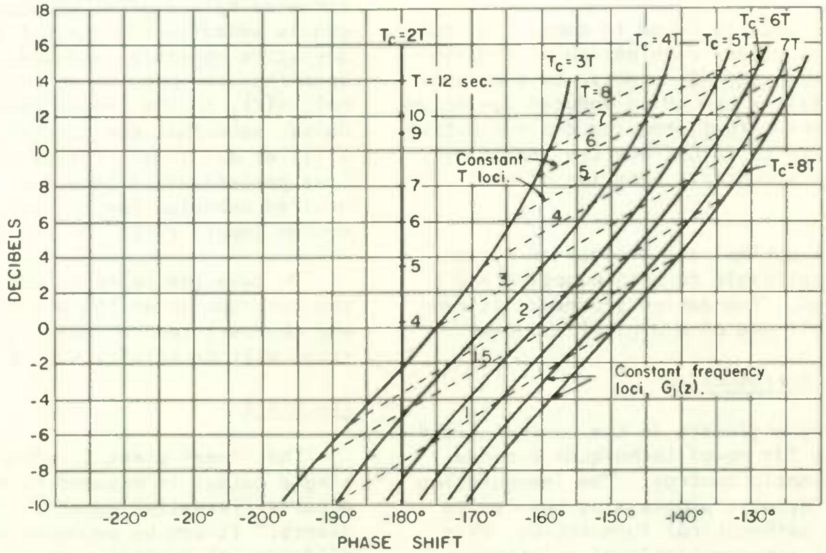


Fig. 10. Frequency loci of $G_1(z)$ in gain-phase plane, $G(s) = \dfrac{1}{s(s+1)}$.

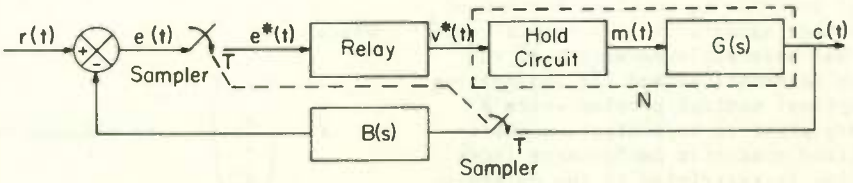


Fig. 11. Relay-type sampled-data control system with two synchronized sampling switches.

# AN ALGORITHM FOR STOCHASTIC CONTROL THROUGH DYNAMIC PROGRAMMING TECHNIQUES

Paul P. Chen
The Boeing Company
Renton, Washington

## Summary

An algorithm based on the concept of state and dynamic programming is derived for designing an optimum controller for a linear plant subject to noise. The controller is optimal in the sense that the behavior of the plant satisfies the expected mean quadratic performance index (EMQPI) defined in the paper.

The optimal control problem is formulated as a problem of multi-stage decision processes in dynamic programming. By solving a functional equation obtained by applying Bellman's principle of optimality to the control process in question, the algorithm is formulated. The algorithm generates the sequence of control signals which minimize the EMQPI. In addition, it gives the minimum of the EMQPI for the specified sequence of control signals.

The control signal is found to consist of two components: (1) a linear combination of the system state variables, and (2) a noise-balance component which minimizes the noise-induced deviation of the actual plant output from the desired output. An example is given to illustrate the iterative procedure and the asymptotic behavior of the algorithm.

The design is optimal for a class of system inputs, and is applicable to both sampling and continuous systems. The design procedure is developed to make full use of a digital computer.

## Introduction

Scientists and engineers in the control field search constantly for novel techniques and new theories for automatic control. The inauguration of the theory of dynamic programming has led to new ways for the mathematical formulation, analytical treatment and computational solution of control problems. The modernization of digital computers has facilitated and accelerated the research in this direction.

Recently a number of papers applying the techniques of dynamic programming to the treatment of control problems have appeared in the literature![1-6] Kalman and Koepcke[1] have achieved essential results in both the mathematical and the engineering aspects of the optimal control problem where a linear, stationary plant to be controlled satisfies the generalized quadratic performance index. Their investigation is restricted to the deterministic case. However, in reality most physical systems are subject to random disturbances. This paper describes an investigation of the optimal control problem similar to that originally discussed by Kalman and Koepcke, but in the stochastic case, that is, when the plant is subject to random noise with a known probability distribution. The performance criterion used for the stochastic case is the expected mean quadratic performance index as defined in the next section of this paper. The algorithm obtained due to dynamic programming is ideally suited to digital computation.

## Problem Formulation

A schematic diagram of the system considered is shown in Figure 1. It is assumed that the linear plant is preceeded by a sample-and-hold element so that the input to the plant will be a piecewise constant function of time. In this way, the problem can be formulated as a descrete time model which is not only a more realistic description of many physical systems, but also readily accepted by a digital computer. The noise input can be added to the control signal or applied at any state variable location of the plant. Roughly speaking, the problem is to find the control signal, m(t), to the linear plant which is subject to noise, such that the output of the plant, c(t), will, at all times, follow as closely as possible some predetermined behavior. For example, the desired behavior for the plant output may be the system input, r(t).

To make the paper reasonably self-contained, the description of the plant, the system input, and the performance index in the deterministic case[1] will be briefly stated below.

### The Plant

The linear plant, X, with a single input and a single output is assumed to be described by an $n^{th}$ order differential equation with constant coefficients. It can be decomposed into n first order differential equations by selecting n state variables, $x_1, x_2, \dots x_n$. The vector-matrix differential equation describing the plant can be written as:

$$\dot{X} = AX + hm \qquad (1)$$

where:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \dot{x}_n \end{pmatrix}, \text{ an n-dimensional vector}$$

$A$ = n x n constant matrix
$m$ = control signal
$h$ = n-dimensional constant vector represent the effect of m on $\dot{X}$

$\dot{X}$ = the derivative of X with respect to time

The variable $x_1$ can always be selected as the plant output, c, that is:

$$x_1(t) = c(t) \quad . \qquad (2)$$

The solution of (1) can be written as[7]:

$$X(t) = \Phi^x(t)\; X(0) + \int_0^t \Phi(t-\tau) h\; m(\tau)\, d\tau \qquad (3)$$

where $\Phi^x(t)$ is the solution of the matrix differential equation:

$$\dot{\Phi}^x = A\Phi^x \text{ with } \Phi^x(0) = I \text{ (unit matrix)} \qquad (4)$$

and is called the transition matrix of the plant. Since the control signal, m, is piecewise constant, it can be shown that[8,9]:

$$X(kT+\tau) = \Phi^x(\tau)\; X(kT) + m(kT) H^x(\tau) \qquad (5)$$

where T is the sampling period, $0 \leq \tau \leq T$, k = 0, 1, 2, ..., and:

$$H^x(\tau) = \int_0^\tau \Phi^x(\tau) h d\tau \quad . \qquad (6)$$

## The Input

Let $\{r(t)\}$ be a class of system inputs for which the system is designed to be optimized. An element, r(t), of this class is defined as:

$$\begin{aligned} r(t) &= y_1(t) \quad \text{for} \quad t \geq 0 \\ &= 0 \qquad\qquad\quad\; < 0 \end{aligned} \qquad (7)$$

where $y_1(t)$ is the first component of the input state vector:

$$Y(t) = \Phi^y(t) Y(0) \quad . \qquad (8)$$

The matrix, $\Phi^y(t)$, is the transition matrix of an $\ell^{th}$ order ordinary linear differential equation with constant coefficients, and Y(0) is an arbitrary constant vector. Thus, $\Phi^y(t)$ and Y(0) determine, respectively, the specific class of system inputs and the particular member of that class. For example, the class of all step and ramp functions $\{r(t)\}$ can be described by the differential equation:

$$\frac{d^2 r(t)}{dt^2} = 0 \quad .$$

Select the input state variable as $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} r \\ \dot{r} \end{pmatrix}$ the vector differential equation describing the input can be written as:

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

and the transition matrix as:

$$\Phi^y(\tau) = e^{\begin{pmatrix} 0 1 \\ 0 0 \end{pmatrix}\tau} = \begin{pmatrix} 1 \tau \\ 0 1 \end{pmatrix} \quad .$$

Thus, the input, r(t), is described by the transition equation as given by equation (8):

$$\begin{pmatrix} r(\tau) \\ \dot{r}(\tau) \end{pmatrix} = \begin{pmatrix} 1 & \tau \\ 0 & 1 \end{pmatrix} \begin{pmatrix} r(0) \\ \dot{r}(0) \end{pmatrix}$$

where r(t) is satisfied by r(t) = at + b for all constants a and b through arbitrary selection of r(0) and $\dot{r}(0)$.

If the actual system input does not belong to the class of signals for which the system is optimized, then the input may be approximated over one sampling period by a member of that class so that the system can be nearly optimized. This approximation has been discussed by Lees[10], and Kalman and Koepcke.[1]

## The Performance Index

The input state vector, Y(t), and the plant state vector, X(t), can be combined to form the system (input-plant) state vector, Z(t), as:

$$Z(t) = \begin{pmatrix} X(t) \\ Y(t) \end{pmatrix} \quad . \qquad (9)$$

By defining

$$\Phi(t) = \left(\begin{array}{c|c} \Phi^x(t) & 0 \\ \hline 0 & \Phi^y(t) \end{array}\right) \qquad (10)$$

$$H(t) = \left(\begin{array}{c} H^x(t) \\ \hline H^y(t) \end{array}\right) \quad , \qquad (11)$$

the system transition equation can be written as:

$$Z(kT+\tau) = \Phi(\tau) Z(kT) + m(kT) H(\tau) \quad . \qquad (12)$$

With the above definition, the quadratic performance index over the interval $0 \leq t \leq NT$ can be defined as:

$$J_N = \int_0^{NT} \left[ Z_t(t) Q Z(t) + \lambda m^2(t) \right] \omega(t)\, dt \qquad (13)$$

where Q is a positive-definite $(n+\ell)$ square matrix characterizing the kinds of errors specified for performance measurement, λ is a positive constant which indicates the weight of control cost with respect to the minimizing errors, and ω(t) is a weighting function of time. The symbol, $Z_t$, denotes the transpose of the vector, Z. If errors are considered only at sampling instants, equation (13) will be reduced to:

$$J_N^* = \sum_{k=1}^{N} \left[ Z_t(kT) Q Z(kT) + \lambda T\; m^2(k-1) T \right] \omega(k) \quad . \qquad (14)$$

The term, $Z_t(0)QZ(0)$, is omitted here in view of its fixed value from the given initial conditions of the input and the plant.

The following example is given to illustrate the nature of Q. In a unit feedback system a plant, whose Laplace transfer function is $\frac{1}{S(S+1)}$, is subject to a ramp input, $r(t) = t$, whose transfer function is $R(S) = \frac{1}{S^2}$. The system can be decomposed as shown in Figure 2. If the system error squared, $e^2 = (r-c)^2$, and the error rate squared, $\dot{e}^2 = (\dot{r}-\dot{c})^2$, are considered as error terms,

$$Q = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}$$

and $e^2+\dot{e}^2 = Z_tQZ$. Observe that Q is generally symmetric.

## Noise Consideration

In general there are three ways by which random variables are introduced into control systems. (1) The control signal, m, can be contaminated by some random noise, f, in which case the plant input is some function of the control signal and noise. Usually the plant input is the control signal plus noise m+f. (2) Noise disturbances due to system environment occur at places other than the plant input location. This may be the case where a plant consists of several component parts and noise occurs at the connection locations of the component parts. In most cases the noise location can be selected as a state variable. (3) Some of the elements of a control system contain parameters which exhibit randomness, that is, there are noisy components in a control system. Only the first two cases are discussed in this paper.

The vector differential equation describing the plant, when noise is considered, can be written as:

$$\dot{X} = AX + hm + gf \tag{15}$$

where g = n-dimensional constant vector represent the effect of noise on $\dot{X}$.

Let the probability distribution of the noise be P(f). Assuming noise occurs only at sampling instants, that is, impulsive noise only, the expected state of the plant denoted by $E[X]$ can be expressed by the probabilistic transition equation:

$$E\left[X(kT+\tau)\right] = \Phi^x(\tau)X(kT) + H^x(\tau)m(kT) + \int G^x(\tau)f(kT)P(f)df \tag{16}$$

where $G^x(\tau)$ is the response of the plant $\tau$ seconds after a unit impulse is applied at the noise location, when the plant is initially in state $X = 0$. When the plant input through the hold is the control signal plus noise, then $G^x = H^x$. The probabilistic system transition equation becomes:

$$E\left[Z(kT+\tau)\right] = \Phi(\tau)Z(kT) + H(\tau)m(kT) \int G(\tau)f(kT)P(f)df \tag{17}$$

where:

$$G(\tau) = \begin{pmatrix} G^x(\tau) \\ \hline 0 \end{pmatrix} . \tag{18}$$

Let $\bar{J}_N$ and $\bar{J}^*_N$ be defined as:

$$\bar{J}_N = \frac{J_N}{NT} , \quad \bar{J}^*_N = \frac{J^*_N}{N} \tag{19}$$

and be called, respectively, the mean quadratic performance index (MQPI) and the sampled mean quadratic performance index (MQPI*). The expected values of MQPI and MQPI* for stochastic processes are given by:

$$E\left[\bar{J}_N\right] = \frac{1}{NT}\int_0^{NT}\left\{\int\left[Z_t(t)QZ(t)\right]P(f)df + \lambda m^2(t)\right\}\omega(t)dt, \tag{20}$$

and

$$E\left[\bar{J}^*_N\right] = \frac{1}{N}\sum_{k=1}^{N}\left\{\int\left[Z_t(kT)QZ(kT)\right]P(f)df + \lambda T\ m^2(k-1)T\right\}\omega(k) \tag{21}$$

and are represented by EMQPI and EMQPI*, respectively.

## Statement of the Problem

The problem discussed in this paper can now be precisely stated below:

Given:
1. A linear plant X described by equation (15)
2. A random noise f of known probability distribution P(f)
3. A class of inputs Y described by equation (8).

Find a sequence of control signals m(o), m(1), m(2), ... m(N-1), such that the EMQPI is minimized for any initial conditions of the plant, X(0), and input, Y(0), as N approaches infinity. Let $\bar{I}_N$ denote the minimum of the EMQPI for this optimum sequence of N control signals, that is:

$$\bar{I}_N = \min_{m(o),m(1),\dots m(N-1)} E\left[\bar{J}_N\right] . \tag{22}$$

As N approaches infinity, $I_N$ is denoted by I, that is,

$$\bar{I} = \lim_{N\to\infty} \bar{I}_N . \tag{23}$$

The problem, then, is to specify m(o), m(1),... to give $\bar{I}$ for given X, f, Y and Z(0).

### Principle of Optimality

The basic principle by which the optimum control signal is obtained was developed by Bellman[11,12] This principle, the "principle of optimality", states: "an optimal policy has the property that, whatever the initial state and the initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision." The problem of deciding m(0), m(1), ..., m(N-1) for N sampling periods can be viewed as an N-stage multidecision process in dynamic programming. When applied to the problem of this paper, the principle of optimality can be stated as: an optimal sequence of control signals, m(0), m(1), ... m(N-1), has the property that, whatever the initial state, Z(0), and the initial control signal, m(0), the remaining sequence, m(1), m(2),... m(N-1), must constitute an optimal sequence with regard to the state, Z(1), resulting from the first choice m(0). Note that $I_N$ is a function of Z(0) and N. The effect of any initial choice of m(0) for the time interval 0 to 1 will be to convert Z(0) into a new state Z(1). It follows then that at time, t = 1, the problem of determining m(1) will be the same as that of determining m(0) at t = 0, except that the initial state, Z(1), will be used instead of Z(0), and the remaining length of time, N-1, will be used instead of N. This argument is the application of the principle of optimality and makes it possible to write the principle mathematically as:

$$\bar{I}_N(Z(0)) = \frac{1}{NT} \min_{m(0)} \Big\{ E\big[J_1(Z(0))\big] + E\big[\bar{I}_{N-1}(Z(1))\big] \Big\}. \tag{24}$$

Starting with N = 1, equation (24) gives:

$$\bar{I}_1(Z(0)) = \frac{1}{T} \min_{m(0)} E\big[J_1(Z(0);m(0))\big] . \tag{25}$$

Equation (24) is the functional equation produced by a mathematical transiteration of this principle.

### The Functional Equation

In this section, the formulation of functional equations by the mathematical transiteration of the principle of optimality will be demonstrated for stochastic control processes. An analytic solution of the functional equations will be given in the next section. The expected system state at the (k+1)th sampling instant for T = 1 can be expressed by (17) as:

$$E\big[Z(k+1)\big] = \Phi(1)Z(k) + H(1)m(k) + \int G(1)fP(f)df . \tag{26}$$

The first two terms on the right side of equation (26) represent the system state at time k+1 for a noise-free condition, and will be denoted by $Z^d(k+1)$. Let the expected (mean) value and the second moment of the noise be denoted respectively by $\bar{f}$ and $\bar{\bar{f}}$, that is,

$$\bar{f} = \int fP(f)df \tag{27}$$

$$\bar{\bar{f}} = \int f^2P(f)df ; \tag{28}$$

equation (26) is reduced to

$$E\big[Z(k+1)\big] = Z^d(k+1) + G(1)\bar{f} . \tag{29}$$

If system errors are interesting only at the sampling instants and are weighted equally at all sampling instants, that is, $\omega(k) = 1$ for k = 1, 2, ...,N, the EMQPI* by virtue of equations (21) and (27) to (29) becomes:

$$E\big[\bar{J}_N^*(Z(0))\big] = \frac{1}{N} \sum_{k=1}^{N} \Big\{ Z_t^d(k)QZ^d(k) + 2\bar{f}G_tQZ^d(k) + \bar{\bar{f}}G_tQG + \lambda m^2(k-1) \Big\}. \tag{30}$$

The principle of optimality yields the functional equation,

$$\bar{I}_N^*(Z(0)) = \frac{1}{N} \min_{m(0)} \Big\{ E\big[J_1^*(Z(0))\big] + E\big[\bar{I}_{N-1}^*(Z(1))\big] \Big\} \tag{31}$$

where:

$$E\big[J_1^*(Z(0))\big] = Z_t^d(1)QZ^d(1) + 2\bar{f}G_tQZ^d(1) + \bar{\bar{f}}G_tQG + \lambda m^2(0). \tag{32}$$

Note that Z(1) is the state resulting from a transition depending upon noise. The last term in equation (31), which represents the expected minimum of the expected quadratic performance index from the random initial state, Z(1), for an N-1 stage process, has the meaning as below:

$$E\big[I_{N-1}^*(Z(1))\big] = \int \Big[I_{N-1}^*(Z^d(1) + Gf)\Big]P(f)df . \tag{33}$$

Equations (31) to (33) are the fundamental equations from which the following analytic results are obtained.

### Solution of the Functional Equation

Analytic results are obtained by proceeding inductively with equation (31). For a single-stage process, N = 1, thus $I_{N-1}^* = 0$, and equation (31) is reduced to:

$$\bar{I}_1^*(Z(0)) = \min_{m(0)} E\big[J_1^*(Z(0))\big] . \tag{34}$$

Observe that $E\big[J_1^*(Z(0))\big]$ is a function of m(0) by virtue of equation (32); its minimum can be obtained by making its derivative with respect to m(0) equal to zero, that is,

$$\frac{dE\left[J_1^*(Z(0))\right]}{dm(0)} = H_tQ\left[\Phi Z(0) + Hm(0)\right] + \bar{f}G_tQH$$

$$+ \lambda m(0) = 0 \ . \tag{35}$$

The optimal control signal starting from the initial state for a single-stage process is the solution of m(0) in equation (35) and will be denoted by $m_1(0)$, thus,

$$m_1(0) = \alpha_t(1)Z(0) + \bar{f}\beta(1) \tag{36}$$

where:

$$\alpha(1) = \frac{a(1)}{\lambda+d(1)}$$

$$\beta(1) = \frac{b(1)}{\lambda+d(1)}$$

$$a(1) = -\Phi_tQH$$

$$b(1) = -H_tQG$$

$$d(1) = H_tQH \ .$$

Since the performance criterion is quadratic, the extremum will necessarily be a minimum.[13] Substituting (36) into (34) and combining terms which are quadratic in Z(0), linear in Z(0), and constants, it follows that the minimum of the EMQPI* from the initial state over one sampling period is:

$$\bar{I}_1^*(Z(0)) = Z_t(0)M(1)Z(0) + 2\bar{f}K_t(1)Z(0)$$

$$+ \bar{f}^2U(1) + \bar{\bar{f}}V(1) \tag{37}$$

where:

$$M(1) = \psi_t(1)Q\psi(1) + \lambda\alpha(1)\alpha_t(1)$$

$$K_t(1) = \xi_t(1)Q\psi(1) + \lambda\beta(1)\alpha_t(1)$$

$$U(1) = \beta^2(1)(H_tQH + \lambda) + 2\beta(1)G_tQH$$

$$V(1) = G_tQG$$

$$\psi(1) = \Phi + H\alpha_t(1)$$

$$\xi(1) = \beta(1)H + G \ .$$

The matrix, M(1), is a symmetric, positive, definite matrix.[14]

Theorem

If a system is subject to random noise, f, with probability distribution, P(f), mean, $\bar{f}$, and second moment, $\bar{f}$, and is described in discrete version by the state-transition equation, $Z(k+1) = \Phi Z(k) + Hm(k) + Gf(k)$, at the initial state, Z(0), the first control signal which minimizes the EMQPI* over the interval $0 \leq t \leq N$ is:

$$m_N(0) = \alpha_t(N)Z(0) + \bar{f}\beta(N) \tag{38}$$

where:

$$\alpha(N) = \frac{a(N)}{\lambda+d(N)} \quad \text{(an } \ell\text{+n vector)}$$

$$\beta(N) = \frac{b(N)}{\lambda+d(N)} \quad \text{(a constant)}$$

$$a(N+1) = -\Phi_t\left[M(N) + Q\right]H \quad \text{(an } \ell\text{+n vector)}$$

$$b(N+1) = -H_t\left[M(N) + Q\right]G - K_t(N)H \quad \text{(a constant)}$$

$$d(N+1) = H_t\left[M(N) + Q\right]H \quad \text{(a constant)}.$$

The minimum of the EMQPI* when the performance measure is weighted equally at all sampling instants, is:

$$\bar{I}_N^*(Z(0)) = \frac{1}{N}\Big[Z_t(0)M(N)Z(0) + 2\bar{f}K_t(N)Z(0)$$

$$+ \bar{f}^2U(N) + \bar{\bar{f}}V(N)\Big] \tag{39}$$

where:

$$M(N+1) = \psi_t(N+1)\left[M(N) + Q\right]\psi(N+1)$$

$$+ \lambda\alpha(N+1)\alpha_t(N+1) \quad \text{(an } \ell\text{+n square matrix)}$$

$$K_t(N+1) = \xi_t(N+1)\left[M(N) + Q\right]\psi(N+1)$$

$$+ \lambda\beta(N+1)\alpha_t(N+1) + K_t(N)\psi(N+1)$$

$$\text{(an } \ell\text{+n vector)}$$

$$U(N+1) = \beta^2(N+1)\left[H_t\left[M(N) + Q\right]H + \lambda\right]$$

$$+ 2\beta(N+1)\left[G_t\left[M(N) + Q\right]H + K_t(N)H\right]$$

$$+ U(N) + 2K_t(N)G \quad \text{(a constant)}$$

$$V(N+1) = G_t\left[M(N) + Q\right]G + V(N) \quad \text{(a constant)}$$

$$\psi(N+1) = \Phi + H\alpha_t(N+1)$$

$$\xi(N+1) = \beta(N+1)H + G$$

for N = 0, 1, 2, ..., $\lambda > 0$, with $M(0) = K_t(0) = U(0) = V(0) = 0$.

Equations (38) and (39) give the algorithm from which the optimal control signal and the minimum of the EMQPI* are generated. As can be seen from (38) the optimal control signal consists of two components: (1) a linear combination of the system state variables, which is a component to guide the plant to follow its desired output, and (2) a component proportional to the mean value of the noise to minimize the noise-induced deviation of the actual plant output from the desired output. The EMQPI*, when the optimal sequence of control signals is used, results in (39) which consists of quadratic and linear terms in Z(0), and constant terms proportional to the squared mean value of the noise and to the second moment of the noise.

Proof: The theorem is verified by mathematical induction. The theorem is true for N = 1, as it can be seen that equations (36) and (37) result from direct substitution of 1 for N in equations

(38) and (39). Let the induction hypothesis be equations (38) and (39). The principle of optimality and equations (39) and (33) yield:

$$\bar{I}^*_{N+1}(Z(0)) = \frac{1}{N+1} \underset{m(0)}{\text{Min}} \Big\{ E\big[J^*_1(Z(0))\big] + E\big[I^*_N(Z(1))\big]\Big\} \tag{40}$$

$$= \frac{1}{N+1} \underset{m(0)}{\text{Min}} \Big\{\Big[\int Z_t(1)QZ(1)P(f)\,df + \lambda m^2(0)\Big] + \int\Big[Z_t(1)M(N)Z(1) + 2\bar{f}K_t(N)Z(1) + \bar{f}^2U(N) + \bar{\bar{f}}V(N)\Big]P(f)\,df\Big\} \tag{41}$$

By virtue of equations (27) to (29), equation (41) becomes:

$$= \frac{1}{N+1} \underset{m(0)}{\text{Min}} \Big\{ Z^d_t(1)\big[M(N)+Q\big]Z^d(1) + 2\bar{f}\big\{G_t\big[M(N)+Q\big] + K_t(N)\big\}Z^d(1) + \bar{\bar{f}}\big\{G_t\big[M(N)+Q\big]G + V(N)\big\} + \bar{f}^2\big[U(N) + 2K_t(N)G\big] + \lambda m^2(0)\Big\}. \tag{42}$$

Notice that $Z^d(1)$ is a function of m(0). It is a straight forward process to obtain the minimum of the expression, $\{\ \}$ in equation (42). Make $\frac{d\{\ \}}{dm(0)} = 0$ and solve for m(0). The solution of m(0), thus obtained, will be $m_{N+1}(0)$, and equals equation (38) if N is replaced by N+1. Substitute $m_{N+1}(0)$ into equation (42); it is easy to show that $\bar{I}^*_{N+1}(Z(0))$ is the same as equation (39) if N is replaced by N+1. Thus, the hypothesis is true for N = N+1, and the theorem is true for all positive integers N.

The theorem gives the first optimal control signal. For successive optimal control signals, the principle of optimality reveals that the optimal control signal at the (k+1)th stage for an N-stage control process, $m_N(k)$, is the same as the optimal control signal at the first stage for an N-k stage process, $m_{N-k}(0)$. Thus,

Corollary 1:

$$m_N(k) = m_{N-k}(0) \quad \text{for } k = 0, 1, 2, \ldots N-1$$

where $m_N(k)$ and $m_{N-k}(0)$ are linear functions of Z(k) and Z(0) respectively.

For a deterministic case, $f = \bar{f} = \bar{\bar{f}} = 0$, equations (38) and (39) reduce to

Corollary 2:

$$m_N(0) = \alpha_t(N)Z(0) \tag{43}$$

$$I^*_N(Z(0)) = Z_t(0)M(N)Z(0)\ . \tag{44}$$

This is essentially the same result as derived by Kalman and Koepcke for the deterministic case.[1]

The theorem formulates the algorithm to generate the necessary constants of the optimal controller described in the next section. Starting with $M(0) = K_t(0) = U(0) = V(0) = 0$, for N = 0, a(1), b(1), d(1) are obtained from (38), then M(1), $K_t(1)$, U(1) and V(1) are obtained from (39). The terms a(2), b(2), d(2) result by substituting M(1) and $K_t(1)$ into (38) for N = 1, and the cycle repeats as desired for any N. Thus, the optimal control signal at the instant when there are N sampling periods left until the end of the process, $m_N(0)$, for any N is readily calculated by (38) for the given knowledge Z(0) which can usually be measured or predicted for physical systems.

## An Alternative Derivation of the Optimum Control Signal

If (1), the expected state instead of the expected error is used to formulate the EMQPI*, that is,

$$E\big[J^*_N\big] = \sum_{k=1}^{N}\Big\{E\big[Z_t(k)\big]QE\big[Z(k)\big] + \lambda m^2(k-1)\Big\} \tag{45}$$

and (2), the expected state is used to formulate the minimum of the expected quadratic performance index, instead of the random state used to formulate the expected minimum of the expected quadratic performance index, that is, the last term in equation (31) be replaced by,

$$I^*_{N-1}(E\big[Z(1)\big])\ , \tag{46}$$

then the functional equation can be formulated with equations (45) and (46) by the principle of optimality and proceed inductively as before. The derivation of the optimal control signal is a straight forward process; it results the same as equation (38). The minimum of the EMQPI* thus obtained is the same as equation (39) except that:

$$U(N+1) = \xi_t(N+1)\big[M(N)+Q\big]\xi(N+1) + 2K_t(N)\xi(N+1) + \lambda\beta^2(N+1) + U(N)$$

$$V(N) = 0\ .$$

## Optimum Synthesis

It can be shown that the iterative process of the algorithm converges.[15] The terms, $\alpha(N)$ and $\beta(N)$, converge to a constant vector, $\alpha$, and a constant, $\beta$, as N approaches infinity. If $\alpha(N)$, $\beta(N)$ converge to $\alpha$, $\beta$ at the $N_T$th iteration, then, for $N \geq N_T$

$$\left.\begin{aligned}\alpha(N) &= \alpha\\ \beta(N) &= \beta\end{aligned}\right\}. \tag{47}$$

From equation (38), it is clear that for all $N \geq N_T$, the optimal control signal at the initial stage for an N-stage process is the same as that for an $N_T$-stage process. That is,

$$m_N(0) = m_{N_T}(0) \ .$$

If a system operates for a reasonable length of time such that N can be considered as infinity, the optimal control signal at any sampling instant, k, can be represented by:

$$m(k) = \alpha_t Z(k) + \bar{f}\beta \tag{48}$$

where $\alpha_t Z(k)$ is a linear combination of the state variables and the term $\bar{f}\beta$ is a constant, called the noise-balance component, W. The additional error in J introduced by using equation (48) instead of equation (38) for m, is negligibly small when N is reasonably larger than $N_T$.

The optimal controller can be designed according to the schematic diagram shown in Figure 3. The feedback coefficients $\alpha_1, \alpha_2, \ldots \alpha_{n+\ell}$, are components of the vector, $\alpha$, defined in equation (47). The additional constant signal, W, to the controller is the noise-balance component defined in equation (48). When the sampling period is reduced to be reasonably small, the system becomes nearly continuous, and the controller thus resulted will minimize the EMQPI* as well as the EMQPI.

## Example

To illustrate the foregoing development of the algorithm, the noise-balance component, the feedback coefficients, and the minimum of the EMQPI* are calculated for the system in Figure 4 by using an IBM 709 digital computer. The noise considered is unit impulse noise (f=1) with Bernoulli distribution, that is, P(f=1) = p and P(f=0) = 1-p where p = 1/4. The noise is added to the control signal (type 1 noise) and to the integrator (type 2 noise) at sampling instants only. The class of inputs for which the system is to be optimized is the class of all ramps and steps. There are two performance indexes. The first is the sum of the squared error and squared error rate, $e^2 + \dot{e}^2$; the second is the squared error, $e^2$, only. The control cost is weighted equally for both indexes, that is $\lambda = 1$ for both. According to the above specification the problem is calculated in three cases defined in Table 1. For example case 3 is to design the controller for the system which is subject to the impulse noise at the integrator so that the sum of the squared error and the squared control signal is minimized for the class of all ramp and step inputs. Let:

$$\Phi_1 = \begin{pmatrix} 1 & 1-e^{-T} & 0 & 0 \\ 0 & e^{-T} & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$H_1 = \begin{pmatrix} T - 1 + e^{-T} \\ 1 - e^{-T} \\ 0 \\ 0 \end{pmatrix} \qquad G_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$Q_1 = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \qquad Z_0 = \begin{pmatrix} 0 \\ 0 \\ 5 \\ 0 \end{pmatrix}$$

$$Q_2 = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

where T = 1. The input data used to formulate the algorithm for the three cases are listed in Table 2. The noise-balance component, W, and the feedback coefficient, $\alpha$, of the optimum controller are calculated with the algorithm. With this controller, the minimum of the EMQPI* is calculated by using equation (39). In this example $\bar{f} = \bar{\bar{f}} = p$, the constant terms in equation (39) can be combined as $\bar{f}^2 U(N) + \bar{\bar{f}} V(N) = pL(N)$, where $L = pU+V$. The results are tabulated in Table 3.

The asymptotic behavior of the convergence of the terms $\alpha$, $\beta$, M, K, L, and $I_N^*$ is discussed below. The terms, $\alpha$, $\beta$, K and L converge with accuracy to the 5th decimal place at the 14th, 16th and 16th iteration for cases 1, 2, and 3 respectively. The terms, $\alpha$ and $\beta$, converge monotonically to constants. Their rate of convergence for case 3 is shown in Figures 5 and 6 and appears to be close to exponential. The same terms for cases 1 and 2 have similar behavior and the corresponding figures are omitted.[16] The constant, L, the last element of the matrix, M, $m_{44}$, and the last element of the vector, K, $k_4$, are asymptotic to straight lines of non-zero slopes. The rate of approach of L, $m_{44}$ and $k_4$ is shown in Figures 7, 8 and 9. The remaining elements of M and K converge to constants faster than do $m_{44}$ and $k_4$. The fact that L increases with the number of stages or time is expected; it can be seen in equation (39) that L associates with the noise and the control cost, both of which contribute to $J_N^*$ as N increases.

The asymptotic behavior of $\bar{I}_N^*$ is shown in Figure 10. In this example the limit of $\bar{I}_N^*$ as $N \to \infty$ is $p\Delta L$ where:

$$\Delta L = \lim_{N\to\infty} \left[ L(N+1) - L(N) \right]$$

This is verified by the fact that $Z_t MZ$ and $pK_t Z$ in (39) are constants at large N and that L increases by $\Delta L$ at each stage as $N \to \infty$.

## Conclusions

A method for designing an optimum controller for a linear plant subject to random noise is introduced. The controller is optimum in the sense that the system will satisfy the expected mean quadratic performance criteria. This investigation serves as another example of the application of dynamic programming techniques to stochastic control problems. Iterative formulas for performance indexes such as the squared system error which is weighted unequally at all times can be derived with the principle of stochastic control developed in this paper. This approach must be extended in the engineering aspect to formulate the algorithms for designing an adaptive controller, that is, a controller for a plant which is subject to random noise of unknown probability distribution, or for a plant which has parameters that vary according to environment.

## References

1. Kalman, R. E. and Koepcke, R. W., "Optimal Synthesis of Linear Sampling Control Systems Using Generalized Performance Indexes", Trans. ASME, Vol. 80, pp. 1820-1826, 1958

2. Joseph, P. D. and Ton J. T., "On Linear Control Theory", Application and Industry, Sept., 1961

3. Freimer, M., "A Dynamic Programming Approach to Adaptive Control Processes", IRE Trans. Automatic Control, V AC-4, Nov. 1959

4. Aoki, M., "Dynamic Programming Approach to a Final-Value Control System with a Random Variable Having an Unknown Distribution Function," IRE Trans. Automatic Control, V AC-5 Sept., 1960

5. Merriam, III, C. W., "A Class of Optimum Control Systems", Journal of Franklin Inst., April, 1959

6. Bellman, R., Adaptive Control Processes, Princeton University Press, 1961

7. Coddington, E. A. and Levinson, N., Theory of Ordinary Differential Equations, McGraw-Hill Book Company, New York, 1955, Chapter III

8. Kalman, R. E. and Bertram, J. E., "General Synthesis Procedure for Computer Control of Single and Multi-Loop Linear Systems", Trans. AIEE, Vol. 77, Pt. II, pp. 602-609, 1958

9. Kalman, R. E. and Bertram, J. E., "A Unified Approach to the Theory of Sampling Systems", Journal of Franklin Inst., May, 1959

10. Lees, A. B., "Interpolation and Extrapolation of Sampled Data", IRE Trans., Information Theory, IT-2, 1956, pp. 12-17

11. Bellman, R., Dynamic Programming, Princeton Hall, 1957

12. Bellman, R. and Kalaba, R., "Dynamic Programming and Adaptive Processes", IRE Trans. Automatic Control, Jan. 1960

13. Bellman, R. and Gross, O. A., "Some Aspects of the Mathematical Theory of Control Processes", Rand Report R-313, Santa Monica, California

14. Guillemin, E. A., The Mathematics of Circuit Analysis, John Wiley and Sons Inc., New York, 1949

15. Adorno, D. S., "Studies in the Asymptotic Theory of Control Systems: II", Tech. Report 32-99, Jet Propulsion Lab., CIT, Pasadena, California, June 1961

16. Chen, P. P., "An Algorithm for Stochastic Control Via Techniques of Dynamic Programming", Boeing Document No. D6-7858, Renton, Washington, Oct. 1961

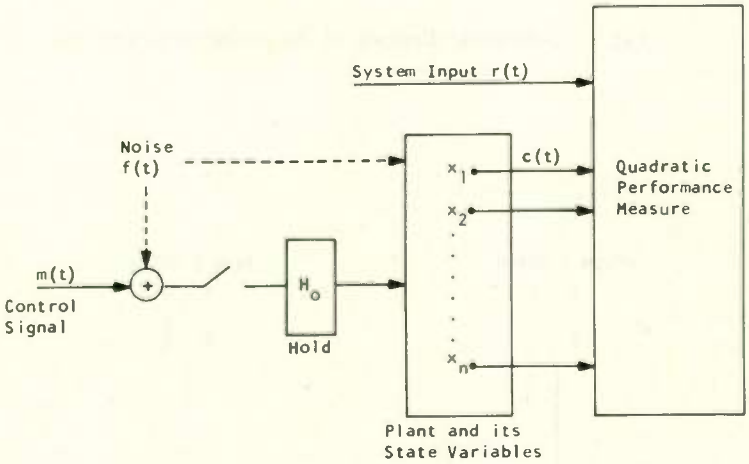


Fig. 1. System schematic diagram.

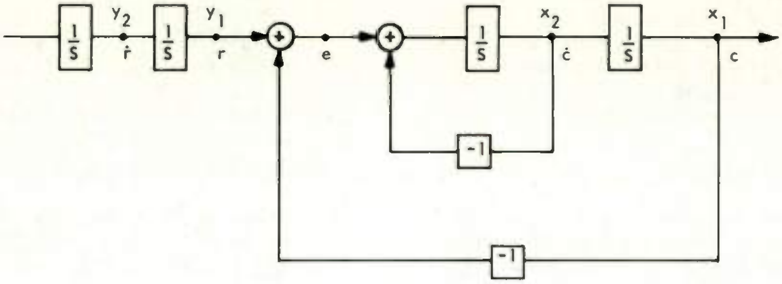


Fig. 2. System decomposition.

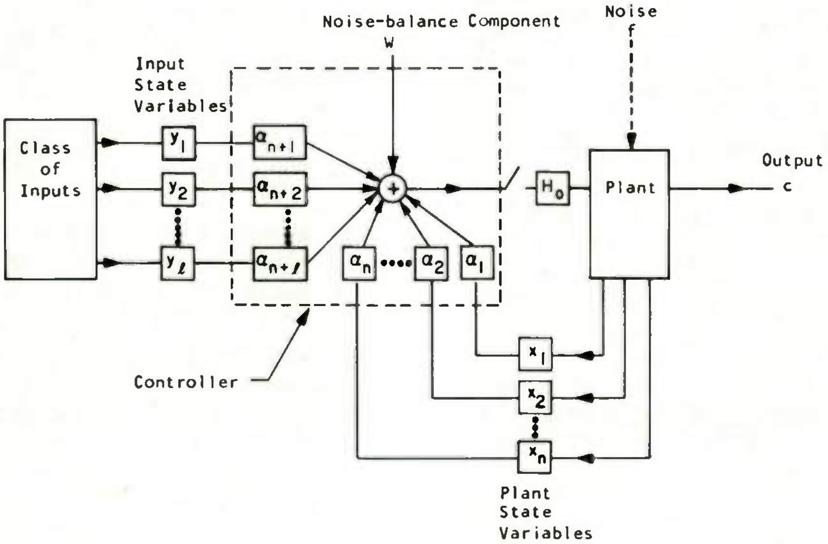


Fig. 3. Schematic diagram of the optimum controller.

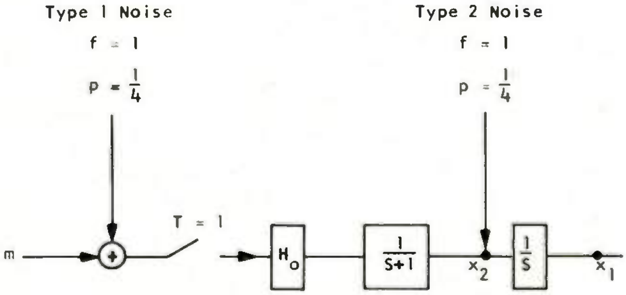


Fig. 4. The system in example.

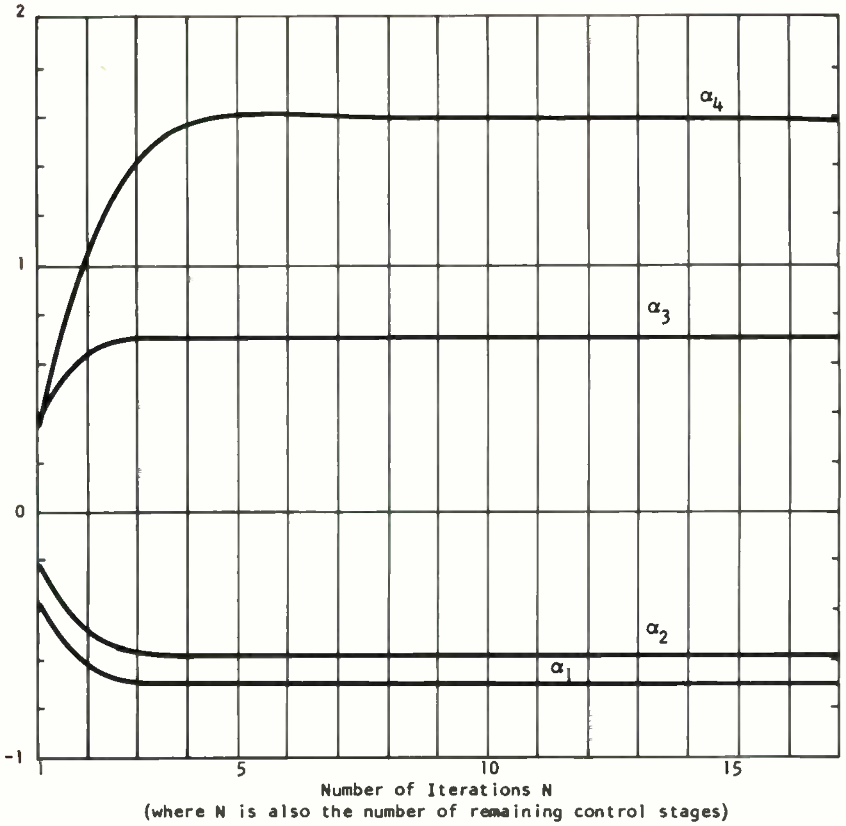


Fig. 5. Feedback coefficients $\alpha$'s for case 3.

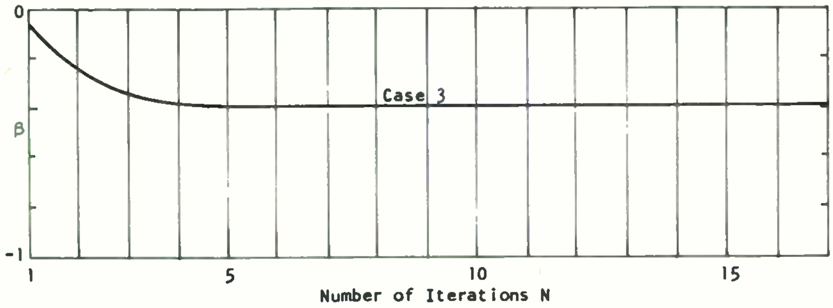


Fig. 6. $\beta$ for case 3.

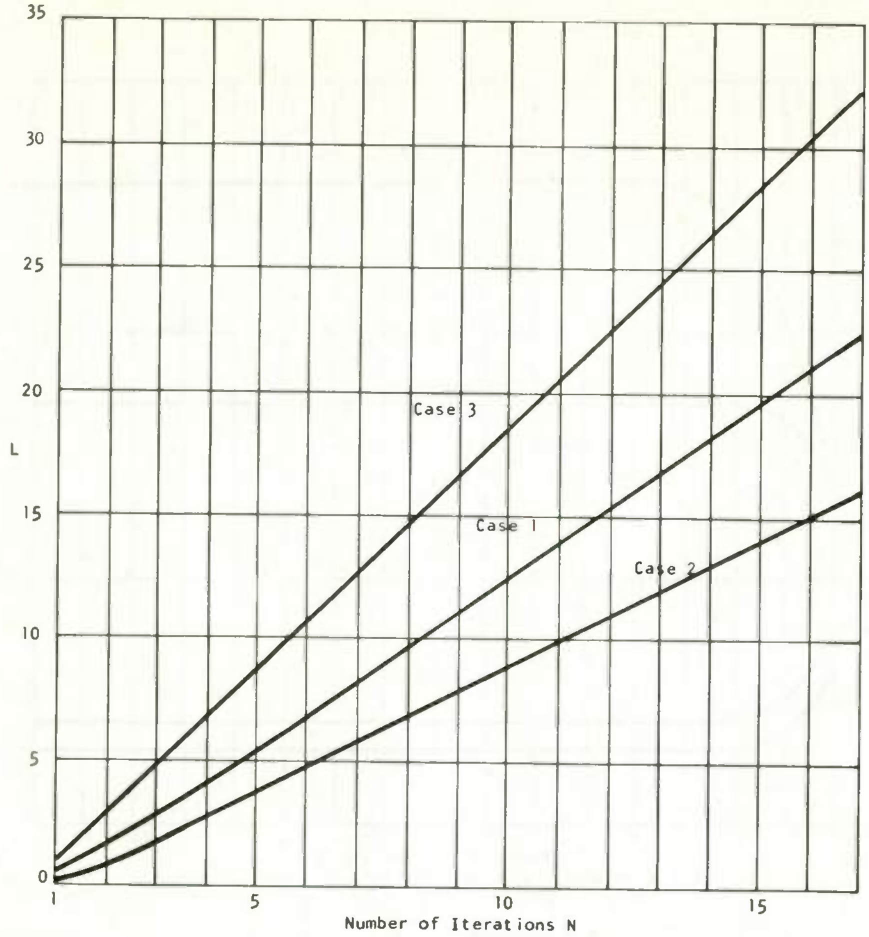


Fig. 7. L for cases 1, 2, and 3.

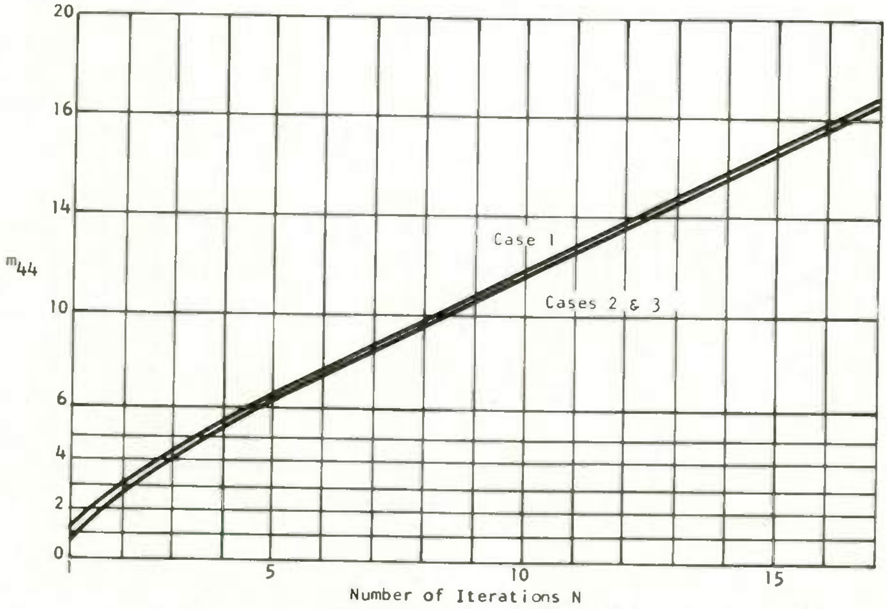


Fig. 8. $m_{44}$ For cases 1, 2, and 3.

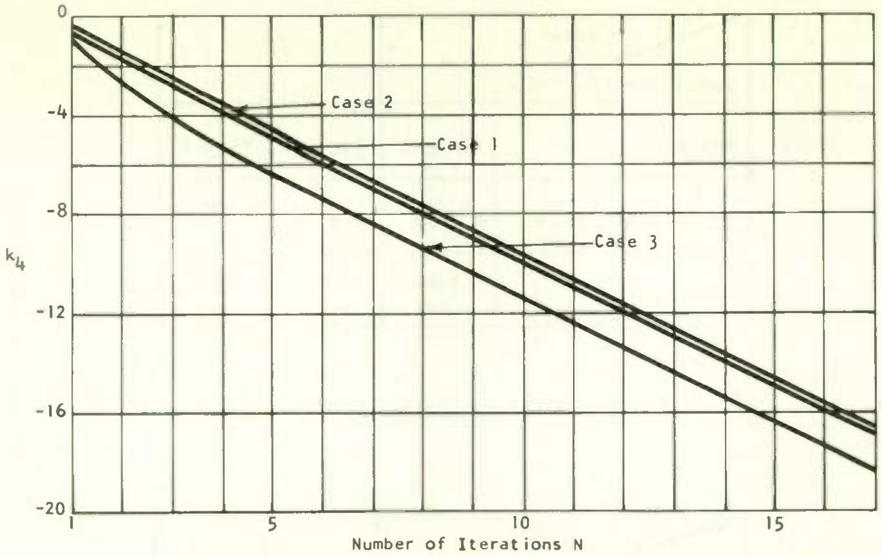


Fig. 9. $k_4$ For cases 1, 2, and 3.

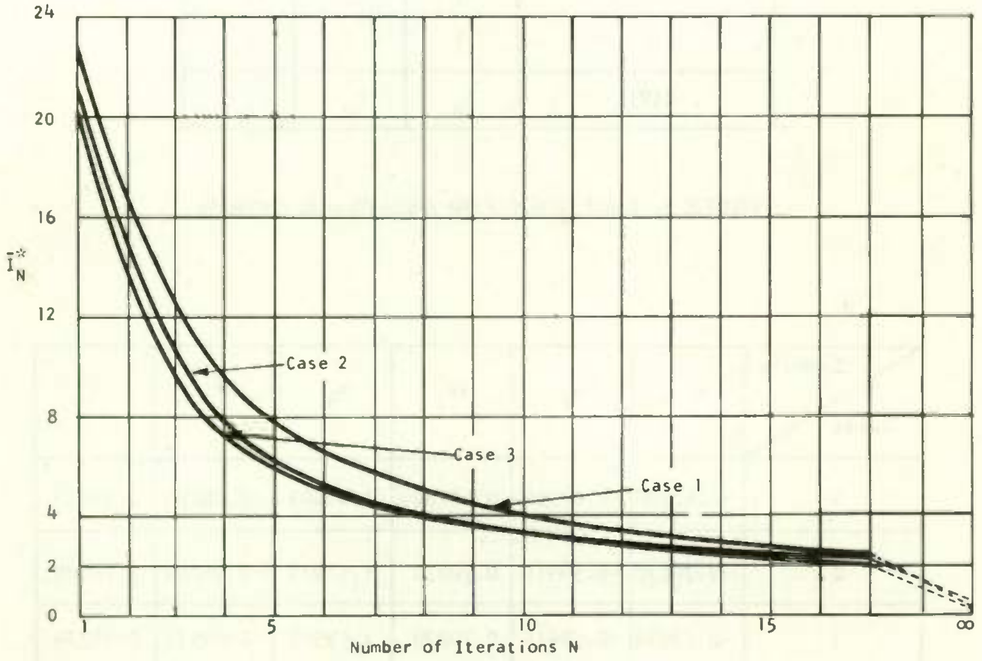


Fig. 10. The minimum of the EMQPI* for cases 1, 2, and 3.

| Cases<br>Specification | 1 | 2 | 3 |
|---|---|---|---|
| Noise | Type 1 | Type 1 | Type 2 |
| Performance Index | $e^2 + \dot{e}^2$ | $e^2$ | $e^2$ |
| Inputs | ramps<br>steps | ramps<br>steps | ramps<br>steps |

TABLE 1. Case Classification in Example

| Cases<br>Input Data | 1 | 2 | 3 |
|---|---|---|---|
| $\Phi$ | $\Phi_1$ | $\Phi_1$ | $\Phi_1$ |
| $H$ | $H_1$ | $H_1$ | $H_1$ |
| $G$ | $H_1$ | $H_1$ | $G_1$ |
| $Q$ | $Q_1$ | $Q_2$ | $Q_2$ |
| $\lambda$ | 1 | 1 | 1 |
| $Z(0)$ | $Z_0$ | $Z_0$ | $Z_0$ |

TABLE 2. Input Data for the Algorithm in Example

| Symbols<br>Cases | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $W$ | $\bar{I}^*$ |
|---|---|---|---|---|---|---|
| 1 | -0.62780 | -0.61504 | 0.62780 | 1.61504 | -0.06250 | 0.35073 |
| 2 | -0.70070 | -0.58915 | 0.70070 | 1.58915 | -0.06250 | 0.25689 |
| 3 | -0.70070 | -0.58915 | 0.70070 | 1.58915 | -0.09932 | 0.48774 |

TABLE 3. Results of the Example.

# DOUBLE MEASUREMENT WITH BOTH SAMPLED AND CONTINUOUS INPUTS

J. C. Hung
The University of Tennessee
Knoxville, Tennessee

## Introduction

A vital problem in the field of control and communication systems is multiple measurement. The problem is raised from the fact that in the modern control systems the instruments used for signal measurement are in general far from ideal due to the unavoidable instrument noise. As a result, the exact value of the measured signal can hardly be obtained. Scientists have been striving to explore the methods of obtaining the best estimate of the true signal from the noise-contaminated signal. It is very often that the required signal appears in several different forms. For instance, a practical problem occurring in guidance and control systems involves the separately obtained measurements of a position signal and its first derivative. It is obvious that the best estimate will be obtained if all the possible measurements are employed and their results are weighted and combined in an optimum way. Measurement of this kind will be referred to as multiple measurement. In the case of single measurement, the fundamental problem of extracting and predicting a signal from a mixture of signal and noise by means of an optimum physically realizable filter was first proposed by Wiener[1] who solved the problem on the basis of three assumptions, namely: (1) the time series representing the true signal and the noise are stationary and their auto-and cross-correlation functions are known; (2) the performance criterion of the filter is to minimize the mean square-error between the estimated value and the true value of the desired signal; and, (3) the operation used is assumed to be linear. Subsequently, further developments were made by Zadeh and Ragazzini,[2] Franklin,[3] and Lees[4] to include the cases of continuous-data finite memory filter, sampled-data infinite memory filter, and sampled-data finite memory filter respectively.

In the general case of multiple measurement, the measurable signals may appear in discrete forms as well as continuous forms. Therefore the measuring system may have continuous inputs and sampled inputs of various sampling rates. Theory on the general optimum multiple measurement, giving the best estimate of the desired signal, has not been developed. Recently, Hsieh and Leondes[5, 6] and Bendat[7] have treated a few special cases of multiple measurements. Their cases are special in the sense that the inputs of a measuring system are either all continuous, or all sampled with same sampling rate.

It is important to note that the theory of the general optimum multiple measurement cannot be obtained by a simple extension of the theories for the special cases mentioned in the last paragraph. The fundamental difference between the general cases and the special cases mentioned lies on the following fact. For multiple measuring system having either only continuous inputs or only sampled inputs of the same sampling rate, the optimum system is time-invariant if the input signals are stationary. But, the optimum multiple measuring system having continuous inputs and sampled inputs of different sampling rates is time-varying even though the inputs are stationary. Therefore the transfer function of an optimum multiple measuring system is, in general, a function of several variables, where the number of variables depends on the number of different input sampling rates. In this paper a procedure of optimum double measurement, having one sampled input and one continuous input as shown in Fig. 1, will be developed.

This system has important applications in the field of linear filtering and prediction, and in reduction of the load on a digital computer in a trajectory tracking system. For example, in guidance control, the available signals guiding the control of the vehicle frequently appear in both sampled and continuous forms. The measuring instrument used for continuous measurement is usually much more noisy than that for discrete measurement. Using the theory developed in this chapter one can weight and combine these two signals in an optimum way to obtain the best information. The second application originated from a missile trajectory tracking problem, where the trajectory of the missile is determined from the available input data. Accurate determination of the discrete points on the trajectory can be made using a digital computer. However, when a large number of points on the trajectory are required, very close to each other, a computer with a very large capacity is needed. To reduce the load capacity and to fill the information between the discrete determination of digital computer, an analog computer can be used in parallel with the former, as shown in Fig. 2. It is known that the analog computer introduces more computation error than the digital computer does. The final results produced by the two computers can be combined in an optimum fashion to give the best continuous determination of trajectory, using the theory developed in this chapter.

In the following, the method of solution for the optimum system under a general input condition and the error analysis of the obtained optimum system are developed. An example is given to illustrate the method. It is quite often that the noise in a sampled-data channel is negligible and the channel may be considered noise-free. Under this condition, a simplified method may be used for solution. Another example is given to illustrate this simplified method. It is also shown in the example how the qualities of the estimated signals obtained by double measurement and single measurements compare.

## Main Assumptions

There are four main assumptions upon which this research is based, namely:

(1) The actual signals measured by the measuring instruments are assumed to be linearly related to the desired signal. The relationships may be represented by the linear time-invariant transfer functions $M_1$ and $M_2$, (Fig. 1).

(2) The time series representing the signals and random noise are assumed ergodic stationary, and have rational spectral density functions.

(3) The criterion of performance used is to minimize the statistical mean square-error between the estimate and the desired signal.

(4) The operation of the optimum filters is assumed linear.

While these four assumptions are not chosen arbitrarily, they can be justified for many practical considerations.

## System Description

Fig. 1 is a general schematic representation of the system. In the figure, r is the desired stationary random signal to be estimated by the measuring system. The measuring system contains two parts, the measuring instruments which are fixed, and the optimum filters which are to be synthesized. The desired signal is measured by two different noisy measuring instruments whose operations are assumed linear and time-invariant having linear stationary transfer functions $M_1$ and $M_2$. In practice, the measuring instruments seldom have exactly linear operations. However, the small nonlinearities, which are usually undesirable, may be considered as equivalent instrument noise. The noise appearing at the output of each measuring instrument is denoted by $n_1$ and $n_2$ which represents the resultant of the internal noise, the external noise, and noise equivalent of the nonlinearity of that particular instrument. Quantities $r_1$ and $r_2$, are the hypothetical continuous outputs of the measuring instruments, which are the mixtures of the clean signals $r_{1c}$, $r_{2c}$, and the noise $n_1$, $n_2$, respectively. One actual output of the measuring instruments appears in sampled form while the other appears in continuous form. These outputs are also the inputs to the optimum filters. The function of the optimum filter $H_1$ and $H_2$ is to weight the output of the measuring instrument $M_1$ and $M_2$ in such a way that the outputs of two filters, when added, give the best continuous estimate, $r_e$, of the desired signal in the least-square sense. The hypothetical error generating scheme is shown on the figure by dashed lines. Before putting the system performance into the analytic expressions, a preliminary thought will be given about the nature of the optimum filters. The thought will lead one to an important observation which governs the number of variables contained in each optimum filter transfer function.

## Time Varying Characteristic of the Optimum Filters ---- An Observation[8]

For single-rate multiple measurement, the optimum filters are stationary if their inputs are stationary. But the optimum filters of multiple measurement are time-varying even though their inputs are stationary. This important characteristic must be included in the system analytic expression.

In Fig. 1, filter $H_1$ receives input $r_1$ at all instants, while filter $H_2$ receives input $r_2^*$ intermittently. Since the output of $H_2$ relies on its input only at sampling instants, the output between the sampling instants is produced entirely by a prediction operation based on the previous sampled input. The farther the time is away from the previous sampling instant (referring to Fig. 3) the more unreliable the prediction is. Therefore, when the input $r_2^*$ is less noisy than the input $r_1$, one would expect that the filter $H_1$ weights its input heavier in between the sampling instants than near the sampling instants, and the filter $H_2$ weights its input heavier at and near the sampling instants than in between the sampling instants. As a consequence, the impulse response of the optimum $H_1$, which is $h_1(t,\tau)$, must be a function of two independent variables t and $\tau$, where t is the time distance between the application time of input impulse and the observation time, and $\tau$ is the time interval between the last sampling instant and the observation time. In the same manner, the impulse response of the optimum $H_2$, which is $h_2(\rho,\tau)$, is also a function of two independent variables $\rho$ and $\tau$. $\tau$ has the same meaning as in $h_1$ and $\rho$ is the number of sampling instants between the application instant of the input impulse and the last sampling instant. In other words, $H_1$ is periodically time-varying filters whose period is the same as the sampling period of the discrete input $r_1^*$.

## Analytic Formulation of System Input-Output Relation

Having considered the nature of the optimum filters, one is in a position to formulate the system analytic expressions. In Fig. 3, T is sampling period, t is the time of observation, $t - \rho T - \tau$ is the instant when an impulse be applied to filter $H_1$. The functional notation of the two impulse responses of these filters are $h_1(t,\tau)$ and $h_2(\rho,\tau)$. These notations assume the following physical meaning. $h_1(t,\tau)$ is the output of filter $H_1$ at time t in response to an input impulse applied t seconds before observation time. $h_2(\rho,\tau)$ is the output of filter $H_2$ at time t in response to an input impulse applied $\rho\tau + \tau$ seconds before the observation time. The observation time is $\tau$ seconds behind the last sampling instant of the sampled channel.

The estimate, which is the output of the measuring system, is the algebraic sum of the outputs of three filters, and is given by

$$r_e(t) = \int_0^\infty r_1(t-t_1)h_1(t_1,\tau)\,dt_1 + \sum_{\rho=0}^{\infty} r_2(t-\rho T)h_2(\rho,\ T) \qquad (1)$$

where t is a dummy variable and $\rho$ is a dummy number. It should be noted that the response of any physical filter cannot depend on its future input, so the condition of physical realizability requires that the integration and the summations in Eq. 1 be taken over all the past inputs only.

## Mean Square-Error

The error of the estimate is

$$e(t) = r(t) - r_e(t) \qquad (2)$$

and its squared value is

$$e^2(t) = r^2(t) - 2r(t)r_e(t) + r_e{}^2(t) \qquad (3)$$

By substituting Eqs. 2, and 3 into 1, one obtains

$$\begin{aligned} e^2(t) &= r^2(t) \\ &+ \int_0^\infty r_1(t-t_1)h_1(t_1,\tau)dt_1 \int_0^\infty r_1(t-t_2)h_1(t_2,\tau)dt_2 \\ &+ 2\sum_{\rho=0}^{\infty} r_2(t-\tau-\ T)h_2(\rho,\tau)\int_0^\infty r_1(t-t_1)h_1(t_1,\tau)dt_1 \\ &+ \sum_{\rho=0}^{\infty} r_2(t-\tau-\rho T)h_2(\rho,\tau)\sum_{\sigma=0}^{\infty} r_2(t-\tau-\sigma T)h_2(\sigma,T) \\ &- 2\,r(t)\int_0^\infty r_1(t-t_1)h_1(t_1,\tau)dt_1 \\ &- 2\,r(t)\sum_{\rho=0}^{\infty} r_2(t-\tau-\rho T)h_2(\rho,\tau) \qquad (4) \end{aligned}$$

where $\sigma$ is a dummy number and $t_2$ is a dummy variable. Averaging both sides of Eq. 4 over the ensemble of all possible combination of inputs and denoting this mean square-error by I, gives

$$\begin{aligned} I &= \widetilde{e^2(t)} \\ &= \widetilde{r^2(t)} - 2\int_0^\infty \widetilde{r(t)r_1(t-t_1)}\,h_1(t_1,\tau)dt_1 \\ &- 2\sum_{\rho=0}^{\infty} \widetilde{r(t)(t-\tau-\rho T)}h_2(\rho,\tau) \\ &+ \int_0^\infty\int_0^\infty \widetilde{r_1(t-t_1)r_1(t-t_2)}h_1(t_1,\tau)h_2(t_2,\tau)dt_1dt_2 \\ &+ 2\sum_{\rho=0}^{\infty}\int_0^\infty \widetilde{r_1(t-t_1)r_2(t-\tau-\rho T)}h_1(t_1,\tau)h_2(\rho,\tau)dt_1 \\ &+ \sum_{\rho=0}^{\infty}\sum_{\sigma=0}^{\infty} \widetilde{r_2(t-\tau-\rho T)r_2(t-\tau-\sigma T)}h_2(\rho,\tau)h_2(\sigma,\tau) \ (5) \end{aligned}$$

The wavy symbol, $\widetilde{\qquad}$, in Eq. 5 means that the quantity under the symbol is ensemble averaged.

In the theory of random processes, the correlation function of two time functions x(t) and y(t) is defined as

$$\phi_{xy}(t_1,\ t_2) = \widetilde{x(t_1)\ y\ (t_2)} \qquad (6)$$

which is a function of two variables $t_1$ and $t_2$. It is called auto-correlation function if y(t) = x(t), and cross-correlation function if y(t) ≠ x(t).

When the random processes are ergodic, the correlation function depends only on the time distance between $t_1$ and $t_2$ rather than on both $t_1$ and $t_2$ themselves. Denote this time distance by $\eta$, then Eq. 6 becomes

$$\phi_{xy}(\eta) = \widetilde{x(t)\ y\ (t+\eta)}$$

Furthermore, under this condition, the ensemble mean, averaged over all possible x(t) y (t+ $\eta$) is equal to the time-average of any x(t) y (t+ $\eta$) over all t. That is

$$\begin{aligned} \phi_{xy}(\eta) &= \widetilde{x(t)\ y\ (t+\eta)} \\ &= \lim_{T_a\to\infty}\frac{1}{T_a}\int_0^{Ta} x(t)\ y\ (t+\eta)\ dt \qquad (8) \end{aligned}$$

where $T_a$ is an arbitrary long time interval over which the time-average is taken. Therefore Eq. 5 can be expressed in terms of the correlation functions of the desired and the actual signals. In practice, the exact time functions of the random signals are never known. The information given for the synthesis of the optimum filters are either the correlation functions of the signals, or their transforms which are also called the spectral density functions. Writing Eq. 5 in term of the correlation functions helps one to obtain the solution of optimum filters as functions of these known quantities.

$$\begin{aligned} I &= \phi_{ee}(o) \\ &= \phi_{rr}(o) - 2\int_0^\infty \phi_{r_1r}(t_1)h_1(t_1,\tau)dt_1 \\ &- 2\sum_{\rho}\phi_{r_2r}(\tau+\rho T)h_2(\rho,\tau) \\ &+ \int_0^\infty\int_0^\infty \phi_{r_1r_1}(t_1-t_2)h_1(t_1,\tau)h_1(t_2,\tau)\ dt_1dt_2 \\ &+ 2\sum_{\rho}\int_0^\infty \phi_{r_1r_2}(t_1-\tau-\rho T)h_1(t_1,\tau)h_2(\rho,\tau)\ dt_1 \\ &+ \sum_{\rho}\sum_{\sigma}\phi_{r_2r_2}(\rho\tau-\sigma T)h_2(\rho,\tau)h_2(\sigma,\tau) \qquad (9) \end{aligned}$$

Eq. 9 is the general form of the mean square-error which is to be minimized in the following section.

## Minimization

In this section the necessary and sufficient condition, that the optimum filters, $h_1$, $h_2$ must satisfy to ensure a minimum mean square-error, is obtained by the method of calculus of variations.

From now on the impulse responses $h_1(t_1,\tau)$ and $h_2(\rho,\tau)$ will be represented by $h_1(t_1)$ and $h_2(\rho)$, respectively, to simplify the mathematical expressions. It is understood that these two functions are also functions of $\tau$.

Let $g_1$ and $g_3$ be any differentiable function satisfying the condition

$$\begin{aligned} g_1(t_1) &= 0 \qquad t_1 < 0 \\ g_2(\rho) &= 0 \qquad \tau < 0 \end{aligned} \qquad (10)$$

where g's are also functions of $\tau$. Then, if h's in Eq. 9 are replaced by $(h + \epsilon g)$'s, where $\epsilon$ is a small real number, the effect will be to increase I by an amount of $\Delta I$ which is called the variation of I. The variation $\Delta I$ is obtained from Eq. 9 as

$$I = \sum_{\rho}\sum_{\sigma} \phi_{r_2r_2}(\sigma T - \rho T)\Big[\epsilon g_2(\rho)h_2(\sigma) + \epsilon g_2(\sigma)h_2(\rho) + \epsilon^2 g_1(\rho)\, g_1(\sigma)\Big]$$
$$+ \int_0^\infty \int_0^\infty \phi_{r_1r_1}(t_2-t_1)\Big[\epsilon g_1(t_1)h_1(t_2) + \epsilon g_1(t_2)h_1(t_1) + \epsilon^2 g_1(t_1)g_1(t_2)\Big]dt_1dt_2$$
$$- 2\sum_{\rho} \phi_{r_2r}(\rho T + \tau)\epsilon g_2(\rho)$$
$$- 2\int_0^\infty \phi_{r_1r}(t_1)\epsilon g_1(t_1)\, dt_1$$
$$+ 2\sum_{\rho}\int_0^\infty \phi_{r_1r_2}(t_1 - \rho T - \tau)\Big[\epsilon g_2(\rho)h_1(t_1) + \epsilon g_1(t_1)h_2(\rho) + \epsilon^2 g_2(\rho)g_1(t_1)\Big]\, dt_1 \qquad (11)$$

Assume that for any physical realizable g's, I has continuous derivatives with respect to $\epsilon$. This implies unique derivative at each point, and assures the differentiability of I with respect to $\epsilon$. h's for which $I = I_o$ is minimum must satisfy the condition

$$\left[\frac{d}{d\epsilon}(I + \Delta I)\right]_{\epsilon=0} = \left[\frac{d}{d\epsilon}(\Delta I)\right]_{\epsilon=0} = 0 \qquad (12)$$

for all physically realizable g's. Differentiating Eq. 11 with respect to $\epsilon$ and then setting $\epsilon$ equal to zero, gives

$$\left[\frac{d}{d\epsilon}(\Delta I)\right]_{\epsilon=0}$$
$$= \sum_{\rho}\sum_{\sigma} \phi_{r_2r_2}(\sigma T - \rho T)\Big[g_2(\rho)h_2(\sigma) + g_2(\sigma)h_2(\rho)\Big]$$
$$+ \int_0^\infty \int_0^\infty \phi_{r_1r_1}(t_2-t_1)\Big[g_1(t_1)h_1(t_2) + g_1(t_2)h_1(t_1)\Big]dt_1dt_2$$
$$- 2\sum_{\rho}\phi_{r_2r}(\rho T + \tau)g_2(\rho) - 2\int_0^\infty \phi_{r_1r}(t_1)g_1(t_1)dt_1$$
$$+ 2\sum_{\rho}\int_0^\infty \phi_{r_1r_2}(t_1 - \rho T - \tau)\Big[g_2(\rho)h_1(t_1) + g_1(t_1)h_2(\rho)\Big]\, dt_1 = 0 \qquad (13)$$

By changing $\rho$ to $\sigma$ and $\sigma$ to $\rho$ in the first term and noting that $\phi_{r_2r_2}$ is an even function, the first term becomes

$$2\sum_{\rho}\sum_{\sigma} \phi_{r_2r_2}(\sigma T - \rho T)g_2(\rho)\; h_2(\sigma). \qquad (14)$$

Similarly, the second term may be written as

$$2\int_0^\infty \int_0^\infty \phi_{r_1r_1}(t_2-t_1)g_1(t_1)h_1(t_2)\; dt_1dt_2. \qquad (15)$$

Combining Eqs. 13, 14, 15 and rearranging the terms

$$2\int_0^\infty g_1(t_1)\Big[\int_0^\infty \phi_{r_1r_1}(t_2-t_1)h_1(t_2)dt_2 + \sum_{\rho}\phi_{r_1r_2}(t_1 - \rho T - \tau)h_2(\rho) - \phi_{r_1r}(t_1)\Big]dt_1$$
$$+ 2\sum_{\rho} g_2(\rho)\Big[\int_0^\infty \phi_{r_1r_2}(t_1 - \rho T - \tau)h_1(t_1)dt_1 + \sum_{\sigma}\phi_{r_2r_2}(\sigma T - \rho T)h_2(\sigma) - \phi_{r_2r}(\rho T - \tau)\Big] = 0 \qquad (16)$$

Since Eq. 16 must hold for any physically realizable g's, it requires that the expressions inside the brackets must equal to zero individually. That is,

$$\int_0^\infty \phi_{r_1r_1}(t_2-t_1)h_1(t_2)dt_2 + \sum_{\rho}\phi_{r_1r_2}(t_1 - \rho T - \tau)h_2(\rho) - \phi_{r_1r}(t_1) = 0 \qquad t \geqq 0 \qquad (17)$$

$$\int_0^\infty \phi_{r_1r_2}(t_1 - \rho T - \tau)h_1(t_1)dt_1 + \sum_{\sigma}\phi_{r_2r_2}(\sigma T - \rho T)h_2(\sigma) - \phi_{r_2r}(\rho T - \tau) = 0 \qquad \rho \geqq 0 \qquad (18)$$

The above derivation has shown that Eqs. 17 and 18 impose the necessary condition that the optimum filter h's must satisfy. The conition is also sufficient. For

$$1/2\left[\frac{d^2}{d\epsilon^2}(\Delta I)\right]_{\epsilon=0} = \sum_{\rho}\sum_{\sigma}\phi_{r_2r_2}(\sigma T - \rho T)g_1(\rho)g_1(\sigma)$$
$$+ \int_0^\infty \int_0^\infty \phi_{r_1r_1}(t_2-t_1)g_1(t_1)g_1(t_2)\; dt_1dt_2$$
$$+ 2\sum_{\rho}\int_0^\infty \phi_{r_1r_2}(t_1 - \rho T - \tau)g_2(\rho)g_1(t_1)dt_1 \geqq 0 \qquad (19)$$

which shows that at the extreme point the variation function concaves upward and the point is a minimum. The $\geqq$ sign, in Eq. 19, holds, since the equation has the form of

$$A^2 + 2AB + B^2 = (A + B)^2 \qquad (20)$$

which is always nonnegative for real A and B.

Eqs. 17 and 18 are two integral equations representing the basic relations in the design of optimum linear filters shown in Fig. 1 on a mean square-error basis. Solution of these equations yields the optimum filters. It is observed that Eqs. 17 and 18 resemble the convolution integrals[21] and summations. This suggests that their solution may be obtained by transforming the equations to and solving the equations in the frequency domain. Since the integral equations hold only for certain ranges of the time variables and since the correlation functions have nonvanishing values outside these ranges, some modifications have to be made in taking the transforms.

## Solution of Integral Equations

Let

$$f_1(t_1) = \int_0^\infty \phi_{r_3r_3}(t_2-t_1)\, h_3(t_2)\, dt_2$$

$$+ \sum_{\rho=0}^{\infty} \phi_{r_2r_1}(\rho T+\tau-t_1)h_2(\rho) - \phi_{r_1r}(t_1) \qquad (21)$$

$$f_2(\rho) = \int_0^\infty \phi_{r_1r_2}(t_1-\rho T-\tau)\, h_1(t_1)\, dt_1$$

$$+ \sum_{\sigma=0}^{\infty} \phi_{r_2r_2}(\sigma T-\rho T)\, h_2(\sigma) - \phi_{r_2r}(\rho T+\tau) \qquad (22)$$

Then, Eqs. 17 and 18 are equivalent to

$$f_1(t_1) = 0 \qquad t_1 \gtreqless 0 \qquad (23)$$

$$f_2(\rho) = 0 \qquad \rho \gtreqless 0 \qquad (24)$$

Taking the two-sided Laplace transforms of Eqs. 17 and 18 with respect to $t_1$ and $\rho$, respectively, one obtains

$$\Phi_{r_1r_1}(s)H_1(s) + \Phi_{r_1r_2}(s)H_2(z)e^{-s\tau} - \Phi_{r_1r}(s) = F_1(s) \qquad (25)$$

$$\left[\Phi_{r_2r_1}(s)\, H_1(s)e^{s\tau}\right]^* + \Phi_{r_2r_2}(z)\, H_2(z)$$

$$- \left[\Phi_{r_2r}(s)e^{s\tau}\right]^* = F_2(z) \qquad (26)$$

where $[\ \ ]^*$ denotes that the term inside the brackets is Z-transformed,

$\Phi_{r_1r_1}(s)$, $\Phi_{r_2r_2}(z)$ = the power-spectral density functions of $r_1$ and $r_2^*$ respectively, and

$\Phi_{r_1r_2}(s)$, $\Phi_{r_1r}(s)$, $\Phi_{r_2r_1}(s)$, $\Phi_{r_2r}(s)$

= the cross-spectral density functions of the signals $r_1$, $r_2$ and the desired signal r.

Note that $H_1(s)$ is a function of s and $\tau$, and $H_2(z)$ is a function of z and $\tau$. The transfer function $F_1(s)$ and $F_2(z)$ should be analytic on the left-half of the s-plane and inside unit-circle of the z-plane, respectively, since $f_1(t)$ and $f_2(\rho)$ vanish over the range $(0,\infty)$ as expressed in Eqs. 23 and 24. In other words, $F_1(s)$ may have poles only in the right-half s-plane, while $F_2(z)$ may have poles only outside the unit-circle of z-plane. Further, from Eq. 24, $f_2(\rho) = 0$ for $\rho \gtreqless 0$, therefore $F_2(z)$ can be expressed as an ascending polynomial in the positive power of z without the constant term.

$$F_2(z) = z \sum_{\rho=0}^{\infty} C_\rho z^\rho = zG_2(z)$$

where all the poles of $G(z)$ are outside the unit-circle of z-plane.

In the following, Laplace transforms are represented by upper case letters, and their complex conjugates are represented by barred upper case letters. For example,

$$R = R(s) = \mathcal{L}[r(t)], \qquad \overline{R} = R(\overline{s})$$

Also, Z-transforms are represented by starred upper case letters, and their complex conjugates by starred upper case letters topped with bars. For example,

$$R^* = R(z) = R^*(s) = \mathcal{L}[r^*(t)],$$

$$\overline{R^*} = R(z^{-1}) = R^*(\overline{s})$$

Using the simplified notation defined Eqs. 25 and 26 become

$$\Phi_{r_1r_1} H_1 + \Phi_{r_1r_2} H_2^* e^{-s\tau} = \Phi_{r_1r} + F_1 \qquad (27)$$

$$z^{-1}\left[\Phi_{r_2r_1}H_1e^{s\tau}\right]^* + z^{-1}\Phi^*_{r_2r_2}H_2^* = z^{-1}\left[\Phi_{r_2r}e^{s\tau}\right]^* + G_2^* \qquad (28)$$

which are to be solved for $H_1$ and $H_2^*$. Because of the mixing of the Laplace transforms and the Z-transforms, Eqs. 27 and 28 cannot be solved directly by the method of matrices. A method of elimination and substitution will be used instead.

Multiply Eq. 27 by $\dfrac{\Phi_{r_2r_1}}{\Phi_{r_1r_1}}$ gives

$$\Phi_{r_2r_1} H_1e^{s\tau} + \frac{\Phi_{r_1r_2}\Phi_{r_2r_1}}{\Phi_{r_1r_1}} H_2^* = \frac{\Phi_{r_1r}\Phi_{r_2r_1}}{\Phi_{r_1r_1}} e^{s\tau}$$

$$+ \frac{\Phi_{r_2r_1}}{\Phi_{r_1r_1}} F_1e^{s\tau}. \qquad (29)$$

Z-transform Eq. 29 and then multiply it by $z^{-1}$.

$$z^{-1}\left[\Phi_{r_2r_1}H_1e^{s\tau}\right]^* + z^{-1}\left[\frac{\Phi_{r_1r_2}\Phi_{r_2r_1}}{\Phi_{r_1r_1}}\right]^* H_2^*$$

$$= z^{-1}\left[\frac{\Phi_{r_1r}\Phi_{r_2r_1}}{\Phi_{r_1r_1}} e^{s\tau}\right]^* + z^{-1}\left[\frac{\Phi_{r_2r_1}}{\Phi_{r_1r_1}} F_1e^{s\tau}\right]^* \qquad (30)$$

By subtracting one from the other, the terms containing $H_1$ in Eqs. 28 and 30 can be eliminated as

$$z^{-1}H_2\left\{\Phi^*_{r_2r_2} - \left[\frac{\Phi_{r_1r_2}\Phi_{r_2r_1}}{\Phi_{r_1r_1}}\right]^*\right\} = z^{-1}\left[\Phi_{r_2r}e^{s\tau}\right]^*$$

$$-z^{-1}\left[\frac{\Phi_{r_1r}\Phi_{r_2r_1}}{\Phi_{r_1r_1}} e^{s\tau}\right]^* + G_2^* - z^{-1}\left[\frac{\Phi_{r_2r_1}}{\Phi_{r_1r_1}} F_1e^{s\tau}\right]^* \qquad (31)$$

or simply

$$z^{-1}H_2^*\left[\Phi_{r_2r_2} - \frac{\Phi_{r_1r_2}\Phi_{r_2r_1}}{\Phi_{r_1r_1}}\right]^*$$

$$= z^{-1}\left[\left(\Phi_{r_2r} - \frac{\Phi_{r_1r}\Phi_{r_2r_1}}{\Phi_{r_1r_1}}\right) e^{s\tau}\right]^*$$

$$+ G_2^* - z^{-1}\left[\frac{\Phi_{r_2r_1}}{\Phi_{r_1r_1}} F_1 e^{s\tau}\right]^* \qquad (32)$$

The factor $\left[\Phi_{r_2r_2} - \frac{\Phi_{r_1r_2}\Phi_{r_2r_1}}{\Phi_{r_1r_1}}\right]^*$ in the left-hand side of Eq. 32 is a rational function in z, and is symmetrical with respect to z and $z^{-1}$. It can be written as the product of two factors

$$Y^* \overline{Y}^* = \left[\Phi_{r_2r_2} - \frac{\Phi_{r_1r_2}\Phi_{r_2r_1}}{\Phi_{r_1r_1}}\right]^* \qquad (33)$$

where $Y^*$has all its poles and zeros inside the unit-circle, while $\overline{Y}^*$ has all its outside. Substituting Eq. 33 into Eq. 32 and dividing both sides by $\overline{Y}^*$ give

$$z^{-1}Y^* H_2^* = \frac{z^{-1}}{\overline{Y}^*}\left[\left(\Phi_{r_2r} - \frac{\Phi_{r_1r}\Phi_{r_2r_1}}{\Phi_{r_1r_1}}\right) e^{s\tau}\right]^* + \frac{G_2^*}{\overline{Y}^*}$$

$$- \frac{z^{-1}}{\overline{Y}^*}\left[\frac{\Phi_{r_2r_1}}{\Phi_{r_1r_1}} F_1 e^{s\tau}\right]^* \qquad (34)$$

In general, a rational algebraic function Q(z) in z can be written as a partial fraction expansion in terms of the roots of the denominator and a polynomial P(z) in z.

$$Q(z) = \sum_{|\alpha_i|<1} \frac{A_i}{z-\alpha_i} + \sum_{|\beta_j|>1} \frac{B_j}{z-\beta_j} + P(z)$$

$$= \Big\{Q(z)\Big\}_i + \Big\{Q(z)\Big\}_o \qquad (35)$$

where

$$\Big\{Q(z)\Big\}_i = \sum_{|\alpha_i|<1} \frac{A_i}{z-\alpha_i} = \sum_{|\alpha_i|<1} \frac{A_i z^{-1}}{1-\alpha_i z^{-1}} \qquad (36)$$

is analytic outside the unit-circle, and

$$\Big\{Q(z)\Big\}_o = \sum_{|\beta_j|>1} \frac{B_j}{z-\beta_j} + P(z) \qquad (37)$$

is analytic inside the unit-circle. Note that the constant term of the polynomial P(z) should be grouped into $Q_o(z)$. The reason for doing so will become apparent later.

In Eq. 34, the term on the left-hand side is analytic outside the unit-circle, since $H_2^*$ is implemented to be a stable function and $Y^*$ is analytic outside the unit-circle by definition. The first term on the right-hand side of Eq. 34, which is completely known, may have poles both inside and outside the unit-circle. Using the method characterized by Eqs. 35, 36, and 37, this term may be written as the sum of two terms, one is analytic outside the unit-circle and the other is analytic inside the unit-circle. The second right side term of Eq. 34 is known to be analytic inside the unit-circle from the definition of $G_2^*$ and $Y_2^*$. The last term

$$\left[\frac{1}{\overline{Y}^*} \frac{\Phi_{r_2r_1}}{\Phi_{r_1r_1}} F_1 e^{s\tau}\right]^* \qquad (38)$$

on the right-hand side of Eq. 34 again may have poles both inside and outside the unit-circle. This term is not completely known, therefore the partial fraction technique of Eq. V-20 cannot be applied. Examining Eq. (38) closely, one sees that all its inside poles are known, since they are the inside poles of the known quantity

$$\left[\frac{\Phi_{r_2r_1}}{\Phi_{r_1r_1}}\right]^* .$$

Denote these poles by $\alpha_i$'s, then the part of Eq. (38) which is analytic outside the unit-circle may be expressed as

$$\left\{\frac{1}{\overline{Y}^*}\left[\frac{\Phi_{r_2r_1}}{\Phi_{r_1r_1}} F_1 e^{s\tau}\right]^*\right\}_i = \sum_i \frac{A_i z^{-1}}{1-\alpha_i z^{-1}} \qquad (39)$$

where $|\alpha_i| < 1$, and $A_i$'s are constants to be determined. Then Eq. 34 can be written as

$$z^{-1} Y^* H_2^* - \left\{\frac{z^{-1}}{\overline{Y}^*}\left[\left(\Phi_{r_2r} - \frac{\Phi_{r_1r}\Phi_{r_2r_1}}{\Phi_{r_1r_1}}\right) e^{s\tau}\right]^*\right\}_i$$

$$- \sum_i \frac{A_i z^{-1}}{1-\alpha_i z^{-1}} = \left\{\frac{z^{-1}}{\overline{Y}^*}\left[\left(\Phi_{r_2r} - \frac{\Phi_{r_1r}\Phi_{r_2r_1}}{\Phi_{r_1r_1}}\right) e^{s\tau}\right]^*\right\}_o$$

$$+ \frac{G_2^*}{\overline{Y}^*} + \left\{\frac{1}{\overline{Y}^*}\left[\frac{\Phi_{r_2r_1}}{\Phi_{r_1r_1}} F_1 e^{s\tau}\right]^*\right\}_o , \qquad (40)$$

the left-hand side of which is analytic outside the unit-circle while the right-hand side of it is analytic inside the unit-circle. In order to satisfy Eq. 40 both sides must equal to a constant $K_1$.

Thus,

$$z^{-1}\,Y^{*}\,H_2^{*} - \left\{ \frac{z^{-1}}{\overline{Y^{*}}} \left[ \left( \Phi_{r_2 r} - \frac{\Phi_{r_1 r}\Phi_{r_2 r_1}}{\Phi_{r_1 r_1}} \right) e^{s\tau} \right]^{*} \right\}_{i}$$

$$- \sum_{i} \frac{A_i z^{-1}}{1-\alpha_i z^{-1}} = K_1$$

Dividing by $z^{-1}Y^{*}$ and rearranging the terms, one obtains

$$H_2^{*} = \frac{z}{Y^{*}} \left\{ \frac{z^{-1}}{\overline{Y^{*}}} \left[ \left( \Phi_{r_2 r} - \frac{\Phi_{r_1 r}\Phi_{r_2 r_1}}{\Phi_{r_1 r_1}} \right) e^{s\tau} \right]^{*} \right\}_{i}$$

$$+ \frac{z}{Y^{*}} \sum_{i} \frac{A_i z^{-1}}{1-\alpha_i z^{-1}} + \frac{z\,K_1}{Y^{*}} \qquad (41)$$

Examining Eq. 41 one sees that the term $H_2^{*}$ does not have a z term, and the factors

$$\left\{ \frac{z^{-1}}{Y^{*}} \left[ \left( \Phi_{r_2 r} - \frac{\Phi_{r_1 r}\Phi_{r_2 r_1}}{\Phi_{r_1 r_1}} \right) e^{s\tau} \right]^{*} \right\}_{i} \quad \text{and}$$

$$\sum_{i} \frac{A_1 z^{-1}}{1-\alpha_i z^{-1}}$$

do not have a constant term. Consequently, $K_1$ must be zero. The optimum $H_2^{*}$ is

$$H_2^{*} = \frac{z}{Y^{*}} \left\{ \frac{z^{-1}}{\overline{Y^{*}}} \left[ \left( \Phi_{r_2 r} - \frac{\Phi_{r_1 r}\Phi_{r_2 r_1}}{\Phi_{r_1 r_1}} \right) e^{s\tau} \right]^{*} \right\}_{i}$$

$$+ \frac{1}{Y^{*}} \sum_{i} \frac{A_i}{1-\alpha_i z^{-1}} \qquad (42)$$

The unknown constants $A_i$'s are remained to be determined. Remembering that in separating the function $Q(z)$ into $\{Q(z)\}_i$ and $\{Q(z)\}_o$, as shown in Eq. 35, the constant term of the polynomial $P(z)$ was grouped into $\{Q(z)\}_o$. This should be so because if $\{Q(z)\}_i$ contained a constant then the expression for $H_2^{*}$, Eq. 42 would have a prediction term z which is not physically realizable.

The transfer function $H_1$ can be obtained by substituting Eq. 42 into 27 and proceeding as follows: In Eq. 27 the power spectral density $\Phi_{r_1 r_1}$, which is a rational and even function in s, may be written as the product of two functions

$$X\,\overline{X} = \Phi_{r_1 r_1} \qquad (43)$$

where X has all its poles and zeros on the LHP while $\overline{X}$ has all its poles on the RHP. Substituting Eq. 43 into 27, dividing both sides by $\overline{X}$, and rearranging the terms,

$$X\,H_1 = \frac{1}{\overline{X}} \left[ \Phi_{r_1 r} - \Phi_{r_1 r_2}\, e^{-s\tau} H_2^{*} \right] + \frac{F_1}{\overline{X}} \qquad (44)$$

Function $H_1$ is then to be obtained using a method similar to that for obtaining $H_2^{*}$.

Let U(s) be a closed function in s, which may have essential singularities on either LHP or RHP but not on both. Further, U(S) vanishes, being at least of the order of $\frac{1}{\omega}$ as ω approaches ∞. Then U(s) may be written as

$$U(s) = \left\{ U(s) \right\}_L + \left\{ U(s) \right\}_R \qquad (45)$$

where $\left\{U(s)\right\}_L$ has all its poles on the LHP while $\left\{U(s)\right\}_R$ has all its poles on the RHP, and both these two functions are closed.

In Eq. 40, the left-hand side term has only LHP poles, since $H_1$ is implemented to be a stable function and X has only LHP poles by definition. The second term on the right-hand side has only RHP poles from the definitions of $F_1$ and $\overline{X}$. The first term on the right-hand side may have poles on both sides of the s-plane. This term can be separated into two parts using the method characterized by Eq. 45. Therefore Eq. 44 may be expressed as

$$X\,H_1 - \left\{ \frac{1}{\overline{X}} \left[ \Phi_{r_1 r} - \Phi_{r_1 r_2}\, e^{-s\tau} H_2^{*} \right] \right\}_L$$

$$= \left\{ \frac{1}{\overline{X}} \left[ \Phi_{r_1 r} - \Phi_{r_1 r_2}\, e^{-s\tau} H_2^{*} \right] \right\}_R + \frac{F_1}{\overline{X}}$$

in which the left-hand side is analytic on the RHP while the right-hand side is analytic on the LHP. In order to satisfy this equation both sides must be equal to a constant $K_2$. Since the output of a physical system must vanish as ω approaches ∞, so $K_2$= o. Thus

$$X\,H_1 - \left\{ \frac{1}{\overline{X}} \left[ \Phi_{r_1 r} - \Phi_{r_1 r_2}\, e^{-s\tau} H_2^{*} \right] \right\}_L = o$$

Dividing this expression by X and rearranging the terms, the optimum $H_2$ is given by

$$H_1 = \frac{1}{X} \left\{ \frac{1}{\overline{X}} \left[ \Phi_{r_1 r} - \Phi_{r_1 r_2}\, e^{-s\tau} H_2^{*} \right] \right\}_L \qquad (46)$$

It remains to determine the constant $A_i$'s contained in Eq. 42. This can be done by substituting both Eqs. 42 and 46 into Eq. 28 and comparing the coefficients of the terms having like poles. It should be remembered that both optimum filters $H_2^{*}$ and $H_1$ are functions of two variables.

It is interesting to note that if the sampled branch does not exist, then $H_2^* = o$. Eq. 46 becomes

$$H_1 = \frac{1}{X}\left\{\frac{\Phi_{r_1 r}}{\overline{X}}\right\}_L \tag{47}$$

which is the well known Wiener filter [1], as it should be. On the other hand, if the continuous branch does not exist, then $H_1 = o$ and all the correlation functions relating to the input $r_1$ do not come into the calculation. Eq. 42 becomes

$$H_2^* = \frac{z}{Y^*}\left\{\frac{z^{-1}\left[\Phi_{r_2 r}e^{s\tau}\right]^*}{\overline{Y}^*}\right\}_i \tag{48}$$

which is similar to the result obtained by Franklin[3].

## Error Analysis

When $h_1$ and $h_2$ are the optimum functions, the necessary and sufficient condition expressed by Eqs. 17 and 18 must be satisfied. Substituting these two equations into Eq. 9, the mean square-error of the optimum system is given by

$$\phi_{ee}(o) = \Phi_{rr}(o) - \int_o^\infty \phi_{r_1 r}(t_1)h_1(t_1,\tau)dt_1$$

$$- \sum_{\rho} \phi_{r_2 r}(\rho T+\tau)h_2(\rho,\tau) \tag{49}$$

Eq. 49 gives the mean square-error of the optimum system in terms of filter impulse responses and the correlation functions. The mean square-error can also be expressed in terms of frequency domain quantities as[8]

$$\phi_{ee}(o) = \left[\mathcal{L}^{-1}\left\{\Phi_{rr}\right\}\right]_{\tau=o} - \left[\mathcal{L}^{-1}\left\{\Phi_{rr_1}\ H_1\right\}\right]_{t_2=o}$$

$$- \left[\mathcal{Z}^{-1}\left\{\left[\Phi_{rr_2}\ e^{-s\tau}\right]^*\ H_2^*\right\}\right]_{\rho=o} \tag{50}$$

Eq. 49 may be used to find the mean square-error when the correlation functions and filter impulse responses are known. On the other hand, when spectral density functions and filter transfer functions are known, one should use Eq. 50.

## Example I

To illustrate the method presented in previous section, consider a simple case where the sampled-input is noise-free and the noise in the continuous input is uncorrelated with the signal. The spectral density functions of the signals and noise, shown in Fig. 1, are

$$\Phi_{rr} = \Phi_{r_2 r_2} = \frac{4}{4-s^2}$$

$$\Phi_{r_1 r_1} = \frac{8-s^2}{4-s^2}$$

$$\Phi_{n_2 n_2} = o$$

$\Phi_{n_1 n_1} = 1$ (white noise)

The sampling period T=1. Since signals and noise are uncorrelated Eq. 33 reduces to

$$Y^*\ \overline{Y}^* = \left[\Phi_{r_2 r_2} - \frac{\Phi_{r_1 r_2}\Phi_{r_2 r_1}}{\Phi_{r_1 r_1}}\right]^*$$

$$= \left[\Phi rr - \frac{\Phi_{rr}^2}{\Phi_{r_1 r_1}}\right]^* = \left[\frac{\Phi_{rr}\Phi_{n_1 n_1}}{\Phi_{r_1 r_1}}\right]^* \tag{51}$$

Using the given data, this equation gives

$$Y^*\ \overline{Y}^* = \left[\frac{\dfrac{4}{4-s^2}}{\dfrac{8-s^2}{4-s^2}}\right]^*$$

$$= \frac{0.705}{(1-0.059z)(1-0.059z^{-1})} \tag{52}$$

Thus,

$$\left.\begin{aligned} Y^* &= \frac{0.839}{1-0.059z^{-1}} \\ \overline{Y}^* &= \frac{0.839}{1-0.059z} \end{aligned}\right\} \tag{53}$$

The factor

$$\left[\left(\Phi_{r_2 r} - \frac{\Phi_{r_1 r}\Phi_{r_2 r_1}}{\Phi_{r_1 r_1}}\right)e^{s\tau}\right]^* = \left[\frac{\Phi_{n_1 n_1}\Phi_{rr}}{\Phi_{r_1 r_1}}\ e^{s\tau}\right]^*$$

$$= \left[\frac{4e^{s\tau}}{8-s^2}\right]^*$$

$$= \frac{0.0839(b+az)}{(1-0.059z)\ (1-0.059z^{-1})} \tag{54}$$

where

$$a = \sinh\sqrt{8}$$

$$b = \sinh\ (1-\tau)\sqrt{8}$$

To find the poles of the last term of Eq. 42, which are inside the unit-circle, one sees that the quantity

$$\frac{\Phi_{r_2 r_1}}{\Phi_{r_1 r_1}} = \frac{\Phi_{rr}}{\Phi_{r_1 r_1}} = \frac{4}{8-S^2}$$

has a LHP pole at $s = -\sqrt{8}$. Therefore $\left[\frac{\Phi_{r_2 r_1}}{\Phi_{r_1 r_1}}\right]^*$

has an inside-pole at

$$z = \alpha = e^{-\sqrt{8}} = 0.059 \qquad (55)$$

Substituting Eqs. 53, 54, and 55 into Eq. 42 results in

$$H_2^* = \frac{(1-0.059z^{-1})z}{0.839}$$

$$\left\{z^{-1} \frac{(1-0.059z)}{0.839} \quad \frac{0.0839(b+az)}{(1-0.059z)\,(1-0.059z^{-1})}\right\}_i$$

$$+ \frac{1-0.059z^{-1}}{0.839} \quad \frac{A}{1-0.059z^{-1}}$$

$$= \frac{(1-0.059z^{-1})z}{0.839} \left\{\frac{0.1z^{-1}\ (b+az)}{1-0.059z^{-1}}\right\}_i + \frac{A}{0.839}$$

Thus, the optimum $H_2^*$ is given by

$$H_2^* = \frac{0.1(b+0.059a) + A}{0.839} = K \qquad (56)$$

where K is a constant with respect to z, but a function of $\tau$. To find the optimum $H_1$, Eq. 43 is first used to give

$$\Phi_{r_1 r_1} = \frac{8-s^2}{4-s^2} = X\,\overline{X}$$

Therefore

$$\left.\begin{aligned} X &= \frac{\sqrt{8}+s}{2+s} \\ \overline{X} &= \frac{\sqrt{8}-s}{2-s} \end{aligned}\right\} \qquad (57)$$

Note that

$$\Phi_{r_1 r} = \Phi_{r_1 r_2} = \Phi_{rr} = \frac{4}{4-s^2} \qquad (58)$$

Substituting Eqs. 56, 57, and 58 into 46

$$H_1 = \frac{s+2}{s+\sqrt{8}} \left\{ \frac{s-2}{s-\sqrt{8}} \left[ \frac{4}{4-s^2} - \frac{4}{4-s^2}\, e^{-s\tau} K \right] \right\}_L$$

$$\frac{s+2}{s+\sqrt{8}} \left\{ \frac{4}{(\sqrt{8}-s)(2+s)} \right\}_L - \frac{s+2}{s+\sqrt{8}} \left\{ \frac{4e^{-s\tau} K}{(8-s)(2+s)} \right\}_L$$

Therefore the optimum $H_1$ is

$$H_1 = \frac{0.828}{s+\sqrt{8}} + \frac{4e^{-s\tau} K}{s^2-8} - \frac{0.828\, e^{-\sqrt{8}\tau}\,(s+2)\,K}{s^2-8} \qquad (59)$$

It should be pointed out that the term inside the second braces of Eq. V-61 does not converge for $\mathrm{Re}\{s\} = -\infty$. Therefore the residue of the LHP term cannot be found. However, the residue of the RHP terms can readily be evaluated. So the part, which is analytic in the RHP, is obtained by taking the difference between the original term and the RHP term. Consequently the residue of $H_1$ at the RHP pole $s = \sqrt{8}$ vanishes resulting in a stable filter.

The constant A, or K, or Eq. 56 is obtained by substituting both $H_2^*$ and $H_1$ into Eq. 28 and equating all the terms having like poles which are inside the unit-circle, one obtains the equation

$$-0.828\, e^{-\sqrt{8}} + 0.705\, K + 0.121\, K\, e^{-2\sqrt{8}\tau} = 0.$$

Solving this equation the value of K is found

$$K = \frac{1.175\, e^{-\sqrt{8}\tau}}{1 + 0.175\, e^{-2\sqrt{8}\tau}} \qquad (60)$$

This completes the solution of this problem. The solution is rewritten in the following.

$$H_2^* = K = \frac{1.175\, e^{-\sqrt{8}\tau}}{1 + 0.175\, e^{-2\sqrt{8}\tau}} \qquad (61)$$

$$H_1 = \frac{0.828}{s+8} + \frac{4e^{-s\tau} K}{s^2-8} - \frac{0.828\, e^{-\sqrt{8}\tau} K(s+2)}{s^2-8} \qquad (62)$$

It is interesting to note that when $\tau = 0$

$$\left.\begin{aligned} H_2^* &= K = 1 \\ H_1 &= 0 \end{aligned}\right. \qquad (63)$$

as they should be, since the sampled-input is noise-free and should have complete transmission at sampling instants.

The impusle responses of $h_2(\rho)$ and $h_1(t)$ are given by

$$h_2(\rho) = \frac{1.175\, e^{-\sqrt{8}}}{1 + 0.175\, e^{-2\sqrt{8}\tau}}\, \delta_{\rho,\,o} \qquad (64)$$

where $\delta_{\rho,o}$ is the Kronecker delta, and

$$h_1(t) = 0.828\, e^{-\sqrt{8}t} + 1.414\, K \sinh\sqrt{8}\,(t-\tau)\, U(t-\tau)$$
$$- 0.585\, e^{-\sqrt{8}\tau} K \sinh\sqrt{8}t - 0.828\, e^{-\sqrt{8}\tau} K \cosh\ \sqrt{8}t \qquad (65)$$

Tables 1 and 2 show the calculated values of $h_2(\rho)$ and $h_1(t)$ as functions of $\rho$, $\tau$, and t, $\tau$, respectively. The impulse response curves are shown in Figs. 4 and 5.

Exact realization of the filters expressed by Eqs. 61 and 62 requires the use of amplifiers having exponentially time-varying gain. A proposed schematic diagram for filter $H_2$ is given in Fig. 6. However, since the impulse response of $H_2^*$ is one dimensional as shown in Fig. 4 , this filter can be approximated by lumped constant networks using curve-fitting technique. Filter $H_1$, as expressed in Eq. 62, can be realized by first plotting the impulse response surface of this filter and then approximate the response surface by means of passive network elements and varying gain amplifier.

The mean square-error can be calculated using Eq. 50 as[8]

$$\phi_{ee}(o) = 0.828 - 0.828\ e^{-\sqrt{8}\,\tau}\,K$$

The value of K is given by Eq. 60. Thus,

$$\phi_{ee}(o) = 0.828 - \frac{0.925}{e^{2\sqrt{8}\,\tau} + 0.175} \tag{66}$$

This equation gives the mean square-error averaged over the entire ensemble, and is a function of $\tau$ , the time distance between the last sampling instant and the observation time. Table 3 lists the values of $\Phi_{ee}(o)$ for various values of $\tau$ , and the result is plotted in Fig. 7. The mean value of $\Phi_{ee}(o)$ averaged over all $\tau$ is found to be

$$\int_o^1 \Phi_{ee}(o)\, d\tau = 0.5745 \tag{67}$$

It is worth while finding out what reduction of the mean square-error has been made with this optimum filtering system compared to the mean square-error of the Wiener's and Franklin's filters. When only the continuous input is used, the optimum filter is derived by Wiener[1] as

$$H_1 = \frac{1}{X}\left\{\frac{\Phi_{r_1 r}}{\overline{X}}\right\}_L \tag{68}$$

where X and $\overline{X}$ are defined in Eq. 43. Using the given spectral density function, one obtains the optimum filter.

$$H_1 = \frac{0.828}{s + \sqrt{8}} \tag{69}$$

The mean square-error is given by

$$\Phi_{ee}(o) = \frac{1}{2\ j}\int_{-j\infty}^{j\infty}\left[\Phi_{rr} - \Phi_{r_1 r_1}\, H_1\, \overline{H}_1\right] ds$$

$$= \frac{1}{2\pi j}\int_{-j\infty}^{j\infty}\left[\frac{4}{4 - s^2} - \frac{8 - s^2}{4 - s^2}\,\frac{(0.828)^2}{8 - s^2}\right] ds$$

$$= 1 - 0.172 = 0.828 \tag{70}$$

Comparing the mean square-error of the double measurement to that of Wiener filter

$$\frac{0.5745 - 0.828}{0.5745} = -\frac{0.2535}{0.5745} = -44.2\ \% \tag{71}$$

shows that the former system reduces the mean square-error by 44.2 per cent.

When only the sampled input is used the optimum filter is found by Franklin[3] as

$$H_2(s) = \frac{1}{W^*}\left[\frac{\Phi_{rr}}{\overline{W}^*}\right]_L , \tag{72}$$

where $\quad W^*\,\overline{W}^* = \Phi^*_{rr}$ ,

and $W^*$ has all its poles and zeros inside the unit-circle while $\overline{W}^*$ has all its outside. Using the given spectral density function,

$$W^*\,\overline{W}^* = \left[\frac{4}{4 - s^2}\right]^*$$

$$= \frac{0.9817}{1 - 0.135\ z^{-1}}\ \frac{1}{1 - 0.135\ Z} ,$$

$$W^* = \frac{0.9817}{1 - 0.135\ Z^{-1}} ,$$

and $\quad \overline{W}^* = \dfrac{1}{1 - 0.135\ Z}$ .

Hence,

$$H_2 = \frac{1 - 0.135\ Z^{-1}}{s + 2} . \tag{73}$$

The mean square-error is

$$\Phi_{ee}(o) = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty}\left[\Phi_{rr} - \Phi^*_{rr}H_2\overline{H}_2\right] ds$$

$$= \frac{1}{2\pi j}\int_{-j\infty}^{j\infty}\left[\frac{4}{4 - s^2} + \frac{0.9817}{(s+2)(s-2)}\right] ds$$

$$= 1 - 0.245 = 0.755 \tag{74}$$

Comparing the results of Eqs. V-78 and V-89 shows that the double measurement reduces the mean square-error by

$$\left|\frac{0.5745 - 0.755}{0.5745}\right| = 31.4\ \% \tag{75}$$

## Simplified Method of Solution for a Special Case

A general method has been presented for solving the set of simultaneous integral equations of a general double measurement having one sampled and one continuous input. The procedure is quite involved as can be seen from Example I. Complication is encountered especially in determining the unknown coefficient $A_i$ in Eqs. 42 and 46 when substituting both equations into Eq. 28 and comparing the coefficients of the partial fraction terms having like poles. However, a shortcut can be used to determine these coefficients for an important special case which occurs quite often in guidance control.

Frequently, the noise in the sampled input is so small that the input may be regarded as white-noise. Furthermore, the functions $M_2$ and $M_1$, representing the characteristics of the measuring instruments, are often minimum-phase. Under such condition

$$\Phi_{r_2r_2} = \Phi_{rr} M_2 \overline{M}_2$$

$$\Phi_{r_1r_1} = \Phi_{rr} M_1 \overline{M}_1 + \Phi_{n_1n_1}$$

$$\Phi_{n_1n_1} = \nu \qquad\qquad \Phi_{n_2n_2} = o$$

$$\Phi_{r_1r_2} = \Phi_{rr} \overline{M}_1 M_2 \qquad\qquad \Phi_{r_2r_1} = \Phi_{rr} \overline{M}_2 M_1$$

$$\Phi_{r_1r} = \Phi_{rr} \overline{M}_1 \qquad\qquad \Phi_{r_2r} = \Phi_{rr} \overline{M}_2$$

then Eqs. 27 and 28 become

$$\Phi_{r_1r_1} H_1 + \Phi_{rr} \overline{M}_1 M_2 H_2^* e^{-s\tau} = \Phi_{rr} \overline{M}_1 + F_1 \qquad (76)$$

$$[\Phi_{rr} \overline{M}_2 M_1 H_1 e^{s\tau}]^* + [\Phi_{rr} \overline{M}_2 M_2]^* H_2^*$$
$$= [\Phi_{rr} \overline{M}_2 e^{s\tau}]^* + zG_2^* \qquad (77)$$

The solution of $H_2^*$ and $H_1$ may readily be obtained from Eqs. 41 and 46 as

$$H_2^* = \frac{z}{Y^*} \left\{ \frac{z^{-1}}{\overline{Y}^*} [\nu \Phi_{rr} \overline{M}_2 e^{s\tau}]^* \right\}_i$$
$$+ \frac{1}{Y^*} \sum_i \frac{A_i}{1 - \alpha_i z^{-1}} \qquad (78)$$

$$H_1 = \frac{1}{X} \left\{ \frac{1}{\overline{X}} \Phi_{rr} \overline{M}_1 (1 - M_2 e^{-s\tau} H_2^*) \right\}_L \qquad (79)$$

where

$$Y^* \overline{Y}^* = \left[ \frac{\Phi_{rr} M_2 \overline{M}_2}{\Phi_{r_1r_1}} \right]^*$$

$$X \overline{X} = \Phi_{r_1r_1}$$

and $\alpha_i$'s are the poles of $\left[ \frac{\Phi_{rr} \overline{M}_2 M_1}{\Phi_{r_1r_1}} \right]^*$,

which are inside the unit circle.

Instead of substituting Eqs. 78 and 79 back into 77 to determine the unknown constants $A_i$'s, a more convenient method will be explored to find these constants. Multiplying Eq. 76 by $\frac{\overline{M}_2}{\overline{M}_1} e^{s\tau}$ and Z-transforming the whole expression, gives

$$\left[\Phi_{r1r1} \frac{\overline{M}_2}{\overline{M}_1} e^{s\tau} H_1 \right]^* + \left[\Phi_{rr} \overline{M}_2 M_2\right]^* H_2^*$$
$$= \left[\Phi_{rr} \overline{M}_2 e^{s\tau}\right]^* + \left[\frac{\overline{M}_2}{\overline{M}_1} e^{s\tau} F_1\right]^* \qquad (80)$$

Subtracting Eq. 77 from 80

$$\left[\left(\frac{\Phi_{r_1r_1}}{\overline{M}_1} - \Phi_{rr} M_1\right) \overline{M}_2 e^{s\tau} H_1\right]^* = \left[\frac{\overline{M}_2}{\overline{M}_1} e^{s\tau} F_1\right]^* - zG_2^*$$

or, after a simple calculation,

$$\nu \left[\frac{\overline{M}_2}{\overline{M}_1} e^{s\tau} H_1\right]^* = \left[\frac{\overline{M}_2}{\overline{M}_1} e^{s\tau} F_1\right]^* - zG_2^* \qquad (81)$$

In general $H_1$ may be written as

$$H_1 = \frac{N}{D_1 D_2^*} \qquad (82)$$

where $D_1$ is a finite polynomial in s, $D_2^*$ is a finite polynomial in z, and N is a mixed polynomial of both s and z. Substituting Eq. 82 into 81, and multiplying both sides by $D_2^*$

$$\nu \left[\frac{\overline{M}_2}{\overline{M}_1} e^{s\tau} \frac{N}{D_1}\right]^* = \left[\frac{\overline{M}_2}{\overline{M}_1} e^{s\tau} F_1\right]^* D_2^* - zG_2^* D_2^*$$

Since $M_1$ and $M_2$ are minimum phase, the right-hand side of this equation has all its poles outside the unit-circle while the left-hand side has poles both inside and outside the unit-circle. Separating the left-hand side term into two parts, one has only inside poles and the other has only outside poles, this equation becomes

$$\left\{\nu\left[\frac{\bar{M}_2}{\bar{M}_1} e^{s\tau}\frac{N}{D_1}\right]^*\right\}_i = -\left\{\nu\left[\frac{\bar{M}_2}{\bar{M}_1} e^{s\tau}\frac{N}{D_1}\right]^*\right\}_o + D_2^*\left[\frac{\bar{M}_2}{\bar{M}_1} e^{s\tau} F_1\right]^* - D_2^* zG_2^*$$

Since the left-hand side of this equation is analytic outside the unit-circle both must be equal to a constant $K_3$. Thus,

$$\left\{\left[\frac{\bar{M}_2}{\bar{M}_1} e^{s\tau} \frac{N}{D_1}\right]^*\right\}_i = K_3 \qquad (83)$$

In Eq. 83, the quantity $\frac{\bar{M}_2}{\bar{M}_1}$ does not have LHP pole, while the quantity $\frac{N}{D_1}$ has only LHP poles which are the roots of $D_1$. Therefore the poles of $\left\{\left[\frac{\bar{M}_2}{\bar{M}_1} e^{s\tau} H_1\right]^*\right\}_i$ are all due to the poles of $H_1$. To satisfy Eq. 83, it is necessary that $\left[\frac{N}{D_1} e^{s\tau}\right]^*$ has zero residue at its various poles on the z-plane. An equivalent statement describing this condition on the s-plane is the following.

(1) The residue of $\frac{N}{D_1}$ at each real pole or each complex pole whose imaginary com- imaginary component is not equal to $\frac{n\pi}{T}$ must vanish.

(2) The sum of the residues of $\left[\frac{\bar{M}_2}{\bar{M}_1} e^{s\tau}\frac{N}{D_1}\right]^*$ at the complex roots of $D_1$ whose imaginary components are equal to $n\frac{\pi}{T}$ must vanish.

Since $\frac{1}{D_2^*}$ is in general not zero at the roots of $D_1$ the above statement can be put into a more convenient form as follows.

$$\text{(1) Residue}\left[H_1\right] \underset{\substack{\text{real pole of } H_1\text{, or}\\ \text{complex pole of } H_1 \text{ whose}\\ \text{imaginary part} \neq n\frac{\pi}{T}}}{=\!=\!=} 0$$

$$\text{(2) }\sum\text{Residues}\left[\frac{\bar{M}_2}{\bar{M}_1} e^{s\tau} H_1\right] \underset{\substack{\text{complex poles of } H_1\\ \text{whose imaginary part}\\ = n\frac{\pi}{T}}}{=\!=\!=} 0 \qquad (84)$$

Eq. 84 offers very helpful information in determining the unknown constants $A_i$'s of Eqs. 78 and 79. This is done by finding the residues of $H_1$ or $\left[\frac{\bar{M}_2}{\bar{M}_1} e^{s\tau} H_1\right]^*$ at various poles of $H_1$ and setting them equal to zero, as expressed in Eq. 84. Since these residues are function of the unknown constants they can be solved for the constants. In the next section an example is given to illustrate this method.

## Example II

Consider a measuring system shown in Fig. 8 where the sampled channel measures the desired signal directly while the continuous channel measures the rate of change of desired signal. So $M_2 = 1$ and $M_1 = s$. Let the various spectral density functions be

$$\begin{aligned}
\Phi_{rr} &= \frac{4}{(4-s^2)(-s^2)} = \Phi_{r_2r_2} = \Phi_{r_2r} = \Phi_{rr_2}\\
\Phi_{r_1r_1} &= \frac{8-s^2}{4-s^2}\\
\Phi_{n_1n_1} &= 1 \qquad\qquad\qquad\qquad (85)\\
\Phi_{rr_1} &= \Phi_{r_2r_1} = \frac{4}{(4-s^2)(-s)}
\end{aligned}$$

The sampling period T is assumed unity. Eqs. 78 and 79 are used to get the solution of the optimum filter transfer functions $H_2^*$ and $H_1$. First,

$$Y^* \bar{Y}^* = \left[ \frac{\Phi_{rr} M_2 \bar{M}_2}{\Phi_{r_1 r_1}} \right]^*$$

$$= \left[ \frac{-4}{s^2(8-s^2)} \right]^*$$

$$= \frac{0.311(1+0.189z)(1+0.189z^{-1})}{(1-z)(1-z^{-1})(1-0.059z)(1-0.059z^{-1})}$$

Therefore,

$$\left.\begin{aligned} Y^* &= \frac{0.558(1+0.189z^{-1})}{(1-z^{-1})(1-0.059z^{-1})} \\ \bar{Y}^* &= \frac{0.558(1+0.189z)}{(1-z)(1-0.059z)} \end{aligned}\right\} \qquad (86)$$

The quantity

$$\frac{\Phi_{rr} M_2 \bar{M}_1}{\Phi_{r_1 r_1}} = \frac{-4}{s(8-s^2)}$$

has an LHP pole at $s = -\sqrt{8}$, so the corresponding inside-pole of

$$\left[ \frac{\Phi_{rr} \bar{M}_2 M_1}{\Phi_{r_1 r_1}} \right]^* \text{ is at}$$

$$z = \alpha = 0.059. \qquad (87)$$

In Eq. 79 the quantity

$$\left[ \nu \Phi_{rr} \bar{M}_2 e^s \right]^* = \left[ \frac{-4 e^{s\tau}}{s^2(4-s^2)} \right]^*$$

$$= \frac{0.5\, d_1(1+d_2 z)}{(1-z)(1-z^{-1})}$$

$$- \frac{0.01043(b+az)}{(1-0.059z)(1-0.059z^{-1})} \qquad (88)$$

where

$$\left.\begin{aligned} d_1 &= 1-\tau \\ d_2 &= \frac{\tau}{1-\tau} \\ a &= \sinh \sqrt{8}\,\tau \\ b &= \sinh\,(1-\tau)\sqrt{8} \end{aligned}\right\} \qquad (89)$$

Substituting Eqa. 85 through 88 into 78,

$$H_2^* = \frac{z(1-z^{-1})(1-0.059z^{-1})}{0.558(1+0.189z^{-1})}\, X$$

$$\left[ \left\{ \frac{(1-z)(1-0.059z)z^{-1}}{0.558(1=).189z)} \left[ \frac{0.5d_1(1+d_2 z)}{(1-z)(1-z^{-1})} - \frac{0.01043(b+az)}{(1-0.059z)(1-0.059z^{-1})} \right] \right\}_i + \frac{A}{1-0.059z^{-1}} \right]$$

$$= \frac{1.27(1-0.059z^{-1})}{0.558(1+0.189z^{-1})} - \frac{\left[0.0176(b+0.059a)-A\right](1-z^{-1})}{0.558(1+0.189z^{-1})}$$

Let

$$K = 0.0315\ (b+0.059a) + \frac{A}{0.558} \qquad (90)$$

Then

$$H_2^* = \frac{1.27(1-0.59z^{-1}) - K(1-z^{-1})}{1 + 0.189z^{-1}} \qquad (91)$$

Now

$$X \bar{X} = \Phi_{r_1 r_1} = \frac{8-s^2}{4-s^2}$$

so

$$\left.\begin{aligned} X &= \frac{8 + s}{2 + s} \\ \bar{X} &= \frac{8 - s}{2 - s} \end{aligned}\right\} \qquad (92)$$

Substituting Eqs. 91, 92 and $\Phi_{rr}$ into 79,

$$H_1 = \frac{s+2}{s+\sqrt{8}} \left\{ \frac{s-2}{s-\sqrt{8}}\, \frac{-4}{s^2(4-s^2)}\ (-s) \left[ 1-e^{-s\tau} \left( \frac{1.27(1-0.059z^{-1})-K(1-z^{-1})}{1 + 0.189z^{-1}} \right) \right] \right\}_L$$

which, after evaluating the LHP braces,

$$H_1 = \frac{0.293s + 1.414}{s(s + \sqrt{8})} + \frac{(s+2)(0.276K - 0.325)\ e^{\sqrt{8}\tau}}{s^2 - 8}$$

$$- \frac{4e^{-s\tau}}{s(s^2-8)} \left[ \frac{(1-z^{-1})\ K - 1.27\ (1 - 0.059z^{-1})}{1 + 0.189z^{-1}} \right] \qquad (93)$$

The last step is to determine the constant K. This is done by applying to Eq. 93 the residue condition expressed in Eq. 84. Residue of $H_1$ at $s = 0$ is

$$\text{Residue} \left[ H_1 \right]_{s=0} = \frac{1.414}{8} - 0.5 = 0 \qquad (94)$$

and the residue of $H_3$ at $s = -\sqrt{8} = -2.828$ is

$$\text{Residue}\left[H_1\right]_{s=-8} = -0.207 - 0.0477e^{-\sqrt{8}\tau}$$

$$+ (0.949e^{\sqrt{8}\tau} + 0.0405e^{-\sqrt{8}\tau})\,K \qquad (95)$$

Letting Eq. 95 equal to zero, the constant K is found as

$$K = \frac{0.0477e^{-\sqrt{8}\tau} + 0.207}{0.949e^{\sqrt{8}\tau} + 0.0405e^{-\sqrt{8}\tau}} \qquad (96)$$

The final solution of the optimum filters is therefore given by Eqs. 91, 93, and 96. Note that when $\tau=0$

$$H_2^* = 1$$

$$H_1 = 0 \quad .$$

Tables 4, 5, and 6 give the calculated values of K, $H_2^*$, and $H_1$ as functions of their variables. The impulse responses of $h_2(n)$ and $h_1(t)$ are shown in Figs. 9 and 10.

## Conclusions

In general a multiple measurement may consist of sampled-inputs of various sampling rates as well as continuous inputs. A system of this type is referred to as multirate multiple measuring system. It is explored in this paper that a multirate multiple measuring system is time-varying even though its input are stationary.

A double measurement with one continuous-input and one sampled-input has been treated in detail. It is proposed that the set of simultaneous integral equations, which impose the necessary and sufficient condition of the optimum filter, be solved in the frequency domain using the method of undetermined coefficients. The solutions are the transfer functions of the optimum filters. Very often the sampled-input of this double measurement may be considered noise-free. A residue condition is developed to simplify the determination of the unknown coefficients under this situation.

Methods of evaluating mean square-error of the optimum system in frequency domain as well as in time domain are given. Two examples are worked out to illustrate the methods. It is found, in a typical double measurement, the mean square-error is 44.2% lower than that of Wiener's filter and 31.4% lower than that of Franklin's filter.

## Acknowledgments

This paper presents partial result of a research performed at New York University in partial fulfillment of the requirements for the degree of Doctor of Engineering Science. The author wishes to express his sincere thanks to Dr. John R. Ragazzini for his guidance and encouragement, and to Dr. Sheldon S. L. Chang for his comments. The research was supported by the United States Air Force under Contract No. AF 49-(638)-586 monitored by the Office of Scientific Research, Air Research and Development Command, Washington 25, D. C. This support is gratefully acknowledged.

## Bibliography

1. Wiener, N., "Extrapolation, Interpolation, and Smoothing of Stationary Time Series", New York, N. Y., John Wiley and Sons, Inc., 1949.

2. Zadeh, L. A., and Ragazzini, J. R., An Extension of Wiener's Theory of Prediction, Jour. Appl. Phy., 21, pp. 645-655, (1950).

3. Franklin, G. F., Linear Filtering of Sampled-Data, Technical Report T-5/B, Department of Electrical Engineering, Columbia University, New York, N. Y., December, 1954.

4. Lees, A.B., Interpolation and Extrapolation of Sampled Data, Transactions of IRE, IT-2, No. pp. 12-17, (1956).

5. Hsieh, H. C., and Leondes, C. T., On the Optimum Synthesis of Sampled-Data Multipole Filters with Random and Nonrandom Inputs, Transactions of IRE, AC-5, No. 3. pp. 193-208, (1960)

6. Hsieh, H. C., and Leondes, C. T., On the Optimum Synthesis of Multipole Control Systems in the Wiener Sense, IRE National Convention Record, 7, part 4, pp. 18-31, (1959).

7. Bendat, J. S., Optimum Filters for Independent Measurements of Two Related Perturbed Messages, I.R.E. Transactions on Circuit Theory, CT-4, No. 1, pp. 14-19, (1957).

8. Hung, J. C., Theory of Optimum Multiple Measurements, doctoral dessertation, Electrical Engineering Department, New York University, New York, N. Y. 1961.

9. Courant, R., "Differential and Integral Calculus", New York, N.Y., Interscience Publishers, Inc., vol. I, revised ed., 1937; vol. II, 1936.

10. Ragazzini, J.R., and Franklin, G. F., "Sampled-Data Control Systems", Chapter IV, New York, N. Y., McGraw-Hill Book Co., Inc., 1958.

11. Chang, S.S.L., Two Network Theorems for Analytical Determination of Optimum-Response Physical Realizable Network Characteristics, Proc. IRE, 43, pp. 1128-1135, (1955).

12. Cramer, H., "Mathematical Methods of Statistics", Chapter 15, Princeton, New Jersey, Princeton University Press, 1946.

13. Laning, J. H., Jr., and Battin, R. H., "Random Processes in Automatic Control", Chapter 7, New York, N. Y., McGraw-Hill Book Co., Inc., 1956.

| $\tau$ | $h_2(\rho, \tau)$ | |
|---|---|---|
| | $\rho = 0$ | $\rho \neq 0$ |
| 0 | 1.000 | 0 |
| 0.2 | 0.633 | 0 |
| 0.4 | 0.372 | 0 |
| 0.6 | 0.214 | 0 |
| 0.8 | 0.122 | 0 |
| 1.0 | 0.069 | 0 |

$$h_2(\rho,\tau) = \frac{1.175\, e^{-\sqrt{8}\tau}}{1 + 0.175\, e^{-2\sqrt{8}}}\, \delta_{\rho,0}$$

Table 1. Calculated Values of $h_2(\rho,\tau)$, Example I

| $t$ second | $h_1(t,\tau)$ | | | | | |
|---|---|---|---|---|---|---|
| | $\tau=0$ | $\tau=0.2$ | $\tau=0.4$ | $\tau=0.6$ | $\tau=0.8$ | $\tau=1.0$ |
| 0.00 | 0 | 0.531 | 0.733 | 0.796 | 0.818 | 0.825 |
| 0.02 | 0 | 0.479 | 0.683 | 9.753 | 0.776 | 0.784 |
| 0.04 | 0 | 0.42 | 0.638 | 0.71 | 0.729 | 0.748 |
| 0.06 | 0 | 0.363 | 0.591 | 0.663 | 0.687 | 0.695 |
| 0.08 | 0 | 0.311 | 0.549 | 0.623 | 0.650 | 0.657 |
| 0.10 | 0 | 0.256 | 0.507 | 0.586 | 0.612 | 0.621 |
| 0.15 | 0 | 0.127 | 0.408 | 0.498 | 0.527 | 0.538 |
| 0.20 | 0 | 0 | 0.311 | 0.413 | 0.447 | 0.459 |
| 0.30 | 0 | 0 | 0.151 | 0.285 | 0.330 | 0.346 |
| 0.40 | 0 | 0 | 0 | 0.180 | 0.230 | 0.250 |
| 0.50 | 0 | 0 | 0 | 0.090 | 0.170 | 0.190 |
| 0.60 | 0 | 0 | 0 | 0 | 0.100 | 0.140 |
| 0.70 | 0 | 0 | 0 | 0 | 0.050 | 0.100 |
| 0.80 | 0 | 0 | 0 | 0 | 0 | 0.050 |
| 0.90 | 0 | 0 | 0 | 0 | 0 | 0.030 |
| 1.00 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 2. Calculated Values of $h_1(t,\tau)$, Example I

| $\tau$ | $\Phi_{ee}(0)$ | $\Phi_{ee}(0)$ mean |
|---|---|---|
| 0.0 | 0.000 | 0.5745 |
| 0.2 | 0.531 | |
| 0.4 | 0.729 | |
| 0.6 | 0.795 | |
| 0.8 | 0.818 | |
| 1.0 | 0.825 | |

$$\Phi_{ee}(0) = 0.828 - \frac{0.925}{e^{2\sqrt{8}\tau} + 0.175}$$

Table 3. Calculated Values of Mean Square Error $\phi_{ee}(0)$ vs. $\tau$ , Example I

| | $e^{-\sqrt{8}\tau}$ | $e^{-2\sqrt{8}\tau}$ | K |
|---|---|---|---|
| 0.0 | 1.000 | 1.000 | 0.265 |
| 0.2 | 0.568 | 0.323 | 0.139 |
| 0.4 | 0.323 | 0.104 | 0.754 |
| 0.6 | 0.183 | 0.0333 | 0.0415 |
| 0.8 | 0.104 | 0.0107 | 0.0229 |
| 1.0 | 0.059 | 0.0035 | 0.0131 |

$$K = \frac{0.0477\, e^{-\sqrt{8}\tau} + 0.207}{0.949\, e^{\sqrt{8}\tau} + 0.0405^{-\sqrt{8}\tau}}$$

Table 4. Calculated Values of K as Function of Example II

| $t$ | $h_1(t,\tau)$ | | |
|---|---|---|---|
| | $\tau = 0.0$ | $\tau = 0.5$ | $\tau = 1.0$ |
| 0.0 | 0.0 | 0.2065 | 0.27143 |
| 0.5 | 0.0 | 0.1531 | 0.373 |
| 1.0 | 0.0 | -0.03167 | 0.178 |
| 1.5 | 0.0 | -0.0290 | -0.0375 |
| 2.0 | 0.0 | 0.0003 | -0.03653 |
| 2.5 | 0.0 | 0.00015 | 0.01319 |
| 3.0 | 0.0 | 0.0000-- | 0.00395 |

Table 6. Calculated Values of $h_1(t,\tau)$, Example II

| $\tau$ | $H(z,\tau)$ |
|---|---|
| 0.0 | 1 |
| 0.2 | $\frac{1.126+0.064z^{-1}}{1+0.189z^{-1}} = 1.126-0.149z^{-1} + 0.0283z^{-2} - 0.00533z^{-3} + 0.00111z^{-4} - \ldots$ |
| 0.4 | $\frac{1.19+0.0004z^{-1}}{1+0.189z^{-1}} = 1.19 - 0.2246z^{-1} + 0.0425z^{-2} - 0.00803z^{-3} + 0.0015z^{-4} - \ldots$ |
| 0.6 | $\frac{1.224-0.0235z^{-1}}{1+0.189z^{-1}} = 1.224 - 0.255z^{-1} + 0.0482z^{-2} - 0.009z^{-3} + 0.00172z^{-4} - \ldots$ |
| 0.8 | $\frac{1.242-0.052z^{-1}}{1+0.189z^{-1}} = 1.242 - 0.32z^{-1} + 0.0605z^{-2} - 0.0114z^{-3} + 0.00216z^{-4} - \ldots$ |

Table 5. Calculated Values of $H_2(z,\tau)$, Example II

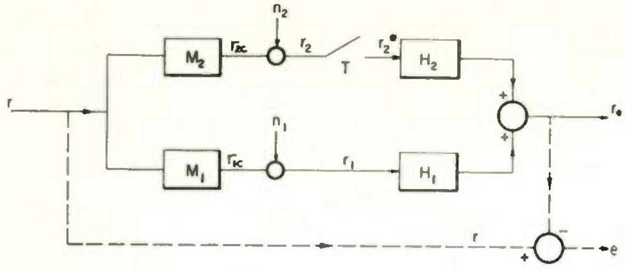


Fig. 1. Double Measurement with Sampled and Continuous Inputs

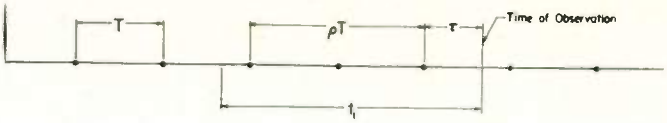


Fig. 3. Definition of Time Variables for System Shown in Fig. 1

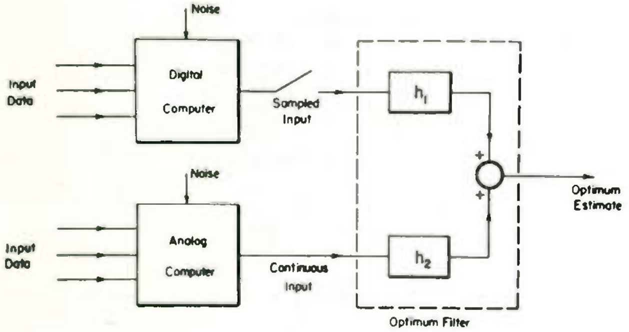


Fig. 2. Parallel Computation Using both Digital and Analog Computers

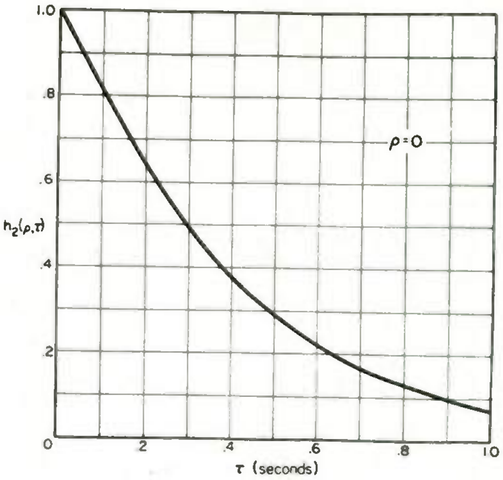


Fig. 4. Impulse Response of $H_2(z,\tau)$, Example I

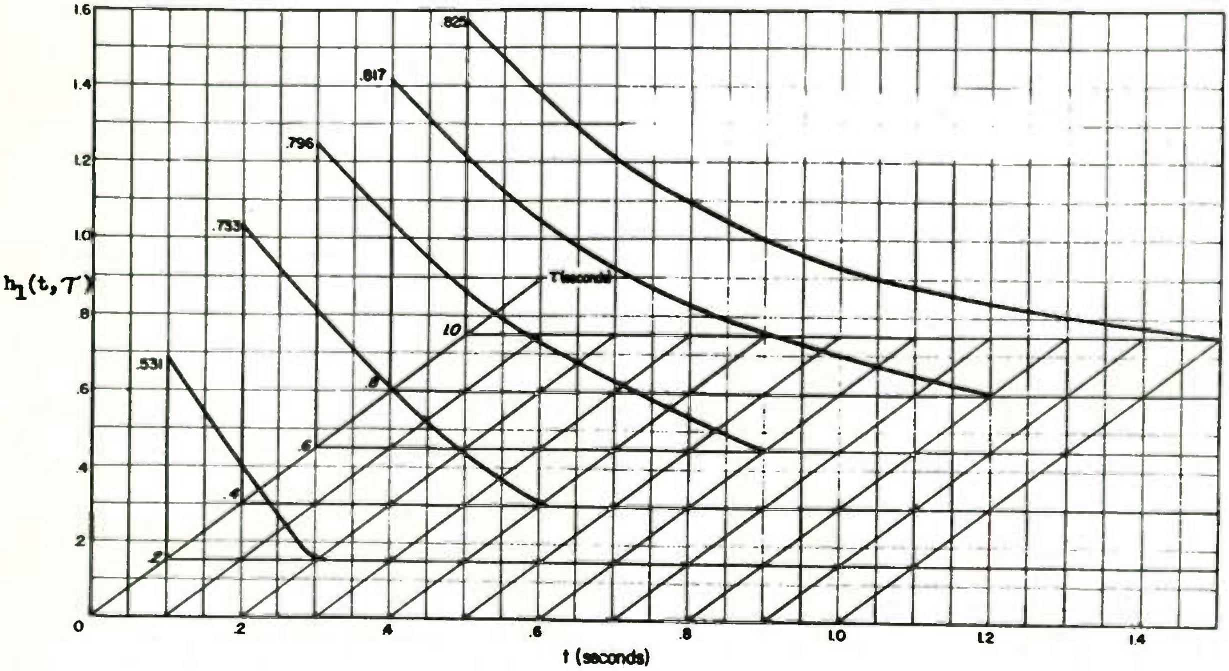
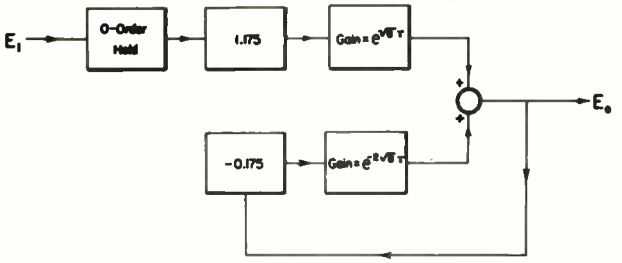


Fig. 5. Impulse Response of $H_1(s,\tau)$, Example I



$$H_2(z,\tau) = \frac{E_o}{E_1} = \frac{1.175e^{-\sqrt{8}\tau}}{1+0.175e^{-2\sqrt{8}\tau}}$$

Fig. 6. Schematic Diagram for $H_2(z,\tau)$, Example I

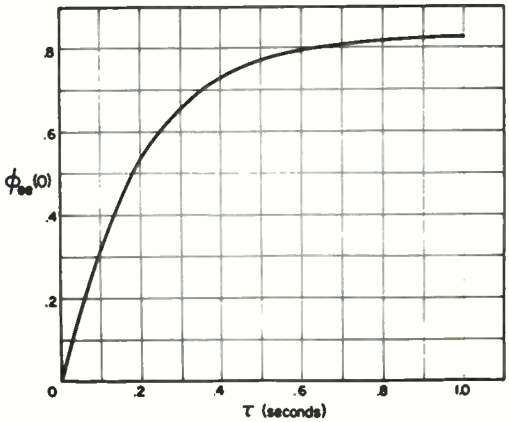


Fig. 7. $\phi_{ee}(0)$ vs. $\tau$ for Example I

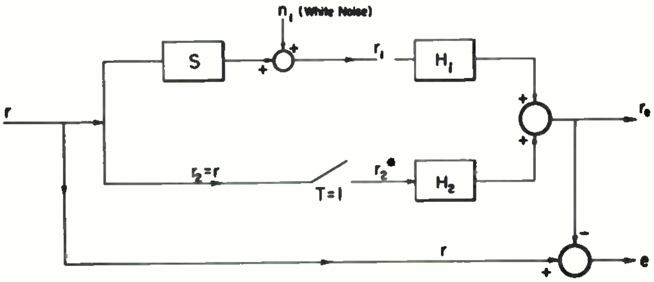


Fig. 8. System of Example II

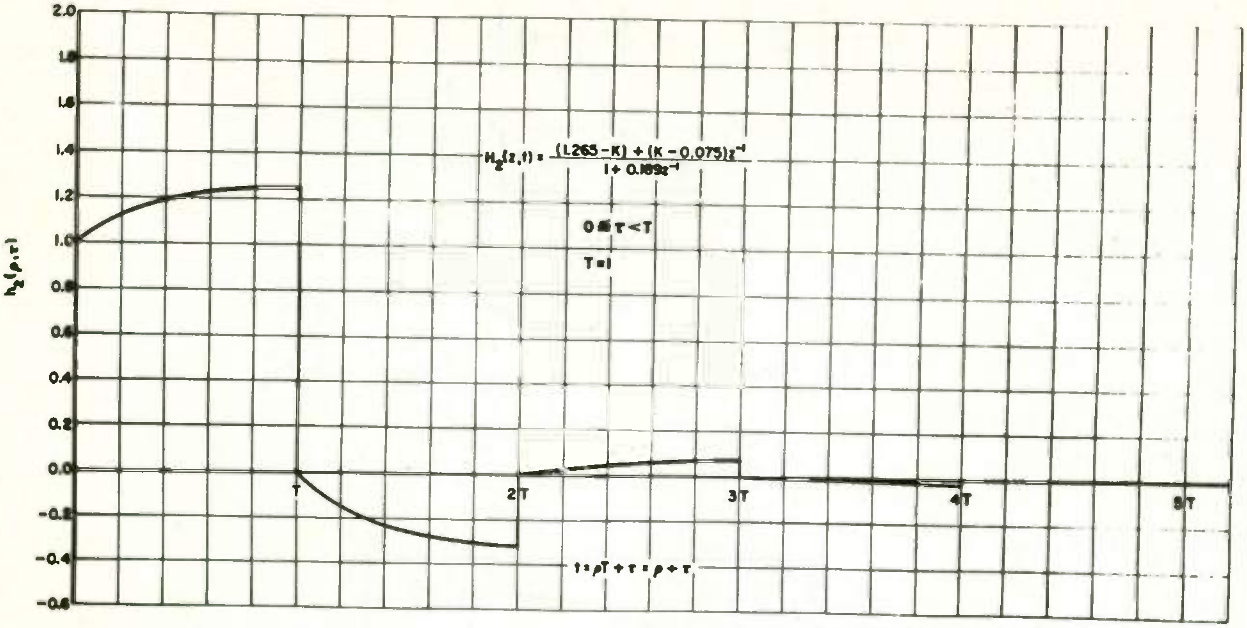


Fig. 9. Impulse Response of $H_2(z,\tau)$, Example II

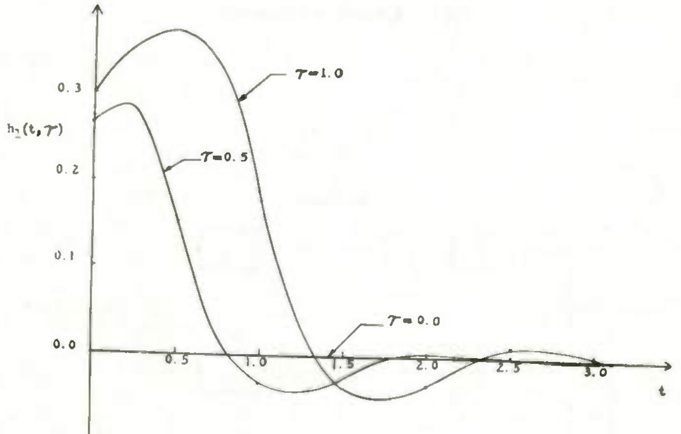


Fig. 10. Impulse Response of $H_1(s,\tau)$, Example II

# MINIMAL TIME CONTROL WITH MULTIPLE SATURATION LIMITS

S. S. L. Chang
New York University
New York 53, New York

## Summary

General rules are proved for minimal time control of a linear system with the constraints that both the manipulated variable $m(t)$ and its derivative $\dot{m}(t)$ are amplitude limited: (1) $\dot{m}(t)$ is always at its extreme value unless $m(t)$ is at its extreme value, (2) the minimal time path is unique and consequently optimum switching boundaries can be defined, and (3) the choice of $\dot{m}(t)$ maximizes a Hamiltonian with a modified adjoint function.

The above rules are applied to third order control systems with decidedly favorable results.

## Introduction

The paper is aimed at removing one essential but impractical condition in the present optimum control theory. In both Pontryagin's maximum principle and the better known "bang-bang" control, the manipulated variable or rudder is assumed to be inertialess.[1-5] Its position can be changed instantly from -a to a. Yet this is never true in actual ships and planes.

The problem can be considered as a special case of a more general problem, that of optimal control in bounded phase space. For instance, in controlling an airplane, the elevator and ailerons are limited in both speed and displacement. One way to remove the multiple limits on the movements of the controls is to consider the velocities of the elevator and ailerons as controls only, and to regard the displacements as phase coordinates together with the other dynamical variables of the airplane. Then the phase coordinates representing the displacements are bounded.

The general problem of optimal control in bounded phase space has been investigated by the writer among others.[6-8] A necessary condition for optimal control was obtained and was also shown to be sufficient under certain conditions. While the result is simple enough, its rigorous proof is quite lengthy and involved.[9]

Using considerably simpler mathematics, the present paper gives an independent proof of a necessary and sufficient condition for minimal time control with multiple saturation limits. The condition is then shown to be identical with the writer's more general result.

In practical terms, the condition means that the minimal time control for a system with multiple saturation limits is a pang-bang system. At all times either the velocity or the displacement of each control is at its maximum value. For the autonomous case optimum switching boundaries are shown to exist and examples are given to illustrate its construction.

## Analytical Preliminaries

### The Problem

The control problem is defined by

$$\dot{x} = Fx + Bm + c \tag{1}$$

$$a_1 \leq m \leq a_2 \qquad b_1 \leq \dot{m} \leq b_2 \tag{2}$$

where $x$ and $c$ are column vectors of $n_1$ dimensions, $m, a_1, a_2, b_1$ and $b_2$ are column vectors of $n_2$ dimensions, and $F$ and $B$ are matrices. The vector relation (2) means that the inequalities hold for each component. The elements of $F, B, a_1, a_2, b_1, b_2$ and $c$ are bounded and continuous functions of time. The vector $x$ is the state vector representing the dynamical state of the system. The vector $m$ is the control vector which can be varied at will within the limitations of (2). It is further assumed that

$$b_1 < \dot{a}_1 < b_2 \qquad b_1 < \dot{a}_2 < b_2 \tag{3}$$

The inequalities (3) assure that the full range of $m$ can be utilized.

The initial condition is represented by $x(0)$ and $m(0)$ at $t = 0$. The terminal condition is given in terms of a vector function $\xi(t)$ of $n_1$ dimensions in two different ways:

The Rendezvous Problem. The function $\xi(t)$ is required to be a possible trajectory which can be traced without using extreme values of $m(t)$ and $\dot{m}(t)$

$$\begin{gathered} \dot{\xi} = F\xi + B\eta + c \\ a_1 < \eta < a_2, \qquad b_1 < \dot{\eta} < b_2 \end{gathered} \tag{4}$$

The problem is to find a $m(t)$ such that

$$x(T) = \xi(T) \tag{5}$$

$$m(T) = \eta(T) \tag{6}$$

for minimum T. Once (5) and (6) are satisfied for some T, it is then possible to make $x(t) = \xi(t)$ for $t \geq T$ by choosing $m(t) = \eta(t)$ for $t \geq T$. The rendezvous of two vehicles is illustrated in Fig. 1a.

The Interception Problem. Let $[x]$ denote a vector made up by $n_1$ or less components of $x$. The problem is to find a $m(t)$ such that

$$[\underset{\sim}{x}]\ (T) = [\underset{\sim}{\xi}]\ (T) \qquad (7)$$

for minimum T. In general $[\underset{\sim}{x}]\ (t) = [\underset{\sim}{\xi}]\ (t)$ cannot be maintained for $t > T$. The interception of one vehicle by another is illustrated in Fig. 1b.

## Reduction of the Problem

For the rendezvous problem, let $\underline{x}(t)$ denote $\underset{\sim}{x}(t) - \underset{\sim}{\xi}(t)$, and $\underline{m}(t)$ denote $\underset{\sim}{m}(t) - \underset{\sim}{\eta}(t)$. Then

$$\underline{\dot{x}} = \underset{\sim}{F}\underline{x} + \underset{\sim}{B}\underline{m} \qquad (8)$$

and

$$\underline{a}_1 \underline{\leq} \underline{m} \underline{\leq} \underline{a}_2 \qquad (9)$$

$$\underline{b}_1 \underline{\leq} \underline{m} \underline{\leq} \underline{b}_2 \qquad (10)$$

where $\underline{a}_1 = \underset{\sim}{a}_1 - \underset{\sim}{\eta}$, $\underline{a}_2 = \underset{\sim}{a}_2 - \underset{\sim}{\eta}$, $\underline{b}_1 = \underset{\sim}{b}_1 - \underset{\sim}{\dot{\eta}}$ and $\underline{b}_2 = \underset{\sim}{b}_2 - \underset{\sim}{\dot{\eta}}$. The problem becomes that of finding $\underline{m}(t)$ so that

$$\underline{x}(T) = 0$$

$$\underline{m}(T) = 0$$

with minimum T.

For the interception problem, a $\underset{\sim}{\eta}(t)$ satisfying (4) may not exist and $\underline{m}(t)$ cannot be defined. Therefore

$$\underline{\dot{x}} = \underset{\sim}{F}\underline{x} + \underset{\sim}{B}\underline{m} + \underline{c}$$

where

$$\underline{c} = \underset{\sim}{c} - \underset{\sim}{\dot{\xi}} + \underset{\sim}{F}\underline{\xi}$$

and the problem is that of finding $\underset{\sim}{m}(t)$ so that $[\underline{x}]\ (T) = 0$ with minimum T.

In the subsequent development, the underlining of the variables will be omitted. It suffices to say that both the rendezvous problem and the interception problem can be reduced to the form of (1) and (2), the inequalities (3) remain valid, and the terminal condition is either

$$\underset{\sim}{x}(T) = 0 \quad \text{and} \quad \underset{\sim}{m}(T) = 0 \qquad (11)$$

or

$$[\underset{\sim}{x}]\ (T) = 0. \qquad (12)$$

## Solution of the Differential Equation by Linear Transform

Consider the homogeneous equation

$$\underset{\sim}{\dot{x}} = \underset{\sim}{Fx} \qquad (13)$$

Let $t_2 \geq t_1$. Due to the linearity of (13), $\underset{\sim}{x}(t_2)$ is related to $\underset{\sim}{x}(t_1)$ by a linear transform $\underset{\sim}{A}(t_2,t_1)$.

$$\underset{\sim}{x}(t_2) = \underset{\sim}{A}(t_2,t_1)\ \underset{\sim}{x}(t_1) \qquad (14)$$

The Dependence of the Function $\underset{\sim}{A}(t_2,t_1)$ on $t_1$ can be Exhibited by Two Conditions: The first condition is obtained by letting $t_1 = t_2$

$$\underset{\sim}{A}(t_2,t_2) = 1 \qquad (15)$$

The second condition is obtained by differentiating (14) with respect to $t_1$:

$$0 = \frac{\partial \underset{\sim}{A}(t_2,t_1)}{\partial t_1}\ \underset{\sim}{x}(t_1) + \underset{\sim}{A}(t_2,t_1)\ \frac{d\underset{\sim}{x}(t_1)}{dt_1}$$

$$= \left[\frac{\partial \underset{\sim}{A}(t_2,t_1)}{\partial t_1} + \underset{\sim}{A}(t_2,t_1)\ \underset{\sim}{F}(t_1)\right] \underset{\sim}{x}(t_1)$$

Since the above equation must be satisfied by arbitrary $\underset{\sim}{x}(t_1)$, it follows that

$$\frac{\partial \underset{\sim}{A}(t_2,t_1)}{\partial t_1} + \underset{\sim}{A}(t_2,t_1)\ \underset{\sim}{F}(t_1) = 0 \qquad (16)$$

The Impulse Response Function of (1) is readily obtainable from $A(t_2,t_1)$: Consider a system initially at rest and an impulse $\underset{\sim}{m}(t)$ is applied at $t_{1-}$:

$$\underset{\sim}{m}(t) = \underset{\sim}{\alpha}\ \delta(t - t_{1-})$$

where $\underset{\sim}{\alpha}$ is a constant vector, and $t_{1-}$ is less than $t_1$ by an infinitesimal quantity, then (1) gives

$$\underset{\sim}{x}(t_1) = \underset{\sim}{B}(t_1)\ \underset{\sim}{\alpha} \qquad (17)$$

and (13) holds for $t > t_1$. From (14) and (17) one obtains:

$$\underset{\sim}{x}(t_2) = \underset{\sim}{A}(t_2,t_1)\ \underset{\sim}{B}(t_1)\ \underset{\sim}{\alpha} \qquad (18)$$

Therefore $\underset{\sim}{A}(t_2,t_1)\ \underset{\sim}{B}(t_1)$ is the response of the system at $t_2$ due to an unit impulse at $t_1$.

The General Solution of (1) is obtained by Superposition: Let $\underset{\sim}{x}(0)$ represent the initial condition of the system at $t = 0$, and $\underset{\sim}{m}(t)$ represent the subsequent input. Then $\underset{\sim}{x}(T)$ at some later instant T can be obtained by adding all the contributions:

$$\underset{\sim}{x}(T) = \underset{\sim}{A}(T,0)\ \underset{\sim}{x}(0) + \int_0^T \underset{\sim}{A}(T,t)\ \underset{\sim}{B}(t)\ \underset{\sim}{m}(t)\ dt$$

$$+ \int_0^T \underset{\sim}{A}(T,t)\ \underset{\sim}{c}(t)\ dt \qquad (19)$$

## The Accessible Region in Enlarged State Space

Let $\underset{\sim}{z}$ be a vector of $n_1 + n_2$ components:

$$z_i = x_i \qquad i = 1, 2 \ldots n_1$$

$$z_{n_1+k} = m_k \qquad k = 1, 2 \ldots n_2$$

The vector $\underset{\sim}{z}$ is the enlarged state vector, and the $n_1 + n_2$ - dimensional $\underset{\sim}{z}$-space is the enlarged state space.

Starting from any given point $\underset{\sim}{z}(0)$, the set of all possible points $\underset{\sim}{z}(T)$ which can be reached at $t = T$ with (1) and (2) satisfied is denoted as $R(T)$. $R(T)$ is the accessible region of the system at $t = T$.

The following Lemmas can be readily proved by treating $\underset{\sim}{\dot{m}}$ as the control vector and $\underset{\sim}{z}$ as the state vector:

Lemma 1. $R(T)$ is convex.

Lemma 2. If a point $\underset{\sim}{\zeta}$ in $\underset{\sim}{z}$-space can be reached at T but not at any time prior to T, then $\underset{\sim}{\zeta}$ is a boundary point of $R(T)$.

## The Rendezvous Problem

For the rendezvous problem, $\underset{\sim}{c} = 0$, and the terminal point is the origin 0. The necessary and sufficient conditions for a control vector $\hat{\underset{\sim}{m}}(t)$ and its resulting path $\hat{\underset{\sim}{z}}(t)$ to be the minimal time control and path pair will be studied:

### Necessary Condition for the Optimal Path

Let $\hat{\underset{\sim}{z}}(t)$ represent a minimal time control which reaches 0 at $t = T$. As $R(T)$ is convex, and 0 is a boundary point of $R(T)$, there is a support plane which passes 0, such that none of the points of $R(T)$ is on the other side of the support plane. Let $\underset{\sim}{h}'$ denote a row vector which is normal to the support plane and points away from $R(T)$. As 0 is on the support plane

$$\underset{\sim}{h}'[\hat{\underset{\sim}{z}}(T) - \underset{\sim}{z}(T)] \geq 0 \tag{20}$$

for any point $\underset{\sim}{z}(T)$ belonging to $R(T)$.

Let $\underset{\sim}{h}_1'$ represent the first $n_1$ components of $\underset{\sim}{h}'$, and $\underset{\sim}{h}_2'$ represent the remaining $n_2$ components of $\underset{\sim}{h}'$. Inequality (20) can be written as

$$\underset{\sim}{h}_1'[\hat{\underset{\sim}{x}}(T) - \underset{\sim}{x}(T)] + \underset{\sim}{h}_2'[\hat{\underset{\sim}{m}}(T) - \underset{\sim}{m}(T)] \geq 0 \tag{21}$$

Let the row vector $\underset{\sim}{\psi}'(t)$ be defined by

$$\underset{\sim}{\psi}'(t) \equiv \underset{\sim}{h}_1' \, \underset{\sim}{A}(T, t) \tag{22}$$

Using (19) and (22), (21) can be written as

$$\int_0^T \underset{\sim}{\psi}'(t)\underset{\sim}{B}(t)[\hat{\underset{\sim}{m}}(t)-\underset{\sim}{m}(t)]\, dt + \underset{\sim}{h}_2'[\hat{\underset{\sim}{m}}(T)-\underset{\sim}{m}(T)] \geq 0 \tag{23}$$

where $\underset{\sim}{m}(t)$ is any other allowed control function.

Let $(\psi'B)_i$ represent the i-th component of the row vector $\underset{\sim}{\psi}'\underset{\sim}{B}$. Written explicitly, (23) becomes

$$\sum_i \left\{ \int_0^T (\psi'B)_i(t)\, [\hat{m}_i(t) - m_i(t)]\, dt + h_{2i}\, [\hat{m}_i(T) - m_i(T)] \right\} \geq 0$$

Since the components $m_i(t)$ can be independently selected, the above inequality must be satisfied for each component:

$$\int_0^T (\psi'B)_i\, (t)\, [\hat{m}_i(t) - m_i(t)]\, dt + h_{2i}[m_i(T) - m_i(T)] \geq 0$$

$$i = 1, 2 \ldots n_2 \tag{24}$$

Inequalities (24) can be used to test the optimality of a given $\hat{m}_i(t)$. The given $\hat{m}_i(t)$ cannot be optimal if (24) is not true with some allowed $m_i(t)$. Thus by choosing different functions for $m_i(t)$, a set of necessary conditions on $\hat{m}_i(t)$ in conjunction with $(\psi'B)_i(t)$ and $h_{2i}$ can be obtained:

Condition 1: $(\psi'B)_i \geq 0$ in any finite interval in which $\hat{m}_i$ is at its upper limit; and $(\psi'B)_i \leq 0$ in any finite interval in which $\hat{m}_i$ is at its lower limit.

Proof: Suppose $\hat{m}_i = a_{2i}(t)$ in an interval $\tau$. Let it be assumed that $(\psi'B)_i < 0$ at $t_1$ in $\tau$. Because $(\psi'B)_i$ is a continuous function of t, there is a finite interval $\tau'$ about $t_1$ in which $(\psi'B)_i < 0$.

Because of (3) it is possible to choose a $\hat{m}_i(t)$ satisfying

$$\hat{m}_i(t) - m_i(t) \quad > 0 \quad \text{in } \tau'$$
$$= 0 \quad \text{elsewhere.}$$

The choice of $m_i(t)$ is illustrated in Fig. 2a. As (24) is contradicted by this particular choice of $m_i(t)$, the assumption that $(\psi'B)_i < 0$ at $t_1$ is not valid.

Condition 2: Let $t_1 < t < t_2$ be an interval in which $m_i(t)$ is not at an extreme value, $t_2 \neq T$, and $\hat{m}_i(t_2)$ is at an extreme value of $m_i$. A function $\varphi_i(t)$ is defined as

$$\varphi_i(t) = \int_t^{t_2} (\psi'B)_i\, (t')\, dt' \tag{25}$$

Then

$$\dot{\hat{m}}_i(t) = b_{2i}(t) \quad \text{if} \quad \varphi_i(t) > 0$$
$$\dot{\hat{m}}_i(t) = b_{1i}(t) \quad \text{if} \quad \varphi_i(t) < 0 \qquad (26)$$

Proof: If $\varphi_i(t) > 0$ but $\hat{m}_i(t) = b_{2i}(t) - \epsilon$ it is then possible to show a contradiction. Let $\dot{m}(t) = b_{2i}(t)$ for an infinitesimal interval $\delta t$, and compensate for the change at $t_2$ as illustrated in Fig. 2b. Inequalities (3) insure the compensation at $t_2$ can always be made. Thus,

$$m_i(t') - m_i(t') = -\epsilon\,\delta t \qquad t < t' < t_2$$
$$= 0 \qquad t' < t, \text{ or } t' > t_2.$$

As $\hat{m}_i(T) = m_i(T)$, only the integral in (24) needs to be evaluated:

$$\int_0^T (\psi'B)_i\,[\hat{m}_i - m_i]\,dt = -\epsilon\,\delta t \cdot \int_t^{t_2} (\psi'B)_i\,dt'$$
$$= -\epsilon\,\delta t\,\varphi_i(t) \le 0$$

Condition 3: If $\hat{m}_i(t_1)$ is at an extreme value of m, then $\varphi_i(t_1) = 0$.

Proof: If $\hat{m}_i(t_1)$ is an extreme value, $\dot{\hat{m}}(t_1)$ is not. It is then possible to choose a $\dot{m}_i(t)$ which differs from $\hat{m}_i(t)$ by $\mp\epsilon$ for an interval $\delta t$ near $t_1$, and by $\pm\epsilon$ for an interval $\delta t$ near $t_2$ as illustrated in Fig. 2c. Then

$$\hat{m}_i(t) - m_i(t) = \pm\epsilon\,\delta t \qquad t_1 < t < t_2$$
$$= 0 \qquad t < t_1,\ t > t_2$$

$$\int_0^T (\psi'B)_i\,[\hat{m}_i - m_i]\,dt = \pm\epsilon\,\delta t \cdot \varphi_i(t_1)$$

The only possible way of satisfying (24) is

$$\varphi_i(t_1) = 0 \qquad (27)$$

In general, $\varphi_i(0) \neq 0$.

Condition 4: In the final interval in which $\hat{m}(t)$ is not at an extreme value, $t_2 = T$, and $\varphi_i(t)$ can be redefined as

$$\varphi_i(t) = \int_t^T (\psi'B)_i\,dt + h_{2i}$$

Then (26) and (27) remain valid.

Proof: Consider any change in $\dot{m}$ for an infinitesimal interval $t - \delta t$ to $t$

$$\dot{m}(t) = \dot{\hat{m}}(t) + \epsilon$$

Then $\hat{m}(t') - m(t') = -\epsilon\,\delta t$ for all $t' \geq t$ as is illustrated in Fig. 2d.

$$\int_0^T (\psi'B)_i\,(\hat{m}_i - m_i)\,dt + h_{2i}\,[\hat{m}_i(T) - m_i(T)]$$
$$= -\epsilon\,\varphi_i(t)\,\delta t$$

The remaining part of the proof is the same as before (conditions 2 and 3).

The above conditions are valid for every component of $\hat{m}(t)$. Since $\dot{m}_i$ is limited, $\hat{m}_i$ is a continuous function of t. The interval 0 to T can be divided for each component i into bang intervals in which $\hat{m}_i$ is at an extreme value and pang intervals in which $\hat{m}_i$ is not. As a result of conditions 2 and 4, $\dot{\hat{m}}_i$ is then at an extreme value. In the bang intervals $\hat{m}_i(t)$ maximizes $(\psi'B)_i\,m_i$. In the pang intervals, $\dot{\hat{m}}_i$ maximizes $\varphi_i \dot{m}_i$. However as $\varphi_i = 0$ in the bang intervals, $\dot{\hat{m}}_i$ can be said to maximize $\varphi_i \dot{m}_i$ at all times. The inequality

$$\varphi_i\,\dot{\hat{m}}_i \geq \varphi_i\,\dot{m}_i$$

is true for all values of t.

The vector $\psi'$ satisfies the matrix differential equation

$$\frac{d\psi'}{dt} + \psi' F = 0 \qquad (28)$$

Eq. (28) follows from (16) since $\psi'(t) = h'A(T,t)$ by definition. The function $\varphi_i(t)$ is continuous in t and satisfies the following conditions:

$$\varphi_i(t) = 0 \qquad \text{in a bang interval}$$
$$\varphi_i + (\psi'B)_i = 0 \text{ in a pang interval} \qquad (29)$$

Let $\varphi'$ be the row vector whose components are $\varphi_i$. In any given problem, the vectors $\psi'$ and $\varphi'$ are unknown as $h'$ is usually unknown. The result of the present section can be summarized as follows:

A necessary condition for $\hat{m}(t)$ to be a minimal time control function is that there exist continuous vector functions $\psi'(t)$ and $\varphi'(t)$ satisfying (28) and (29) such that

$$(\psi'B)_i\,\hat{m}_i \geq (\psi'B)_i\,m_i \text{ in bang intervals}$$
$$\varphi_i\,\dot{\hat{m}}_i \geq \varphi_i\,\dot{m}_i \quad \text{for all } t \qquad (30a)$$
$$i = 1, 2 \ldots n_2 \qquad (30b)$$

## Uniqueness of the Solution

A system is called "normal" if none of the components of $\psi'B$ can be zero over a finite interval unless $\psi'(t) = 0$ for all t. As $\psi'$ is required to satisfy (28), the condition of normality is a condition on the matrices F and B. For example, if (1) describes two independent systems, then some components of $\psi' B$ can be zero while others are not. A more detailed discussion of this condition is given in Lasalle's paper.[3]

For a normal system, if a control vector $\hat{m}(t)$ causes the enlarged state vector $z$ to move from its initial value to the origin in time T, and $\psi'$ and $\varphi'$ exist such that (28), (29), and (30) are satisfied by the set $\hat{m}$, $\psi'$ and $\varphi'$, then it is not possible for any control vector to move $z$ to the origin in time less than T, and $\hat{m}$ is the only one which does the job in T.

To prove the above assertion, assume that there is a control vector $m(t)$ which moves the enlarged state vector $z$ to the origin in time T', T' < T. Let $\dot{m} = 0$ for the duration $T' < t \le T$. Then $m = 0$ and $x = 0$ for the same duration.

Let $\hat{x}$ and $x$ denote the path resulting from $\hat{m}$ and $m$ respectively. Then

$$\frac{d}{dt}\left[\psi'(\hat{x} - x)\right]$$

$$= -\psi' F(\hat{x} - x) + \psi' F(\hat{x} - x) + \psi' B(\hat{m} - m)$$

$$= \psi' B(\hat{m} - m)$$

Since $\hat{x}(0) = x(0)$, and $\hat{x}(T) = x(T) = 0$ the integral of the left hand side of the above equation vanishes and

$$0 = \int_0^T \psi' B(\hat{m} - m)\, dt = \sum_i \int_0^T (\psi' B)_i\, (\hat{m}_i - m_i)\, dt$$

$$= \sum_i \left\{ \sum_j \int_{\tau_{ij}} (\psi' B)_i\, (\hat{m}_i - m_i)\, dt + \sum_k \int_{\tau'_{ik}} (\psi' B)_i\, (\hat{m}_i - m_i)\, dt \right\} \qquad (31)$$

where $\tau_{ij}$ represents the bang intervals, and $\tau'_{ik}$ represents the pang intervals of $\hat{m}_i$. In the pang intervals

$$\int_{\tau'_{ik}} (\psi' B)_i\, (\hat{m}_i - m_i)\, dt = -\int_{\tau'_{ik}} \dot{\varphi}_i\, (\hat{m}_i - m_i)\, dt$$

$$= -\varphi_i(\hat{m}_i - m_i)\Big]_{\tau'_{ik}} + \int_{\tau'_{ik}} \varphi_i(\dot{\hat{m}}_i - \dot{m}_i)\, dt$$

An interval $\tau'_{ik}$ ends either as $\hat{m}_i$ reaches an extreme value, in which case $\varphi_i = 0$, or at t = 0, and t = T, in which case $\hat{m}_i - m_i = 0$. Therefore

$$\varphi_i(\hat{m}_i - m_i)\Big]_{\tau'_{ik}} = 0$$

and (3) becomes

$$0 = \sum_i \left\{ \sum_j \int_{\tau_{ij}} (\psi' B)_i\, (\hat{m}_i - m_i)\, dt + \sum_k \int_{\tau'_{ik}} \varphi_i(\dot{\hat{m}}_i - \dot{m}_i)\, dt \right\} \qquad (32)$$

Referring to (30) the only possibility for (32) to hold is $m_i = \hat{m}_i$ in $\tau_{ij}$, and $\dot{m}_i = m_i$ in $\tau'_{ik}$. Thus there is only one way of reaching $(m, x) = 0$ in T or sooner and the assertion is proved.

## The Interception Problem

For the interception problem, the terminal state is not a single point $z = 0$ in the enlarged state space, but is a hyperplane defined by $[x] = 0$. At some T, R(T) touches the hyperplane, and the tangent point represents the terminal state of minimal time control. Following the same steps as before, the same necessary conditions can be proved with additional boundary conditions on $\psi'$ and $\varphi'$:

$$\psi_i(T) = 0 \qquad \text{for all components } i \text{ with } x_i \text{ not in } [x] \qquad (33)$$

$$\varphi'(T) = 0$$

These conditions replace the boundary conditions on $x$ and $m$, as $x_i(T)$ and $m(T)$ are now unknown.

But given a set of functions $\hat{m}$, $\psi'$, $\varphi'$ satisfying all the necessary conditions (28), (29), (30), and (33), it cannot be proved that no other $m(t)$ causes the system to reach a point on $[x] = 0$ at some earlier time. The only uniqueness condition one can prove is that no other $m(t)$ causes the system to reach the same $x(T)$ at T.

## The Hamiltonian Formulation

Let the $n_1 + n_2$ - dimensional square matrix $G$ be defined by

$$G = \begin{pmatrix} F & B \\ 0 & 0 \end{pmatrix} \qquad (34)$$

Let the $(n_1 + n_2) \times n_2$ matrix $K$ be defined by

$$K = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \qquad (35)$$

where $\underset{\sim}{0}$ is $n_1 \times n_2$ dimensional, and $\underset{\sim}{1}$ is $n_2 \times n_2$ dimensional. Then (1) can be written as

$$\dot{\underset{\sim}{z}} = \underset{\sim}{G}\,\underset{\sim}{z} + \underset{\sim}{K}\,\dot{\underset{\sim}{m}} \qquad (36)$$

The inequalities (2) become bounds on the state variables $z_{n_1+1}, z_{n_1+2} \cdots z_{n_2}$ and control vector $\dot{m}$ and define allowed regions Z and U in $\underset{\sim}{z}$ and $\dot{\underset{\sim}{m}}$ spaces. Let the $n_1 + n_2$ dimensional adjoint function be denoted $\underset{\sim}{\chi}'$, then $\underset{\sim}{\chi}'$ is a row vector:

$$\underset{\sim}{\chi}' = (\underset{\sim}{\psi}', \underset{\sim}{\varphi}') \qquad (37)$$

From the general theory,[6] $\underset{\sim}{\chi}'$ satisfies

$$\dot{\underset{\sim}{\chi}}' + \underset{\sim}{\chi}' G = \zeta(t)\,\underset{\sim}{\eta}' \qquad (38)$$

where $\eta'$ is a row vector of $n_1 + n_2$ dimensions perpendicular to a support plane at $\underset{\sim}{z}$ pointing outward if $\underset{\sim}{z}$ is a boundary point of Z, and $\eta' = 0$ if $\underset{\sim}{z}$ is an interior point of Z; and $\zeta(t) \geq 0$.

A necessary and sufficient condition for $\hat{\dot{\underset{\sim}{m}}}(t)$ to be optimal is that a function $\underset{\sim}{\chi}'$ satisfying (38) can be found such that the choice of $\hat{\dot{\underset{\sim}{m}}}$ maximizes the Hamiltonian

$$H = \underset{\sim}{\chi}'(\underset{\sim}{G}\,\underset{\sim}{z} + \underset{\sim}{K}\,\dot{\underset{\sim}{m}}) = \underset{\sim}{\chi}'\,\underset{\sim}{G}\,\underset{\sim}{z} + \underset{\sim}{\varphi}'\,\dot{\underset{\sim}{m}}$$

Since $\underset{\sim}{\chi}'\,\underset{\sim}{G}\,\underset{\sim}{z}$ is independent of $\dot{\underset{\sim}{m}}$, maximizing H is equivalent to (30b).

Because $z_1, z_2 \ldots z_{n_1}$ are unbounded, the first $n_1$ components of $\underset{\sim}{\eta}'$ are always zero. Eq. (38) gives

$$\dot{\underset{\sim}{\psi}}' + \underset{\sim}{\psi}'\underset{\sim}{F} = 0 \qquad (39)$$

$$\dot{\underset{\sim}{\varphi}}' + \underset{\sim}{\psi}'\,\underset{\sim}{B} = \zeta(t)[\underset{\sim}{\eta}'] \qquad (40)$$

Eq. (39) is identical with (28). Since $\eta_{n_1+i} = 0$ in a pang interval of $m_i$, (29) is satisfied in a pang interval. Since $\zeta(t) \geq 0$, $\varphi_i(t) = 0$ is a solution of (40) in a bang interval if and only if (30a) is true. Thus the equivalence between the general solution and the present special solution is established.

## Examples

A system with multiple saturation limits is illustrated in Fig. 3. The integrator and two nonlinear blocks represent that both $\dot{m}$ and m are amplitude limited: $|m| \leq a$, $|\dot{m}| \leq b$. The plant is assumed to be linear and time-invariant (autonomous).

As there is a unique minimal time trajectory from each point in the enlarged state space, the optimum value of $\hat{\dot{m}}$ is defined at each point. Furthermore, the optimum $\dot{m}$ can only be ± b in the interior of the state space, and 0, ± b at the boundaries where $|m| = \pm a$. The state space can be divided into regions according to the sign of $\dot{m}$. The dividing boundaries are the optimum switching surfaces and curves.

### Example 1.

$$G(s) = \frac{1}{s^2}$$

This example has been worked out by Doll and Stout.[10] Their result is in complete agreement with the present theory.

### Example 2.

$$G(s) = \frac{0.25}{s(s + .5)}$$

$$a = b = 1$$

In terms of normalized coordinates $x_1$, $x_2$, and $x_3$, the controlled system as specified above can be written as

$$\dot{x}_1 = x_3$$

$$\dot{x}_2 = -\,0.5\,x_2 + 0.5\,\dot{m}(t)$$

$$\dot{x}_3 = \dot{m}(t)$$

$$e(t) = c(t) - r(t) = 0.5\,x_1 + 2x_2 - x_3$$

$$|x_3(t)| \leq 1$$

$$|\dot{m}(t)| \leq 1$$

The optimum boundaries are obtained by tracing back in time from the origin. Fig. 4 shows the projections of the optimum boundaries on the $x_1 - x_2$ plane. The heavy curves ABC and DEF are the upper and lower edges of the optimum boundary on the two planes $x_3 = +1$ and $-1$ respectively. BO and EO give the final switching boundaries which are also extremal paths leading to the origin. The boundary surface below EB is formed by extremal paths $\dot{m} = 1$ originating from points on FE and ending on AB or BO. The boundary surface above EB is formed by extremal paths ($\dot{m} = -1$) originating from points on CB and ending on DE or EO.

Fig. 5 gives the responses c(t), $\dot{m}(t)$, and m(t) of the optimum nonlinear system to a unit step input at t = 0. In contrast, the response c(t) of the same controlled system with a linear controller is shown in a broken curve. The linear controller is specified by

$$M(s) = \frac{(s + .05)(s + .5)}{(s + 1)(s + 1.5)}\,[R(s) - C(s)]$$

It is to be noted that the comparison is made at an input amplitude most favorable to the linear system as it is just about saturating. At a lower input amplitude, the responce of the linear system does not change in shape, but the response of the optimum system becomes even faster. At a higher input amplitude, the linear system saturates and its response deteriorates rapidly. The response of the optimum system becomes slower but does not change in character.

## Acknowledgment

The work upon which this paper was based was sponsored by the Office of Scientific Research, Air Research and Development Command, Washington 25, D. C., under Contract No. AF 49(638)-586.

## References

1. McDonald, D.C. "Nonlinear Techniques for Improving Servo Performance." Proc. National Electronics Conference, vol. 6 (1950), pp. 400-21.

2. Fluglotz, I. "Discontinuous Automatic Control." Princeton University Press, Princeton, New Jersey, (1953).

3. LaSalle, J. P. "Time Optimal Control Systems." Contributions to the Theory of Nonlinear Oscillations, vol. 5, Annals of Mathematical Study, Princeton University, Princeton, New Jersey (1960), and Proc. Nat. Acad. Sci., vol. 45, no. 4 (1959), pp. 573-77.

4. Boltyanskii, V. G., Gamkrelidze, R. V., Pontryagin, L. S. "The Theory of Optimal Processes I, Maximum Principle." Izvestia Akad. Nauk SSSR, Ser. Mat., vol. 24, no. 1 (1960) translated by L. W. Neustadt, Space Technology Laboratories, Report no. 9810.32-01 (October 1960).

5. For a textbook reference of these subjects see Chang, S.S.L., Synthesis of Optimum Control Systems. New York: McGraw Hill Co., 1961.

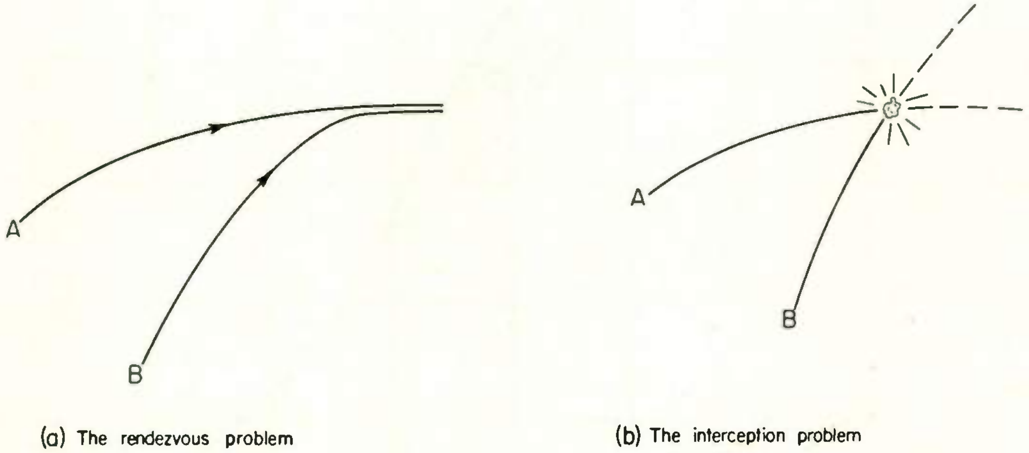
6. Chang, S.S.L. "Optimal Control in Bounded Phase Space." AFOSR Report No. 1238 (1961). Also to be published in the first issue of Automatica.

7. Gamkrelidze, R.V. "Optimal Control Processes with Restricted Phase Coordinates." Izvestia Akad. Nauk SSSR, Ser. Mat. vol. 24 (1960), pp. 315-56.

8. Dreyfus, Stuart. "Variational Problems with Inequality Constraints." Report P2357 (1961) Rand Corporation, Santa Monica, California.

9. Chang, S.S.L. "An Extension of Ascoli's Theorem and its Applications to the Theory of Optimal Control." AFOSR Report 1973, January 1962.

10. Doll, H. G. and T. M. Stout. "Design and Analog-Computer Analysis of an Optimum Third Order Nonlinear Servomechanism." Trans. A.S.M.E., (1957), pp. 513-25.



(a) The rendezvous problem

(b) The interception problem

Fig. 1. Two types of minimal-time control problems.

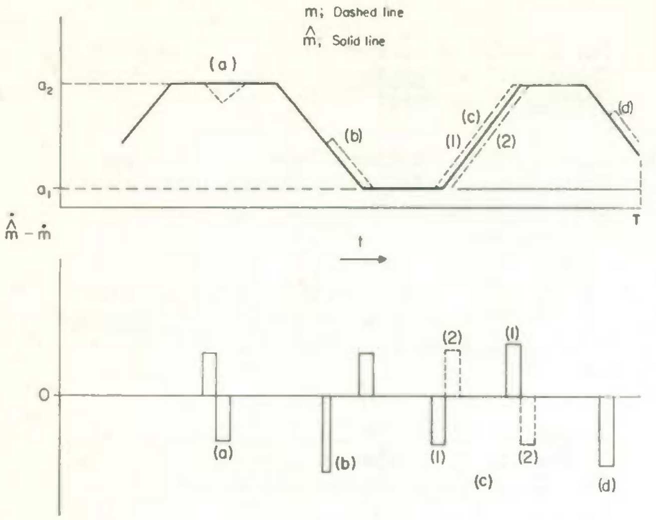


Fig. 2. Four basic types of allowed displacements from $\hat{m}$.

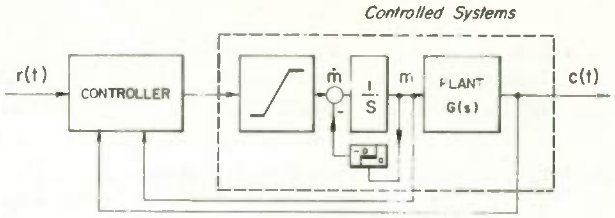


Fig. 3. Block diagram of an autonomous system with multiple saturation limits.

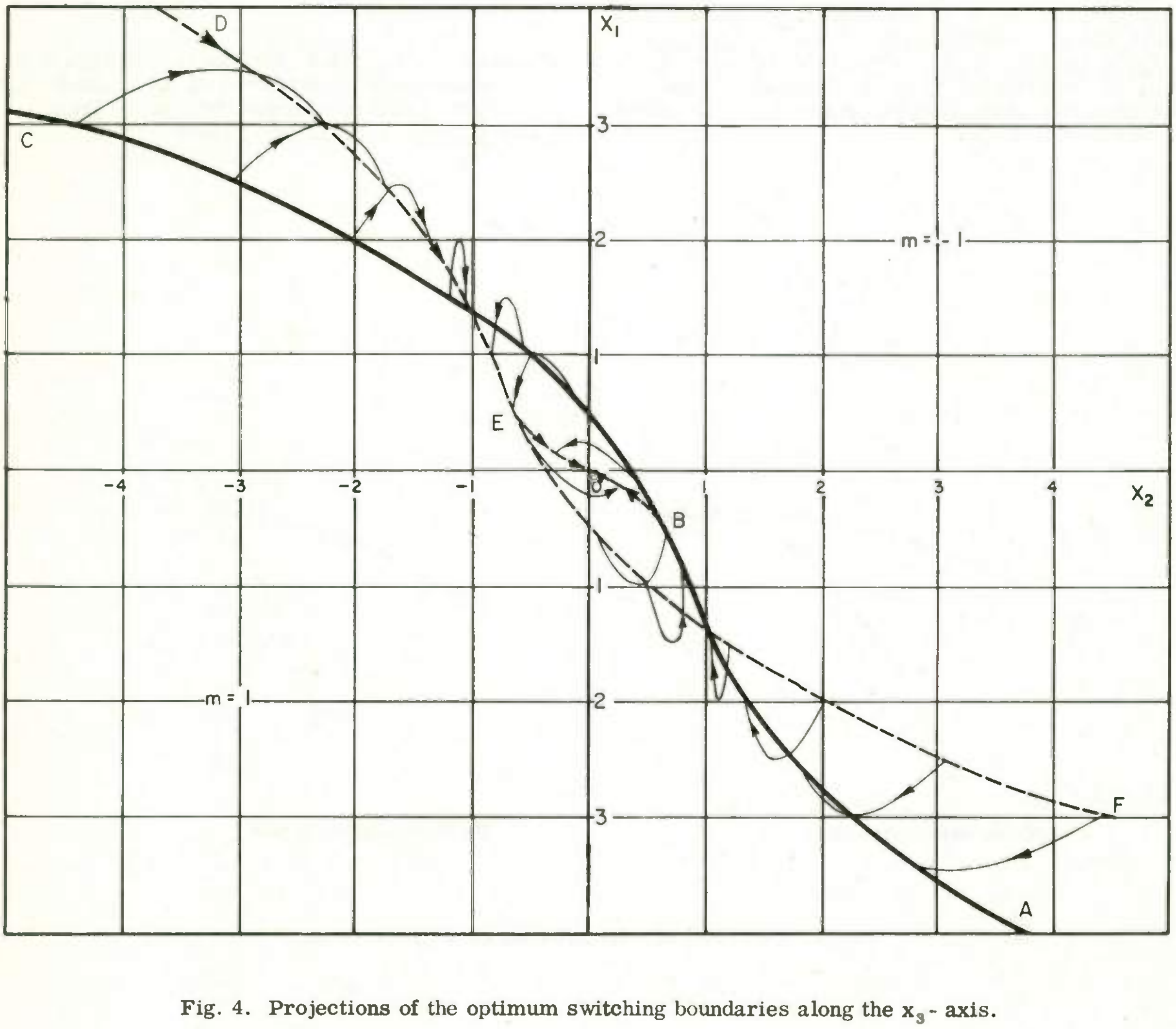


Fig. 4. Projections of the optimum switching boundaries along the $x_3$ - axis.

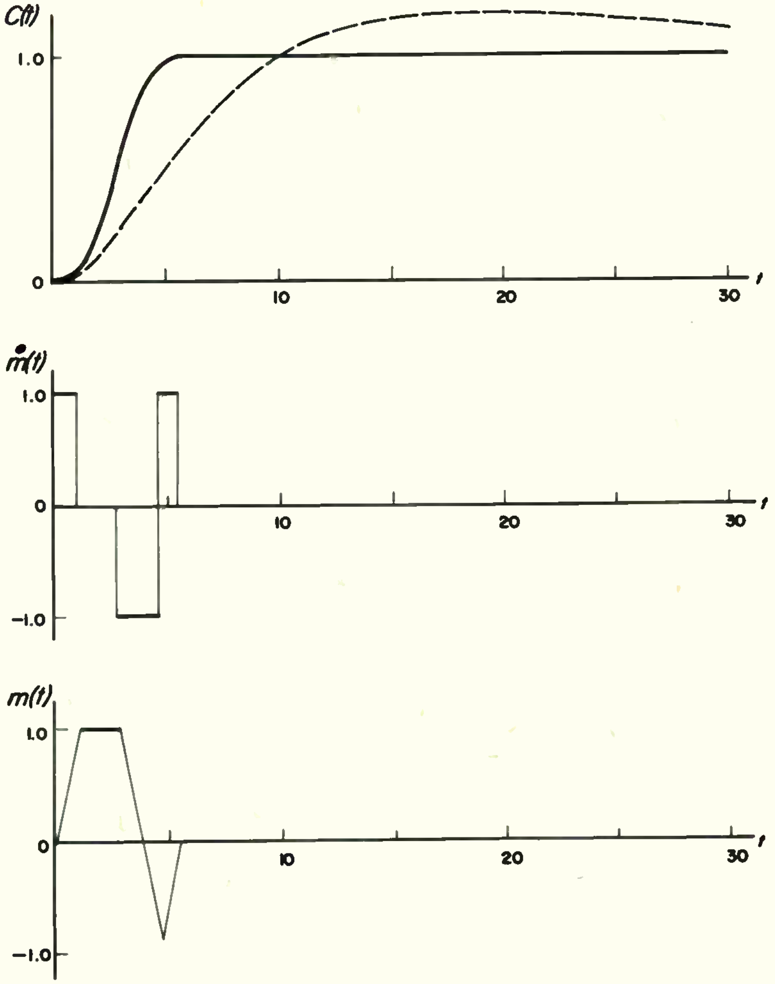


Fig. 5. Unit step responses of two types of control systems.

| | |
|---|---|
| solid curve: | optimum controller |
| broken curve: | linear controller |

# RADAR TRACKING UTILIZING OPERATIONAL DYNAMIC REGENERATION

Stephen Adelman and Stanley M. Shinners
Surface Armament Division
Sperry Gyroscope Company
Division of Sperry Rand Corporation
Great Neck, L.I., N.Y.

## Introduction

The increasing complexity of modern warfare has resulted in the need for highly sophisticated techniques for tracking targets in a countermeasures environment. This paper describes the application of a technique known as Operational Dynamic Regeneration (ODR) that aids the tracking phase of a tracking radar by predicting future target position. The degree of improvement it effects in enabling the smooth and accurate tracking of a target is illustrated.

There are three major sequential phases of operation of a tactical tracking radar. These are designation, acquisition, and tracking. In the designate phase of operation, the radar is programed to the general location of the target within a certain accuracy that depends upon the available information. In the acquisition phase, the radar beam is scanned around the area in a preprogramed scan pattern. When the beam crosses a target, the radar stops moving in the preprogramed pattern and commences tracking. This transition may be either manual or automatic.

In a realistic environment, tactical tracking radars, have difficulty in continuously tracking targets, because of fades caused by natural phenomenonor man-made jamming interferences. However, with tracking radars used for instrumentation, fades caused by natural phenomenon and high target dynamics are the primary concern.[1] There is a need for techniques that will accurately predict future position for reasonable intervals of time, so that redesignation and reacquisition by the radar will be unnecessary. The ODR system described accomplishes this by means of regenerative techniques that continuously generate angular and range tracking signals for the radar from prior knowledge of target position, velocity, and acceleration.

Before deriving a mathematical model for any aided-tracking scheme, however, it is necessary to decide what conditions are most likely to exist in the anticipated environment. In addition, the equipment should operate in a set of coordinates that require the least amount of equipment for practical implementation.

It is questionable whether continuous range tracking can take place in an actual countermeasures environment. Therefore, for a tactical system, a realistic design to aid angle tracking should not depend on range tracking. In addition, it is assumed that the target is flying a straight-line path and is maintaining a constant velocity. The last two assumptions are necessary in order to minimize the complexity of the equipment. If the target were to deviate from a straight-line path, or change its velocity, it would be necessary for the operator to introduce new information into the aided tracking equipment. However, should range tracking be possible over positions of the target's flight path, future range position could also be predicted by means of ODR.

Discontinuities in the predicted target paths are inherent in many regenerative tracking systems for certain target positions and maneuvers; it is desirable to eliminate these. In addition it is desirable for an aided-tracking scheme to produce information that is fairly straightforward and easily obtainable. ODR, a regenerative system that meets these requirements, has two main functions. Its first function is to perform aided angle tracking for a target flying a straight-line path with a constant velocity. This portion of the system assumes that range tracking has never taken place. Should range tracking be possible over a portion of the target's path, the second function of ODR is to predict future range position by solving the target-trajectory equation for a constant-velocity target.

To develop an aided-tracking system for an instrumentation tracking radar, a different environment exists. In all probability, range and range rate will be known accurately for most of the target's path. Therefore, an aided-tracking system, in angle, could depend on range tracking. A system that could aid tracking for such a situation also will be illustrated in this paper. The ODR system will be able to predict future target position by assuming that the target is flying a straight-line path and maintaining a constant velocity. Range, range-rate, and angular-velocity information will be utilized in the regenerative device. As will be illustrated later, the instrumentation required to predict future angular position is considerably reduced when range and range-rate information are available.

## Precision-Tracking Considerations

The ODR system is applicable to precision, tracking radars used for instrumentation and tactical missions with both land-based and shipboard installations. This paper analyzes ODR in a set of coordinates used by shipboard radars, since this is generally a more difficult and interesting case. In addition, both a tactical and an instrumentation tracking radar will be considered for the shipboard case.

One of the most critical problems associated with the design of a naval tracking radar is stabilization against ship-motion dynamics with a load that exhibits severe resonance effects in an environment of external torque loading due to winds. In order to achieve acceptable performance, a stable line-of-sight independent of such factors as ship motion, structural resonances, and external torque loading must be established. The magnitude of this problem is increased greatly because of the frequency characteristics of ship-motion dynamics and nonlinear characteristics of mechanical resonances. [2]

Ship-motion frequencies are usually in the same portion of the frequency spectrum as target dynamics. However, an optimum tracking loop requires that the target dynamics be separated from the ship-motion dynamics. A simple method of achieving this, without introducing additional errors, is to place a minor gyro feedback loop around the external input. Then, the tracking loop need compensate only for target dynamics, while the minor gyro feedback loop compensates for ship-motion dynamics and external torques. A general configuration illustrating this construction is shown in Fig. 1. The minor and major feedback loops are called the stabilization loop and track loop, respectively.

## ODR Model for a Tactical Tracking Radar.

To derive a model for regenerative tracking, it is necessary first to consider the properties of a constant-velocity target flying in a straight-line path as viewed from the radar-antenna coordinates. In addition, it is necessary to consider that the radar, which is stabilized about the line-of-sight, tracks a target in two stabilized planes: elevation and traverse. Figure 2 illustrates the system coordinates.

The position of the target relative to the ship is defined by the vector $\bar{R} = \bar{R}r$ where R is the magnitude of the range from the ship to the target, and $\bar{r}$ is a unit vector directed along the line-of-sight to the target. The rate of change of $\bar{R}$ in space, $d\bar{R}/dt$, is given by eq. (1):

$$\frac{d\bar{R}}{dt} = \frac{d(\bar{r}R)}{dt} = \bar{r}\frac{dR}{dt} + R\frac{d\bar{r}}{dt} \qquad (1)$$

The target's acceleration is given by eq. (2):

$$\frac{d^2(\bar{r}R)}{dt^2} = \bar{r}\frac{d^2R}{dt^2} + 2\frac{dR}{dt}\frac{d\bar{r}}{dt} + R\frac{d^2\bar{r}}{dt^2} \qquad (2)$$

The total angular rate of rotation of the target about the line-of-sight is defined by eq (3):

$$\bar{\omega} = \omega_t\bar{t} + \omega_r\bar{r} + \omega_e\bar{e} \qquad (3)$$

where

$\bar{t}$ = unit vector in a plane perpendicular to r and is directed along the traverse axis

$\bar{e}$ = unit vector perpendicular to both $\bar{r}$ and $\bar{t}$ and is directed along the elevation axis

$\omega_t$ = traverse angular rate

$\omega_r$ = range angular rate

$\omega_e$ = elevation angular rate

By means of eq (2) and (3), it can be shown that the total acceleration of $\bar{r}R$ in space may also be written as shown in eq (4):

$$\frac{d^2(\bar{r}R)}{dt^2} = \bar{r}\left[\ddot{R} - R(\omega_t^2 + \omega_e^2)\right]$$
$$+\bar{t}\left[-R\dot{\omega}_e - 2\dot{R}\omega_e + R\omega_r\omega_t\right] \qquad (4)$$
$$+\bar{e}\left[R\dot{\omega}_t + 2\dot{R}\omega_t + R\omega_r\omega_e\right]$$

Since the target is assumed to move at a uniform velocity, the acceleration of $\bar{r}R$ is zero. The components of $d^2(\bar{r}R)/dt^2$ projected onto a set of mutually orthogonal axes must also be zero and are defined by eq (5), (6), and (7):

$$\left.\frac{d^2(\bar{r}R)}{dt^2}\right|_{\bar{r}} = \ddot{R} - R(\omega_t^2 + \omega_e^2) = 0 \qquad (5)$$

$$\left.\frac{d^2(\bar{r}R)}{dt^2}\right|_{\bar{t}} = -R\dot{\omega}_e - 2\dot{R}\omega_e + R\omega_r\omega_t = 0 \qquad (6)$$

$$\left.\frac{d^2(\bar{r}R)}{dt^2}\right|_{\bar{e}} = R\dot{\omega}_t + 2\dot{R}\omega_t + R\omega_r\omega_e = 0 \qquad (7)$$

The ratio of range-rate to range, $\dot{R}/R$, may be written as shown in eq (8) from the relations given in eq. (6) and (7):

$$\frac{\dot{R}}{R} = -\frac{1}{2}\frac{\dot{\rho}}{\rho} \tag{8}$$

The angular rate of rotation of the target in a plane perpendicular to the line-of-sight ($\rho$) is defined by eq.(9):

$$\rho^2 = \omega_e^2 + \omega_t^2 \tag{9}$$

The ratio of range-rate to range also may be shown to be equivalent to that given by eq.(10). This is obtained from eq. (5) and (8):

$$\frac{\dot{R}}{R} = -\frac{(2\rho^3 + \ddot{\rho})}{3\dot{\rho}} \tag{10}$$

If eq.(8) and (10) are combined to eliminate the dependency of the tactical tracking radar upon range[3] and range rate, the following results:

$$-\frac{1}{2}\frac{\dot{\rho}}{\rho} = -\frac{(2\rho^3 + \ddot{\rho})}{3\dot{\rho}}$$

If this expression is solved for $\ddot{\rho}$, eq.(11) results:

$$\ddot{\rho} = \frac{3\dot{\rho}^2}{2\rho} - 2\rho^3 \tag{11}$$

Equation (11) describes the angular motion of a constant-velocity target flying a straight-path course. By solving this equation, the ODR system may generate angle-tracking signals for a tactical radar, independent of range and range rate.

The angular rate $\rho$, as defined previously, is greater than zero except in the unusual case where both $\omega_e$ and $\omega_t$ are simultaneously zero. Therefore, the inverse of $\rho$ ordinarily will be finite, a very desirable feature in generating eq.(11). If the orthogonal components of $\rho$ were generated separately, a singularity of this kind would be more likely to occur. In addition, the instrumentation required to generate separately the components of $\rho$ would be considerably more complex than that required to generate $\rho$ from eq.(11).

A method for ODR in range is to assume a constant-speed target; specifically, the value of target speed prior to a target fade or jamming. Before the loss of range track, the speed will be computed by means of (12) and its value would be held fixed. From kinematics, the motion of a particle in space is given by eq.(12):

$$V = \sqrt{(\dot{R})^2 + (\rho R)^2} \tag{12}$$

where

$\dot{R}$ = target range rate

$\rho$ = angular rate of rotation of the target in a plane perpendicular to the line-of-sight

$R$ = target range

With this implementation, regardless of any target maneuvers (as long as it maintained a constant velocity), the solution would generate the true range of the target.

Should the radar be usable to normally track in range, the predicted value of range rate, $R_c$, could be computed by means of eq.(13).

$$\dot{R}_c = \sqrt{V_o^2 - (\rho R_c)^2} \tag{13}$$

where

$V_o$ = value of target speed when normal range track is lost

$R_c = R_o + \int_{t_o}^{t} \dot{R}_c \, dt$ = predicted value of target range

$R_o$ = value of target range when normal range tracking is lost

$t_o$ = time when normal range tracking is lost

It is important to note that there is an ambiguity with these computations when the target maneuvers to a 90-degree crossing. Beyond these points it is impossible to determine by this technique whether the target has maneuvered inward or outward (whether $\dot{R}_c$ goes negative or positive from zero). To eliminate this problem without the need for performing any additional computations, it probably could be assumed that the target range rate always reverses direction after crossing at 90 degrees to the target line-of-sight.

## ODR for an Instrumentation Tracking Radar

When deriving a model of regenerative tracking radar, it is reasonable to assume that range tracking is possible over most of the target path. Therefore, it can be assumed that ODR will have range and range-rate information available. The instrumentation required to predict the future position of a target flying a straight path at a constant velocity is greatly simplified when this information is available in addition to angular velocity.

A regenerative tracking model can be derived for this case starting with eq.(6) and (7). Solutions for $\dot{\omega}_e$ and $\dot{\omega}_t$ are obtained from (6) and (7), respectively. The results are illustrated in (14) and (15):

$$\dot{\omega}_e = -2\frac{\dot{R}}{R}\omega_e + \omega_r\omega_t \qquad (14)$$

$$\dot{\omega}_t = -2\frac{\dot{R}}{R}\omega_t - \omega_r\omega_e \qquad (15)$$

It is the instrumentation of these simple equations that makes ODR possible for this case. By calculating the terms $\omega_e$ and $\dot{\omega}_t$ it is possible to cause the radar to predict future angular positions for a target flying a straight-line path at a constant velocity.

The same ODR technique used for range prediction, illustrated for the case of the tactical tracking radar, can also be used for the instrumentation, tracking radar case. Equations (12) and (13) define the required instrumentation. As a matter of fact, regenerative range tracking is a necessity for this situation since eq.(14) and (15), the regenerative equations for angle tracking, depend on range and range rate. Should range track be lost for a small portion of the target's trajectory, predicted values of range and range rate, $R_c$ and $\dot{R}_c$, would be used in eq.(14) and (15).

## Implementation for Tactical Tracking Radar

During the normal angle-tracking mode of operation, either automatic or manual, the radar transmits angular traverse and elevation rates of the line-of-sight axis to ODR. These traverse and elevation angular rates, $\omega_t$ and $\omega_e$ are transformed to the stabilized horizontal and vertical angular rates $\rho_h$ and $\rho_v$ by a rotation of coordinates through $\omega_r$. These horizontal and vertical angular rates are then vectorially summed to yield the total angular rate of the target line-of-sight in the boresight plane $\rho$ , where the angle $\rho$ = arc tan $\rho_h/\rho_v$. Before $\rho$ and $\dot{\rho}$ can be fed to the function generator for storage, it is advisable to filter these quantities. A block diagram illustrating the operation of the system under these conditions is shown in Fig. 3.

When normal angle tracking ceases, the system is switched into the ODR mode. The solution to the equation of constant-velocity motion, eq.(11), is generated by the system shown in Fig. 4. The values of $\dot{\rho}$ and $\rho$, which are stored in ODR when this mode of operation commences operation, are the initial conditions required to solve equation 11. The angle $\theta$ is held fixed at the last value established prior to entering the ODR mode. The angular rate $\rho$, which is computed in the function generator during this mode, is resolved through the angle $\theta$ into the horizontal and vertical components of the angular rate $\rho_h$ and $\rho_v$. These stable horizontal and vertical angular rates are transformed into the deck-oriented lateral and vertical angular rates $\rho_t$ and $\rho_e$ by a rotation of the coordinate system through the angle $\omega_r$. These deck-oriented angular rates are introduced as rate commands into the corresponding traverse and elevation stabilization loops. The radar then moves in accordance with these ordered rates.

The angular rates provided by ODR may be modified at the discretion of the radar operator to compensate for changes in target velocity and direction. Manual controls incorporated into these circuits could allow the operator to modify $\rho_t$ and $\rho_e$.

During the normal range track mode of operation, traverse angular rate ($\omega_t$), elevation angular rate ($\omega_e$), target range (R), and target range rate ($\dot{R}$) are received as inputs to the ODR system from the radar. The angular rates $\omega_t$ and $\omega_e$ are added vectorially to obtain the combined angular rate $\rho$. Target speed V is then computed as indicated by eq.(12) from the vector sum of $\dot{R}$ and the product of R with $\rho$. Target speed and range are stored for use in the ODR mode. A block diagram illustrating the operation of the system under these conditions is shown in Fig. 5.

If range tracking is lost, ODR is used to solve eq. (13) for the predicted value of target range rate $\dot{R}_c$. Angular rates $\omega_t$ and $\omega_e$ continue to be sent from the radar just as in the normal mode of operation. In order to continuously predict target range ($R_o + R_c$) when normal range tracking is lost, the integral of the predicted value of the target range rate is integrated and combined with the value of target range known at the time range tracking was interrupted. Figure 6 illustrates the operation of the system.

## Implementation for Instrumentation Tracking Radar

During the normal angle-tracking mode of operation, either automatic or manual, the radar transmits range R, range rate $\dot{R}$, range angular rate $\omega_r$, traverse angular rate $\omega_t$, and elevation angular rate $\omega_e$ to the ODR system, which generates, and stores $\omega_e$ and $\omega_t$ . It is advisable to filter R, $\omega_r$, $\omega_r$, and $\omega_e$ before feeding it to the function generator. A block diagram illustrating the operation of the system under these conditions is shown in Fig. 7.

When normal angle tracking ceases, the system is switched into the ODR mode. The solution to the equations of constant-velocity motion, eq.(14) and (15), is generated by the system shown in Fig. 8. The deck-oriented angular rates $\omega_t$ and $\omega_e$ are introduced as rate commands into the corresponding traverse-and elevation-stabilization loops.

The radar then moves in accordance with these commands.

The angular rates provided by ODR may be modified at the discretion of the radar operator in case of changes in target velocity or direction. Manual controls incorporated in these circuits could allow the operator to modify $\omega_t$ and $\omega/e$.

The same ODR implementation used for range prediction, illustrated for the case of the tactical tracking radar, can also be used for the instrument tracking radar case. Figures 5 and 6 illustrate the operation of this system.

## Possible Tracking Modes Utilizing ODR

ODR is a very powerful tool for either automatic or manual tracking. For automatic tracking, ODR updating is utilized only during intervals of desired or inadequate tracking-signal information. For the case where the tracking loop is closed manually via the intelligence of a well-trained operator, ODR eases the task considerably. For example, all the operator needs to do when manually tracking a target flying a straight-line path at a constant velocity is to initially adjust his tracking controls until he establishes the proper tracking rates. Once he has manually locked onto the target, ODR will predict the future target position automatically, and the operator need not have to continually adjust his controls to track the target. However, it may be necessary to update the ODR information should the target change its velocity or direction. This situation easily can be displayed to the operator for corrective action, via a bipolar video display (A scope) or an error display (F scope), or both.

Manual tracking offers considerable improvement in the signal-to-noise ratio required for tracking a target.[4] This desirable feature lends itself readily to the tactical tracking radar case, where the normal tactical environment is very noisy. Considerable literature has been written on the effective human transfer function. [5, 6, 7, 8]. An operator may be called upon to function in several different ways, depending upon the design of the radar. He may function as a simple lag element, an equivalent differentiator, a single or double integrator, or a sampled-data system. If it were possible to know every minute detail of the operator's characteristics, the differential equation for a specific control situation would be different for different kinds of inputs such as periodic or random functions. In addition, the differential equation of the operator would have variable coefficients, since his characteristics change according to his learning, motivation, fatigue, and instructions. Due to the complexity of the problem, tracking lags and errors have been unavoidable in a manual-tracking mode of operation. A simple, manual, rate-aided tracking system is shown in Fig. 9, where the operator is considered to be a nonlinear, sampled-data system.

ODR techniques can compensate effectively for some of the lags and errors in manual tracking. Because of ODR prediction capabilities, the task of the operator is greatly minimized; his primary role simply is to update the ODR system according to observed changes in target velocity or direction. This is not a very critical operation, and nonperiodic lags and errors can be tolerated. Therefore, manual tracking in conjunction with ODR represents a very powerful method for tracking in a countermeasures environment with relatively low signal-to-noise ratios and with most of the tracking lags and errors compensated to a very large degree.

## Evaluation Criteria for ODR

The tracking error, $E_1(s)$, as defined in Fig. 1, can be described as an analytic function of complex frequency, s, in the simple case of a linear tracking loop. In the case of ODR, the tracking error becomes much more complicated. At best, the error can be described in terms of statistical properties of possible target dynamics. For these properties, an expectation of the error can be determined and a realistic evaluation or ODR can proceed. But to proceed with a realistic system evaluation of ODR, a suitable error criterion must first be chosen to measure system performance. Several useful criteria have been suggested in the literature. One very useful technique for expressing system error is known as the integral-square-error criteria, $E_{IS}$, defined by eq.(16):

$$E_{IS} = \int_0^\infty [e(t)]^2 \, dt \qquad (16)$$

where

$$e(t) = r(t) - e(t)$$

A modified $E_{IS}$ can be defined by using a suitable weighting function, W(t):

$$E_{IS_m} = \int_0^\infty W(t) [e(t)]^2 \, dt \qquad (17)$$

W(t) is chosen to bring into prominence the error in that portion of time primarily of interest. If the error is considered in the static time sense; then:

$$W(t) = 0 \quad , \quad 0 \le t \le T \qquad (18)$$

Thus, in essence, the error before a time T is not of interest, for only the steady-state value is of importance. This criteria is valid, with a small modification, for ODR. If the error becomes extremely large, the target may be lost entirely. The criteria is subsequently modified such that:

$$W(t) = 0 \quad , \qquad 0 \leq t \leq T$$

provided that $e(t) \leq A$, where A is the critical break-track point.

In order to determine the characteristics of the system itself, one can use the method of Wiener and consider the effect of a target moving in a Brownian fashion.[9] Therefore:

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} r(t) r(t+\tau) dt = A \delta (0) \qquad (19)$$

where A is the spectral power density and $\delta$ (0) is a dirac delta function at t=0. Assuming stationarity, or at least short-time stationary behavior, the dynamic characteristics of the system are completely describable in terms of the coefficients of a set of orthogonal functions. Wiener describes the general system in terms of Laguerre and Hermite spectra. However, any set of orthogonal functions can be normalized and used to characterize the system. The expression for this characterization is as follows:

$$r(t) = \sum_{n=1}^{N} C_n \Phi_n (t) \qquad (20)$$

where the $\Phi_n$'s represent a set of orthogonal functions, and the $C_n$'s are their corresponding coefficients.

In addition, the following constraints are imposed to insure orthogonality and the proper normalization of the spectra in the region defined by a and b:

$$\int_a^b \Phi_n^2 (t) dt \triangleq 1 \qquad (21)$$

$$\int_a^b \Phi_n (t) \Phi_m (t) dt = 0 \qquad n \neq m \qquad (22)$$

It also can be shown that an optimum match in the integral-square sence can be achieved in the region of interest defined by the weighting function W(t) when:

$$C_n = \int_a^b \Phi_m(t) r(t) W(t) dt \qquad (23)$$

It is now possible to evaluate $C_n$ from the behavior to the Brownian input, and therefore to evaluate the response to any individual component of the spectra. From this it becomes a relatively simple matter to evaluate the response to the various components of the input and to thereby obtain a measurement of system accuracy and performance. The procedure for evaluating the $C_n$'s for a set of orthogonal Laguerre and Hermite polynomials is described in detail by Wiener.[8]

To proceed with the evaluation of a typical ODR system, it can be assumed that the system excitation has a known statistical distribution. From this input, using a Wiener type of Laguerre and Hermite spectrum analyzer, the coefficients of the output spectra may be analyzed readily for the particular system and parameters in question. From a direct comparison of the input and output spectra coefficients, the error may be simply evaluated. This technique is primarily an analysis, rather than a synthetic method, and the designer must have considerable insight into the fundamentals of the problem in order to simplify the derivation of useful results.

## Conclusion

A unique approach to the design of a class of radar tracking loops has been described. ODR models for tactical and instrumentation tracking radars have been derived. The characteristics of ODR have been discussed and their implementation for the continuous tracking of a target is illustrated. Considered are cases where the dynamic characteristics are suitably described in three coordinates, and where certain information in one coordinate is missing or inadequate. The minimization of the error in the predicted tracking data is considered in the Wiener sense, from a mean-square point of view, and from a static-time point of view.

ODR is a very powerful tool that can enhance the smooth and continuous automatic and manual tracking capabilities of a tracking radar. The additional cost and complexity entailed are relatively small compared to the over-all cost of a modern, precision, tracking radar. Above all, the resultant improvement in tracking performance that can be derived from this technique will easily pay for itself by reducing the number of times missile tracking or guidance is lost due to normal target fades and countermeasures.

## Acknowledgment

The authors wish to thank the many members of the engineering staff of the Sperry Gyroscope Company who contributed to the concept discussed in this paper. Special acknowledgement is due Messrs. R. Belluck, C.F. Chubb, R.C. Colbert, and D.P. Meyer, for the invaluable aid in the development of some of the concepts discussed in this paper. In addition, the authors wish to extend their thanks and appreciation to Messrs. A. Cincotta and P.C. Miller for helping to prepare and edit this paper.

## References

1. S. Adelman and S.M. Shinners, "Automatic Tracking Considerations for Ballistic Targets," presented at the 5th National Convention on Military Electronics, Washington, D.C., June 26-28, 1961.

2. S.M. Shinners, "Minimizing Servo Load Resonance Errors," CONTROL ENGINEERING, January 1962.

3. Unpublished Notes of R. Belluck and D.P. Meyer of the Sperry Gyroscope Company.

4. L.V. Blake, "Recent Advancements in Radar Range Calculation Technique," IRE TRANSACTIONS ON MILITARY ELECTRONICS, April 1961.

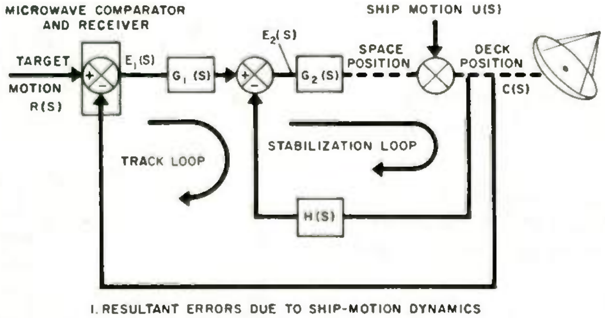
5. A. Tustin, "The Nature of the Operator's Response in Manual Control and its Implications for Controller Design," JOURNAL OF THE INSTITUTION OF ELECTRICAL ENGINEERS (London), part IIA, 1947.

6. J.R. Ragazzini, "Engineering Aspects of the Human Being as a Servomechanism," presented at the American Psychological Association annual meeting, 1948.

7. R. Mayne, "Some Engineering Aspects of the Mechanism of Body Control", ELECTRICAL ENGINEERING, March 1951.

8. N.D. Diamantides, "Man as a Link in a Control Loop, "ELECTRO-TECHNOLOGY, January 1962.

9. N. Wiener, NONLINEAR PROBLEMS IN RANDOM THEORY, The Technology Press (M.I.T.), Cambridge, 1958.



1. RESULTANT ERRORS DUE TO SHIP-MOTION DYNAMICS

(a) IN THE STABILIZATION LOOP:

$$E_2(S)=U(S)\frac{H(S)}{1+G_2(S)H(S)} \text{ FOR } G(S)H(S)\gg 1, E_2(S)=U(S)\frac{1}{G_2(S)}$$

(b) IN THE TRACK LOOP:

$$E_1(S)=E_2(S)\frac{1}{1+G_1(S)}$$

2. RESULTANT ERRORS DUE TO TARGET DYNAMICS;

ASSUME THAT $G_2(S)H(S)\gg 1$:

$$E_1(S)=R(S)\frac{1}{1+G_1(S)/H(S)}$$

FIG. 1 CONFIGURATION FOR SEPARATING TARGET DYNAMICS FROM SHIP-MOTION DYNAMICS

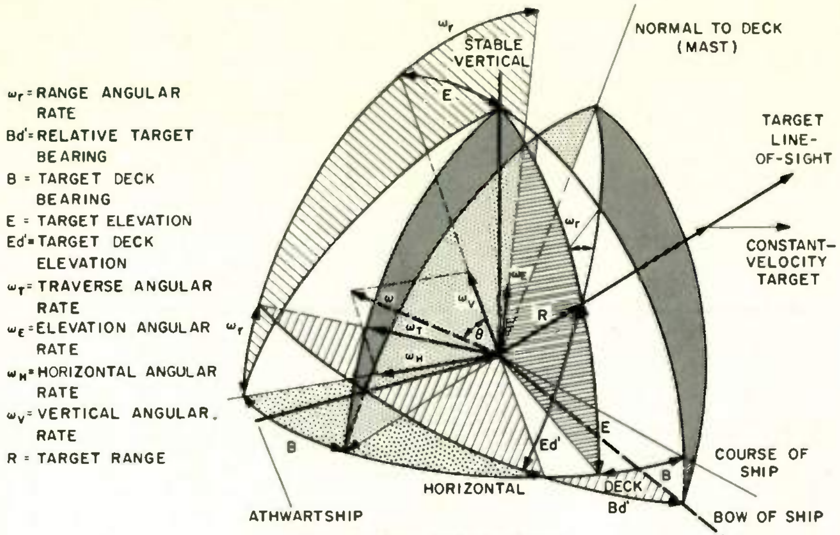


FIG. 2 SYSTEM COORDINATES

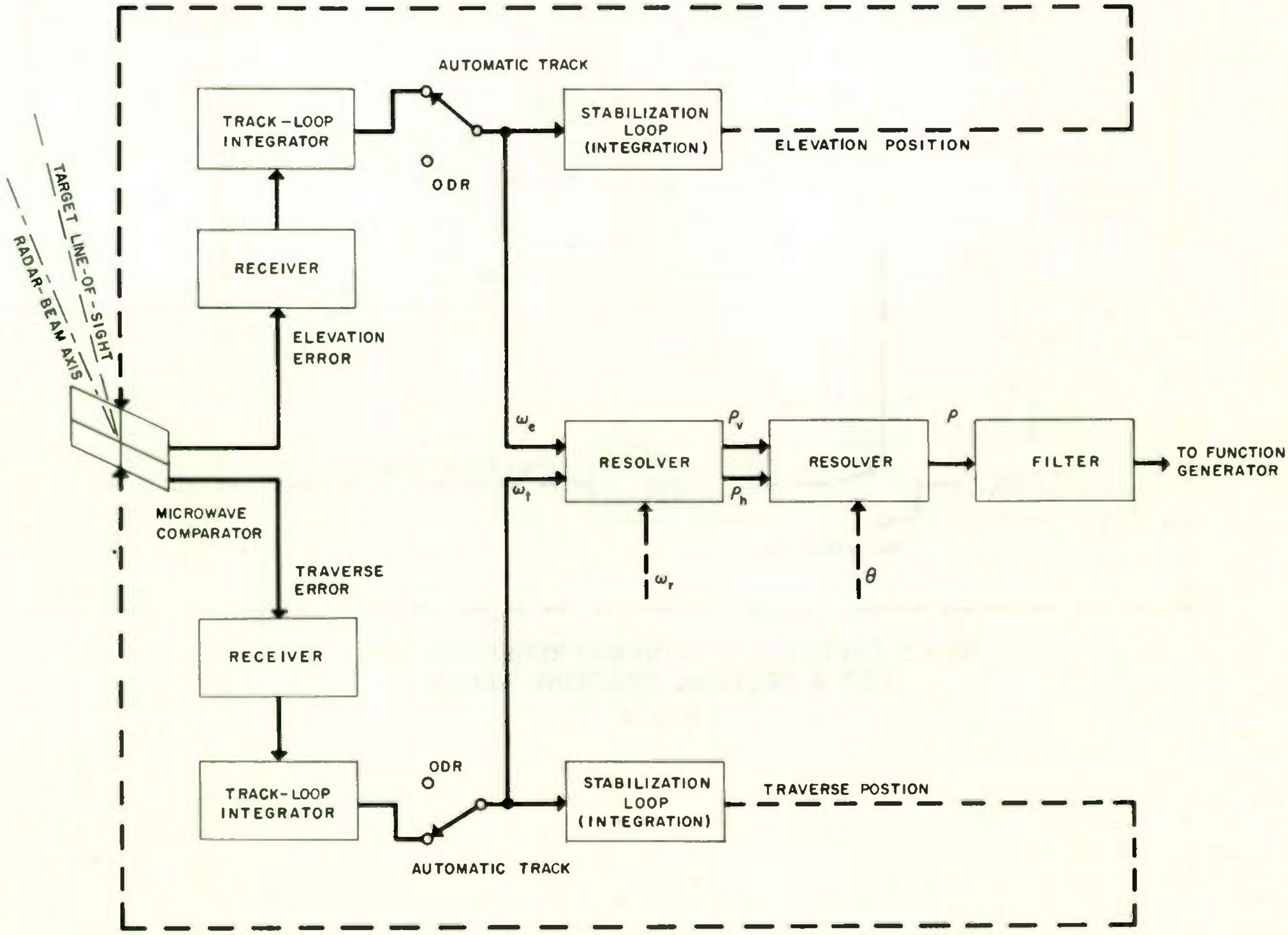
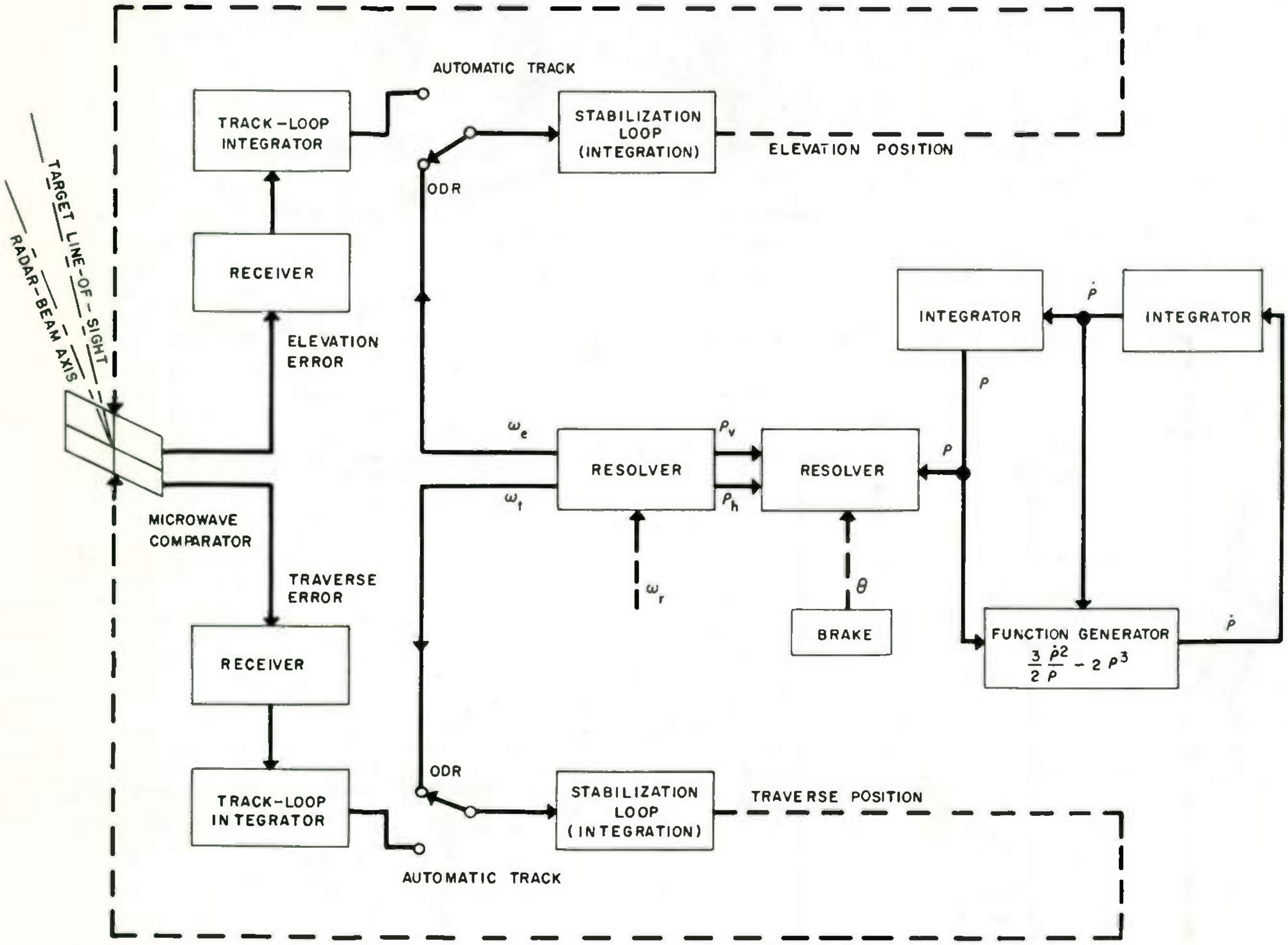
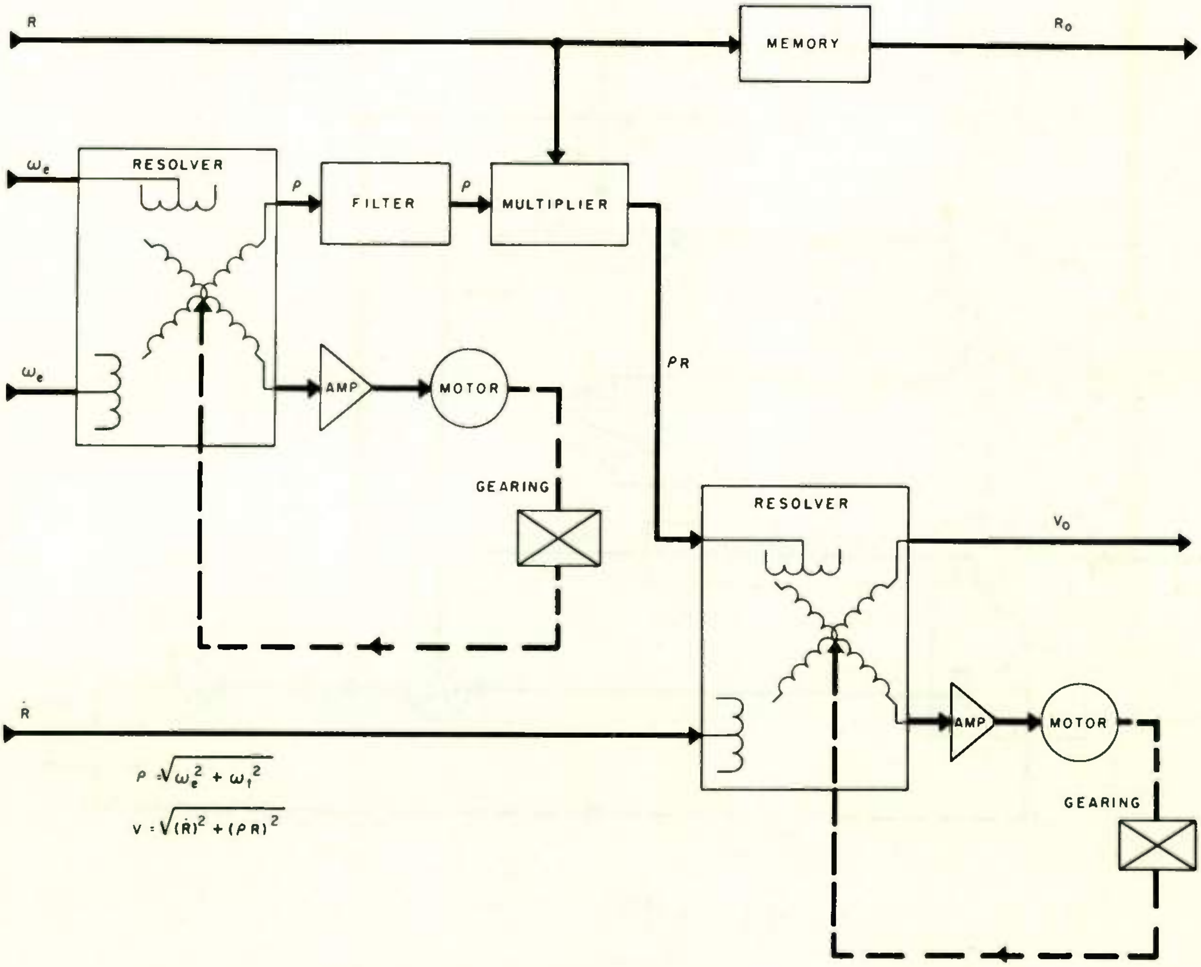


FIG. 3 NORMAL ANGLE-TRACKING OPERATION
OF A TACTICAL TRACKING RADAR



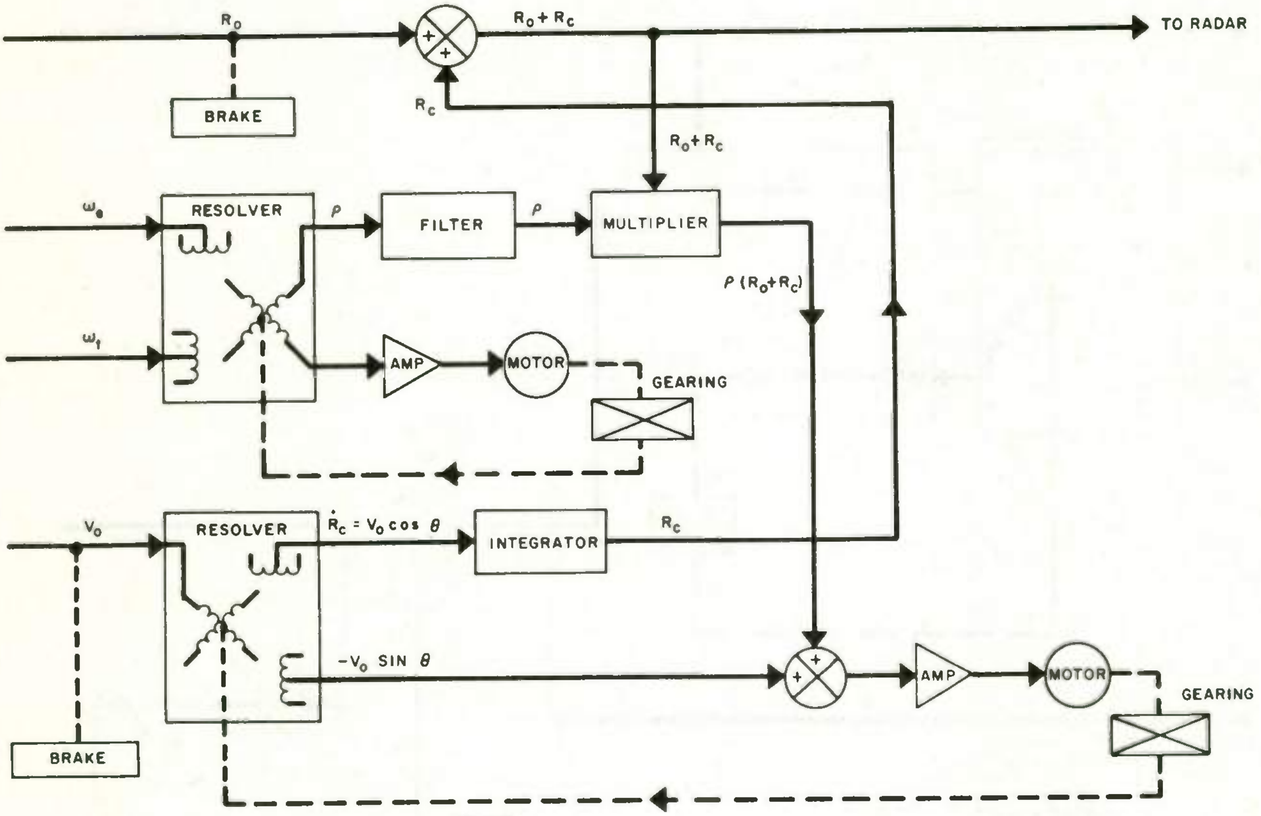
ANGLE OPERATIONAL DYNAMIC REGENERATION
FOR A TACTICAL TRACKING RADAR
FIG. 4



$$\rho = \sqrt{\omega_e^2 + \omega_t^2}$$

$$V = \sqrt{(\dot{R})^2 + (\rho R)^2}$$

FIG. 5 OPERATION OF ODR DURING NORMAL RANGE TRACKING



$$\rho = \sqrt{\omega_e^2 + \omega_t^2}$$

$$R_c = R_o + \int_{t_o}^{t} \left[ V_o^2 - (\rho R_c)^2 \right]^{1/2} dt$$

FIG. 6 RANGE ODR

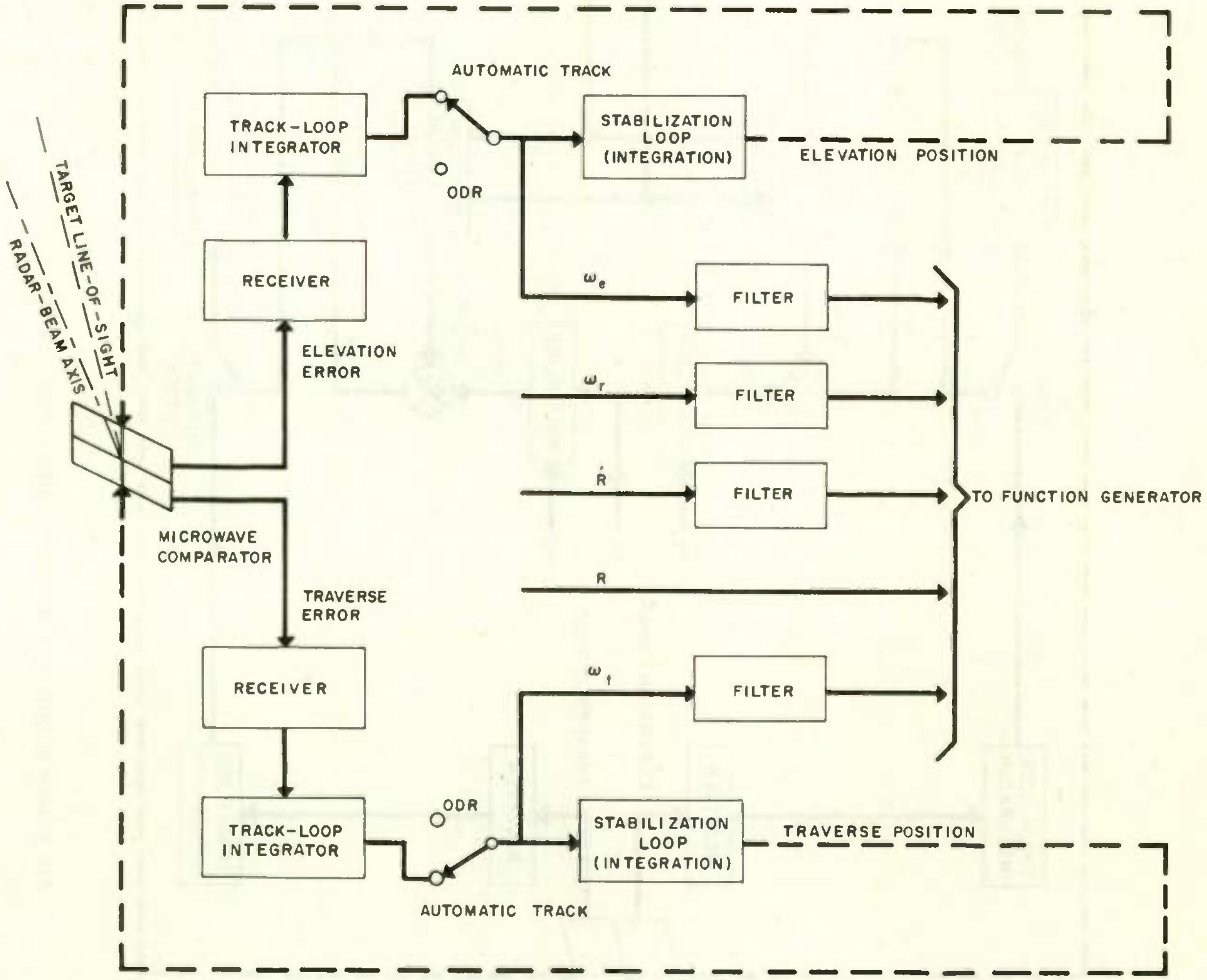


FIG. 7 NORMAL ANGLE-TRACKING OPERATION OF AN INSTRUMENT TRACKING RADAR

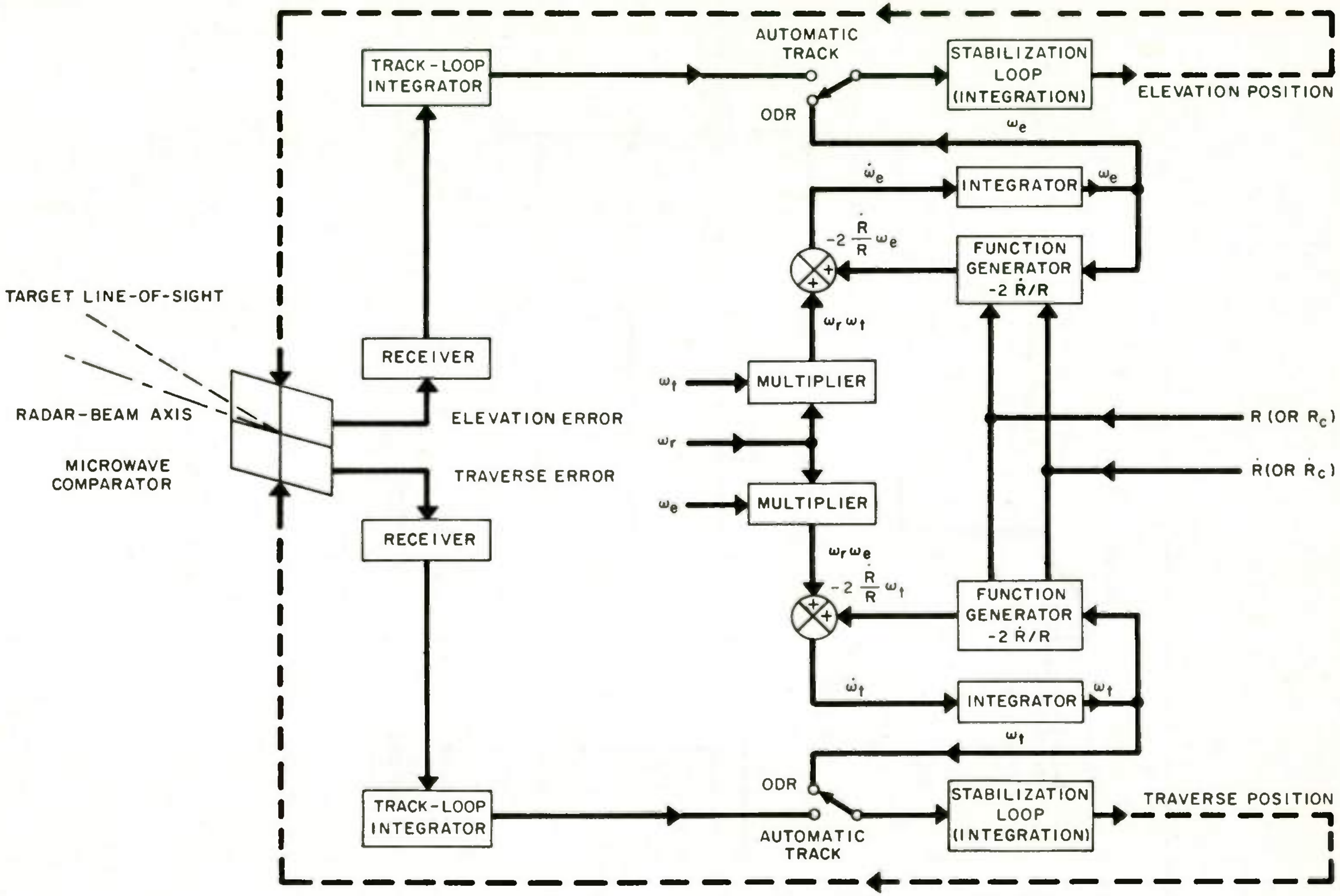


FIG. 8 ANGLE ODR FOR AN INSTRUMENTATION TRACKING RADAR

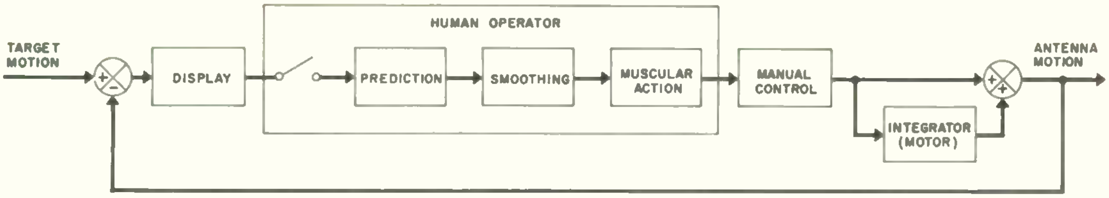


FIG. 9 MANUAL RATE-AIDED TRACKING SYSTEM

# AUTOMATIC STEERING TECHNIQUES[*]

Donald Barrick[**]

Antenna Laboratory
Department of Electrical Engineering
The Ohio State University
Columbus 10, Ohio

## Summary

This paper examines the feasibility and characteristics of several automatic steering systems for automobiles. It compares these systems on the basis of the following performance characteristics: (1) stability, (2) lateral acceleration, (3) error in tracking. The systems differ in that each has a different type of actuating signal for the front wheel positioning servo.

Methods are proposed for the generation of signals required for the various systems. The transfer function for each of the systems is derived and stability requirements are specified. Both transient and steady state responses of all systems are then determined so that the characteristics of the systems can be compared. The systems are compared on the basis of positional error and lateral accelerations.

## I. Problem Statement

In the development of an automatic steering system for automobiles one is concerned with two problems, namely, the generation of error signals which indicate position of the vehicle relative to the roadway and, secondly, the synthesis of a system which will provide a suitable dynamic response. The latter problem is not as simple as it might appear at first glance in that one must not only realize the usual requirements concerned with minimization of the displacement of the vehicle from the center of the roadway but also one must limit the magnitude of the lateral accelerations which may be involved. The lateral accelerations can be quite severe and very uncomfortable if acceleration considerations are neglected.

## II. Input Signals

All systems basically employ as input signals one or some combination of the following variables: (a) distance from the center of the lane, (b) radius of curvature, (c) angle between automobile axis and line of sight at some point a given distance ahead on the road. In this paper, the systems studied will be referred to as the Type A, Type B, Type C, and Type D systems. The input signal for the Type A system is simply the displacement of the automobile from the center of the lane. The Type B system uses the same input as the Type A system but adds a second signal proportional to the angle $\beta$ between the frame of the vehicle and the centerline of the lane. The Type C system adds to the displacement input of the Type A system an input proportional to the angle $\gamma$ between the front wheels of the car and the centerline of the lane. The Type D system uses as its input signal the angle between the centerline of the automobile and the line of sight to a point a given length "$\ell$" ahead on the road. This system is similar to radar in its ability to anticipate future conditions and changes and is similar in nature to the mode used by the human driver.

All these inputs, after amplification, are fed to the wheel positioning servo which controls the angular position of the front wheels.

Ideally a system should perform as nearly like a human driver as possible. The human attempts not only to minimize and eliminate error, but also to reduce lateral acceleration due to steering as much as possible. From actual tests made by the author it was found that a human driver may experience lateral accelerations of up to 1G[+] for brief periods, but in general the acceleration must be kept below about G/2 if skidding is to be avoided. The systems will be analyzed with these requirements in mind.

[*] The work involved in this paper was sponsored in part by the Ohio Department of Highways in cooperation with the Bureau of Public Roads.

[**] Mr. Barrick was formerly with the Antenna Laboratory, Department of Electrical Engineering, The Ohio State University, and is now serving in the U.S. Navy.

[+] The terminology "G" refers to a unit of acceleration normalized with respect to the acceleration of gravity.

## III. Detection Methods[+]

A possible scheme for obtaining the input signals needed for the Types A, B, and C systems is briefly described in this section. Consider a cable buried beneath the centerline of the lane of travel and excited by a low-frequency current. The resultant field will be circular in form. Let two coils be placed in a plane parallel to the cable and the centerline of the lane but equidistant from the center of the vehicle as shown in Fig. 1. The difference voltage in the two will be zero but as the two coils move out from the center, keeping their same difference spacing relative to each other, the difference voltage will rise. This difference voltage can be used as a measure of the displacement error input called for in the Types A, B, and C systems.

Now, if a third coil is placed so that the plane of the coil is perpendicular to the cable as shown in Fig. 2, the voltage induced in this coil is zero. The voltage induced in the coil as it is rotated from the perpendicular position varies nearly linearly with rotation for small angles. Thus this coil, if mounted to the frame of the auto, would give the second input called for in the Type B system; if the coil were to rotate with the front wheels it would generate the second input called for in the Type C system.

Ideally this input signal proportional to angle should not vary with the distance from the center, but with this arrangement it does to some extent, just as the linearity of the difference voltage from the first two coils becomes distorted as they are moved considerably from the center. In order to optimize these signals by proper choice of parameters, it is necessary to analyze these voltages further. The Biot-Savart Law for a long uniform wire and Faraday's Law give the desired relationships.

$$e = e_2 - e_1 = Kh\left[\frac{1}{\left(\frac{w}{2} - d\right)^2 + h^2} - \frac{1}{\left(\frac{w}{2} + d\right)^2 + h^2}\right] \qquad (1)$$

where

$$K = \frac{\mu \; NA \; C_1 \omega}{2\pi}$$

$\mu$ = permeability, assumed the same for air and ground

$r_i$ = distance from cable to coil "i"

$i$ = current = $C_1 \sin \omega t$

$w$ = separation of coils (fixed)

$\phi_i$ = angle between coil "i" and the centerline

$h$ = depth from coil centerline to cable

$d$ = instantaneous displacement of coils from the centerline

$A$ = area of coils (small compared to $h^2$)

$N$ = turns in each coil

$e = e_2 - e_1$ = difference voltage between coils.

See Fig. 3. Typical values of parameters are: $w/2 = 2.5$ ft, $h = 5$ ft, $I = .5$ amp (RMS), $\mu = 4\pi \times 10^{-7}$, $A = 10$ cm$^2$, $\omega = 500$ rad/sec A plot of e vs. d is shown in Fig. 4.

The voltage induced in the third coil varies with distance from the cable and is given below.

$$e_\beta = \frac{Kh}{d^2 + h^2} \sin \beta \cos \omega t \qquad (2)$$

$\sin \beta \approx \beta$ for $\beta << 1$.

A plot of $e_\beta / \beta$ is shown in Fig. 5.

There are many practical considerations not mentioned in this detection scheme as described above which may seriously affect the practical performance of such a system, but basically it can provide the input signals required.

The Type D system incorporates the usual complexities and problems of any radar system, and would have to be ruled out for reasons of size and economy at the present time. The system is considered here for the sake of comparison since it causes a dynamic response similar to that of a vehicle manually steered.

[+] The detection techniques involved in the Type A system is identical to that employed by RCA and General Motors and used in the automobiles and test track at Princeton, N.J.

## IV. System Transfer Functions[2]

In order to derive transfer functions for the four systems, certain assumptions will be made and then the response of the systems will be related to time. For the sake of analysis, it will be assumed that the test roads are laid out along the x axis. Next the automobile velocity will be assumed constant. Thus if the angle of the roadway with respect to the x axis is small, the x axis can be replaced by a time axis, relating the lateral displacement y of the vehicle to time. The actual lateral deviation of the roadway from the x axis at time t is given by $y_r$.

The transfer function* that will be used for the wheel positioning servo is given by

$$\frac{\alpha}{\alpha_d} = \frac{\omega_o^2}{s^2 + 2\zeta\omega_o s + \omega_o^2} \qquad (3)$$

$\alpha$ = actual wheel angle with respect to the auto

$\alpha_d$ = input to servo, desired wheel angle

$\omega_o$ = undamped natural frequency of servo = 20 rad/sec

$\zeta$ = damping factor of servo = .5

s = operator d/dt.

Now, given $\alpha$, the problem remains to find y, the lateral position of the automobile. Referring to Fig. 6, the velocity of the front wheels may be broken into two components.

$$\begin{aligned} V_y &= V \sin(\theta+\alpha) \approx V(\theta+\alpha) \\ V_x &= V \cos(\theta+\alpha) \approx V \approx x/t \end{aligned} \qquad (4)$$

for $\alpha$ and $\theta$ small

V = speed of auto

$\theta$ = angle of auto centerline with respect to x axis.

Here tire slippage is neglected. See Reference 2 for more exact representation.

From Fig. 7 it can be seen that $V_\theta$ is the component of front wheel velocity which causes the angle $\theta$ to increase. Assuming the car pivots about its rear wheels on a turn, it can be seen that

$$\frac{d\theta}{dt} = \frac{V\theta}{b} ,$$

$$V_\theta = V \sin \alpha \approx V\alpha \text{ for } \alpha << 1 ,$$

b = distance between front and rear wheels ,

$$\frac{V\alpha}{b} = \frac{d\theta}{dt} = s\theta$$

where

$$s = \frac{d}{dt}$$

$$Y = \int V_y \, dt \approx \int V(\theta+\alpha)dt .$$

The relationship between y and $\alpha$ becomes

$$\frac{Y}{\alpha} = \frac{V}{s}\left(1 + \frac{V}{sb}\right) . \qquad (5)$$

The block diagram showing this relationship is illustrated in Fig. 8. Now all the relationships are at hand for analyzing the four systems.

In the Type A system, the lateral error of the automobile d from the center of the lane is given by d = $y_r$ = y. Here $y_r$ is the position of the centerline of the roadway. Thus the block diagram for the system is shown in Fig. 11 and the transfer function is given by

$$\frac{y}{y_r} = \frac{K_d\omega_o^2 V(bs + V)}{bs^2(s^2+2\zeta\omega_o s+\omega_o^2)+K_d\omega_o^2 V(bs+V)} . \qquad (6)$$

In the Type B system, as seen again from Fig. 9, the second input proportional to the angle of the car with the centerline of the road is given by $K_\beta(\phi-\theta)$. Noting that $K_\beta$ is a constant and $\phi$ is the slope of centerline of lane relative to the x-y coordinates

$$\phi = \frac{dy_r}{dx} \text{ and } x \approx vt$$

*This is a common representation for a servo system. The damping factor $\zeta$ determines the transient overshoot and $\zeta\omega_o$ determines the rate at which the transient dies out. See Reference 1 or any book dealing with control systems.

it follows that

$$\phi = \frac{1}{V}\frac{dy_r}{dt} = \frac{s}{V}y_r \ . \tag{7}$$

The block diagram for this system is shown in Fig. 12. The transfer function is given by

$$\frac{Y}{Y_r} = \frac{\omega_o^2 V\left(K_d + K_\beta \frac{s}{V}\right)\ (V + bs)}{bs^2(s^2+2\zeta\omega_o s+\omega_o^2)+\omega_o^2 V(K_\beta+K_d b)s+K_d\omega_o^2 V^2} . \tag{8}$$

From Fig. 9, the second input for the Type C system proportional to the angle of the wheels with the centerline of the lane is given by $K_\gamma(\phi-\theta-\alpha)$. The block diagram for this system is shown in Fig. 13, and the corresponding transfer function is given by

$$\frac{Y}{Y_r} = \frac{\omega_o^2 V\left(K_d + K_\gamma \frac{s}{V}\right)(V+bs)}{bs^2(s^2+2\zeta\omega_o s+\omega_o^2)+K_\gamma\omega_o^2 bs^2+\omega_o^2 V(K_\gamma+K_d b)s+K_d\omega_o^2 V^2} . \tag{9}$$

The analysis of the input to the Type D system is somewhat different. See Fig. 10. This input signal, proportional to the angle between the centerline of the car and a point on the road ahead a distance "$\ell$" is given by $\psi-\theta$. The ordinate of the road a distance ahead "$\ell$" is $Y_r e^{\tau s}$ in operator notation where $\tau$ is a time to be determined by stability requirements. Since "$\ell$" is never more than a few degrees from the x axis, it can be approximated by

$$\ell = \tau V$$

$$\sin\psi = \frac{Y_r e^{\tau s} - Y}{\ell}$$

for $\psi << 1$

$$\alpha_d = \frac{Y_r e^{\tau s} - Y}{\ell} - \theta \ . \tag{10}$$

The block diagram for this system is shown in Fig. 14 and the transfer function is given by

$$\frac{Y}{Y_r} = \frac{\omega_o^2 (V + bs)e^{\tau s}}{\tau bs^2(s^2 + 2\zeta\omega_o s+\omega_o^2)+\omega_o^2\left[s(\tau V+b)+V\right]} . \tag{11}$$

## V. Stability Analysis

Using the transfer functions previously derived, stability criteria are found using the Routh's method.[3] For the sake of brevity, dedetails have been omitted and the final inequalities which must be satisfied are listed below.

Type A System

(1) $2\zeta\omega_o > K_d V$

(2) $2\zeta\omega_o > K_d V + 4\zeta^2 \frac{V}{b}$ .

Type B System

(1) $2\zeta\omega_o > \left(K_d + \frac{K_\beta}{b}\right) V$

(2) $2\zeta\omega_o\left(K_d+\frac{K_\beta}{b}\right) > V\left(K_d+\frac{K_\beta}{b}\right)^2 + 4\zeta^2 K_d \frac{V}{b}$ .

Type C System

(1) $2\zeta\omega_o(1+K_\gamma) > \left(K_d + \frac{K_\gamma}{b}\right) V$

(2) $2\zeta\omega_o(1+K_\gamma)\left(K_d + \frac{K_\gamma}{b}\right) > V\left(K_d+\frac{K_\gamma}{b}\right)^2 + 4\zeta^2 K_d \frac{V}{b}$ .

Type D System

(1) $2\zeta\omega_o > \left(\frac{1}{\tau} + \frac{V}{b}\right)$

(2) $2\zeta\omega_o\left(\frac{1}{\tau} + \frac{V}{b}\right) > \left(\frac{1}{\tau} + \frac{V}{b}\right)^2 + 4\zeta^2 \frac{V}{\tau b}$ .

Note that in the first three systems, velocity appears only on the right side of the inequalities. Therefore if the systems are stable at higher speeds, they will always be stable at lower speeds as would be expected. In the fourth system, decreasing velocity will improve stability but does not always guarantee it.

Throughout the analysis an automobile speed of 100 ft/sec or about 70 mps will be used since at lower speeds the stability margin is increased and the systems respond in a shorter distance to errors in position. Also, since lateral acceleration varies as $a = v^2/r_c$ where $r_c$ is the radius of curvature of the road, it can be seen that at lower speeds the lateral acceleration will be reduced considerably.

Assigning typical numerical values to the parameters gives a more accurate quantitative description of stability demands upon gain constants. Let $V = 100$ ft/sec, $b = 20$ ft, $\omega_o = 20$ rad/ft, $\zeta = 0.5$. For the Type A system the inequality which must be met for stability is

$$K_d < .15 \text{ rad/ft} \quad .$$

A reasonably safe value to assume for $K_d$ is 0.1. This value will be used for $K_d$ in the second and third systems so that the merits of the addition of the second input may be evaluated. For the Type B system, the following inequality results:

$$\left[K_\beta < 2\right] .$$

Thus for stability an additional constraint must be imposed upon $K_\beta$ as well as $K_d$ in the Type B system.

For the Type C system, the inequality is:

$$\left[K_\gamma > - .67\right] .$$

Note here that the system can never become unstable for variations in $K_\gamma$ as long as $K_\gamma$ is positive.

The inequality for the Type D system reduces to

$$\left[\tau > .087 \text{ sec}\right] .$$

Thus the lead time must remain above a minimum safe value.

In order to verify and study more fully the effect of variation of the gain constants upon stability and transient response, root position plots[4] vs. gain constants were derived and plotted as shown in Fig. 15. Only the first can be called a true root locus plot, for in all the other systems, changing the gain constant affects not only the locations of the poles but also of the zeros as may be seen from the transfer functions. These plots verify the stability criteria derived by the Routh method and show the effects of variation of gain parameters. Note the effect of increasing $K_\gamma$ in the Type C system.

## VI. Steady-State Response

For circular tracts the angle $\alpha$ of the front wheels relative to the body is given by $\sin \alpha \; b/r_c$ and $\sin \alpha \approx \alpha$ for $r_c >> b$. See Fig. 16. In the steady state, $\alpha_d = \alpha$. In this analysis a 300 ft diameter circle was chosen. For the Type A system, with $\alpha_d = \alpha = K_d d = b/r_c = 20/150$, the displacement d from the center line is given by $d = 2/15 \cdot 1/K_d$ ft.

For the Type B system, $K_d d + K_\beta \beta = \alpha_d$. But here, $\beta$, the angle between the centerline of the body and the tangent to the centerline of the lane is given by $\alpha$.

$$d = 2/1.5\,(1-K_\beta)\text{ft for } K_d = 0.1.$$

For the Type C system, the angle between the front wheels and the centerline of the lane must be zero if the car is following the curve accurately one has

$$K_d d + K_\gamma \cdot 0 = \alpha_d = 20/150 .$$

In this case

$$d = 2/1.5 = 1.33 \text{ ft}$$

for $K_d = 0.1$.

For the Type D system, see Fig. 17, it is evident that the angle between the centerline of the vehicle and the line "$\ell$" is $\alpha$.

$$d \approx \ell \sin \alpha \approx \alpha \ell \qquad \text{for } \alpha << 1$$

$$\alpha = 2/15$$

and

$$d = 2/15\ell = 2/15\,\tau V .$$

If V is chosen to give 1/3 G lateral acceleration, $d = 5.16\tau$. Plots of the relationships between gain and distance from the centerline are given in Fig. 18.

In steady-state analysis of a roadway varying sinusoidally, care was taken to choose values for amplitude and frequency such that (1) lateral peak acceleration was held to 1/3G at 70 mph, (2) the lateral distance travelled was always small compared to the distance along the x axis.

$Y/Y_r = G(s)$,  $Y_r = A \sin \omega t$
$G(s)$ = system transfer function

$$d = Y_r - Y = Y_r\,[1 - G(s)] = D \sin(\omega t - \Omega)$$

$$D = \text{maximum displacement} = A\,|1 - G(j\omega)|\ .$$

For $Y_r = 250\,(\text{ft}) \sin .2t$

$$\therefore\ D = 250\,|1 - G(j.2)|\ .$$

Plots of D vs. the various gain constants of the systems were computed and are shown in Fig. 19. Note that these plots correspond very closely to those for the circular plot, indicating that the steady state response for nearly any type of curve can be predicted fairly well from these charts. For the Type A system, error is reduced by increasing $K_d$. At $K_d = 0.1$ steady state error for the circle was held at 1 1/3 ft, which is quite tolerable. In the Type B system with $K_d = 0.1$, the steady state error can be eliminated by setting $K_\beta = 1.0$. For the Type C system, nothing is gained over the Type A system as far as reduction in steady state error. In the Type D system, error rapidly becomes intolerable as $\tau$ is increased. Notice that in the first three systems, which "see" only present error, the response lags the input and the automobile tracks on the outside of the circle. However, in the Type D system, the response leads the input and the automobile tracks on the inside of the circle.

## VII. Transient Response

The transient response of a system may be measured in many ways. In this analysis the systems were simulated on an analogue computer and a step function was used as the input signal. This might be visualized in an actual highway system as being a sharp displacement in the centerline. The displacement was held to 0.3 ft so that the results as far as error and lateral accelerations at 70 mph would be reasonable. The response traces, which were compared with calculated curves as a check, are shown for the four systems at various gain constants in Fig. 20. Simultaneously traces were made of the lateral acceleration experienced from the discontinuity as measured in G's.

Notice from the two sets of traces that for the Type A system, response becomes more oscillatory as $K_d$ is increased resulting in higher peak acceleration and longer damping time. For the Type B system, response becomes worse due to longer damping time and higher frequency of oscillation when $K_\beta$ is increased. This type of a ride would be quite unbearable to a passenger in the car. In the Type C system, oscillations and resultant lateral accelerations can be nearly eliminated by increasing $K_\gamma$ without impairing stability. The Type D system certainly gives the smoothest ride of all as $\tau$ is increased, and the leading response can be noted in the trace origins.

By studying the block diagrams for the various systems one notes that the effect of the second input in the Type B system is to introduce error-rate feedback. This tends to make for a closer tracking system at the expense of a very oscillatory response. The second input of the Type C system adds to this error-rate feedback an accelerometer feedback which in this case gives a much smoother response.

## VIII. Conclusions

Of all the systems considered only two of the systems studied have acceptable dynamic and static characteristics judging the response upon the basis of both displacement errors and lateral accelerations. The two acceptable systems are the Type C and Type D systems. The Type D system is not practical and for this reason it is concluded that the Type "C" system is most promising of the automatic steering systems.

## References

1. Servomechanisms and Regulating System Design , by Chestnut and Mayer, Vol. I, 2nd Ed., John Wiley and Sons, Inc., 1959.

2. "The Application of Mathematics to a Basic Study of Automobile Control and Stability Problems," paper by Robert H. Kohr (A.M. Research Laboratories) printed in "General Motors Engineering Journal," April-May-June 1959.

3. Transients in Linear Systems by Gardner and Barnes, Vol. I, John Wiley and Sons, Inc., 1957.

4. Chestnut and Mayer, op. cit.

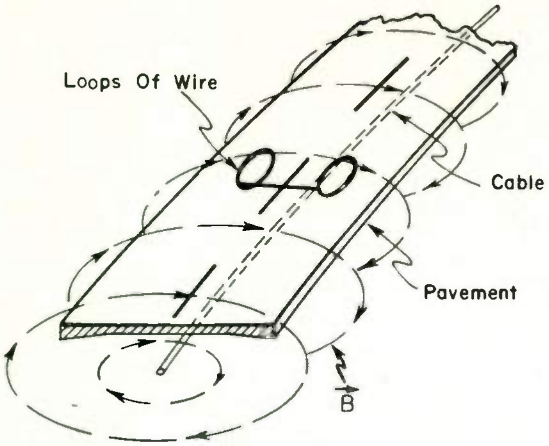


Fig. 1. Detection coils for displacement error.

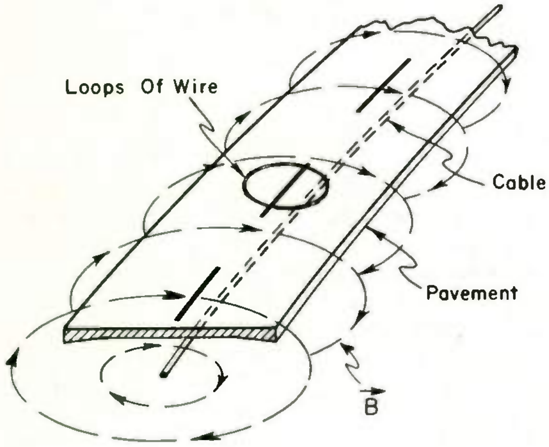


Fig. 2. Detection coil for angular error.

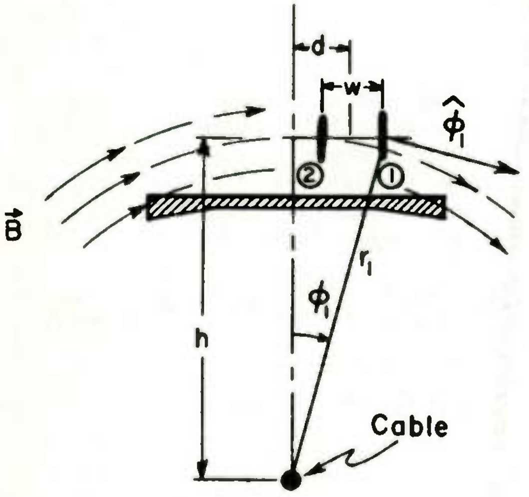


Fig. 3. End view of displacement coils with symbols.

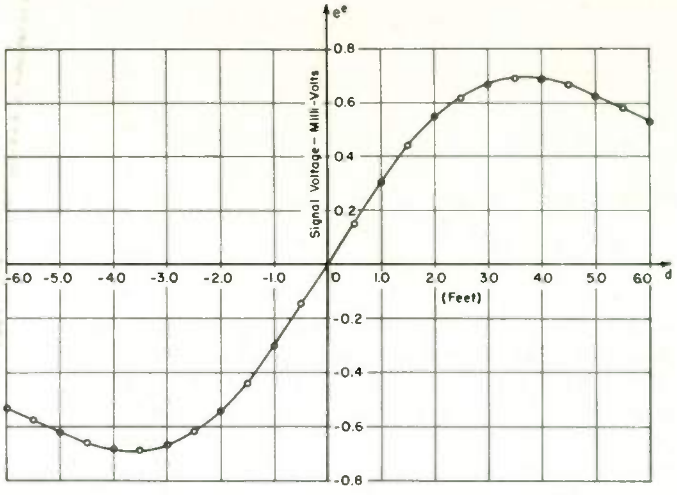


Fig. 4. Displacement error voltage vs displacement.

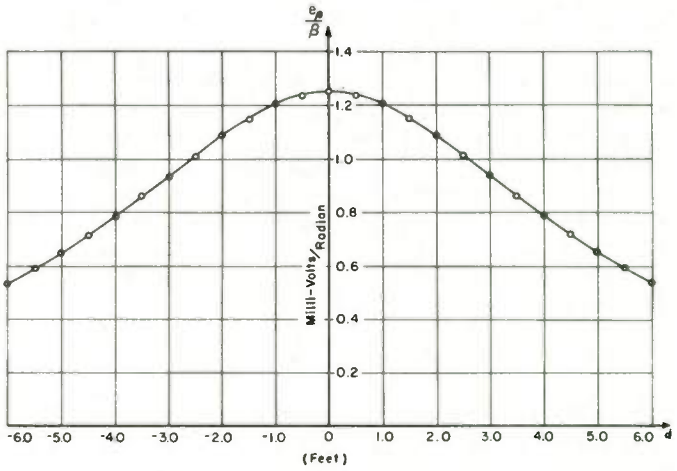


Fig. 5. Angular error voltage vs displacement from center.

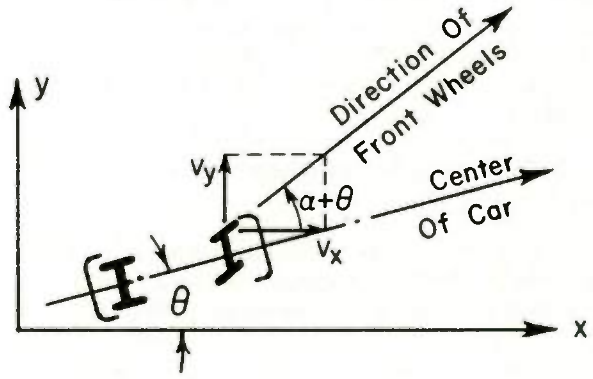


Fig. 6. Velocity of front wheels.

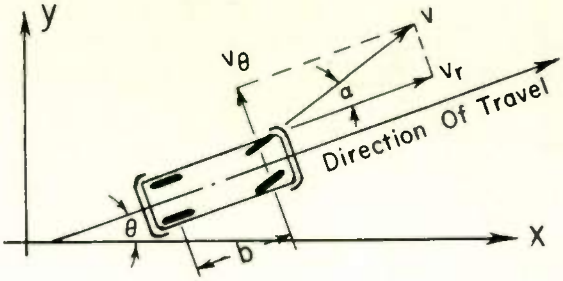


Fig. 7. Rate of change of angle of frame "θ".

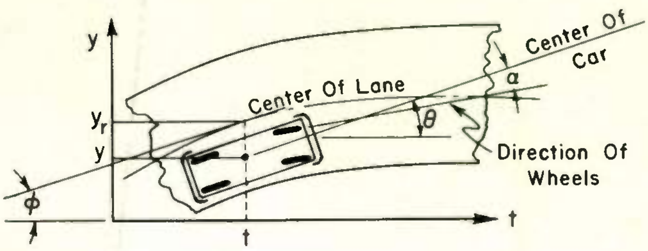


Fig. 9. "Y" vs "t" coordinates showing displacement and front wheel angle errors.

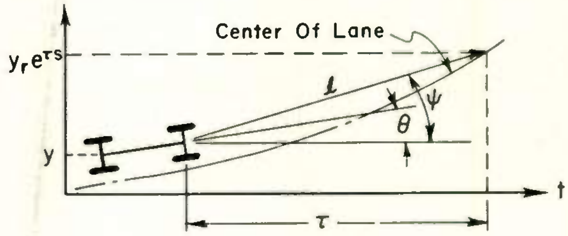


Fig. 10. "Y" vs "t" coordinates showing future angle error.

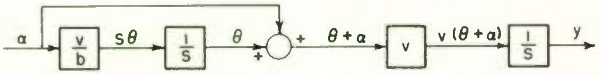


Fig. 8. Block diagram for front wheel angle-auto position system.

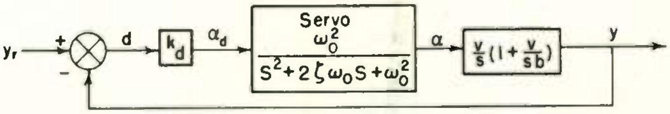

$y_r$
+
-
d
$k_d$
$\alpha_d$
Servo
$\frac{\omega_0^2}{S^2+2\zeta\omega_0 S+\omega_0^2}$
α
$\frac{v}{s}(1+\frac{v}{sb})$
y

Fig. 11. Block diagram for Type A system.

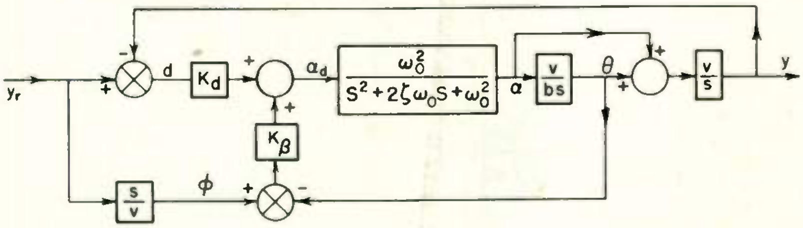


Fig. 12. Block diagram for Type B system.

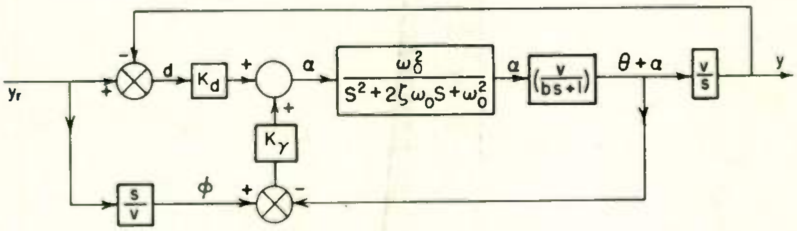
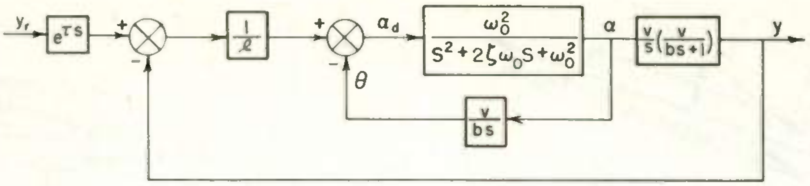

$y_r$
-
+
d
$K_d$
+
+
α
$\frac{\omega_0^2}{S^2+2\zeta\omega_0 S+\omega_0^2}$
α
$\left(\frac{v}{bs+1}\right)$
θ + α
$\frac{v}{s}$
y
$K_\gamma$
$\frac{s}{v}$
φ
+
-

Fig. 13. Block diagram for Type C system.



$y_r$
$e^{\tau s}$
+
−
$\frac{1}{\ell}$
+
−
$\theta$
$a_d$
$\frac{\omega_0^2}{S^2+2\zeta\omega_0 S+\omega_0^2}$
$\alpha$
$\frac{v}{s}\left(\frac{v}{bs+1}\right)$
y
$\frac{v}{bs}$

Fig. 14. Block diagram for Type D system.

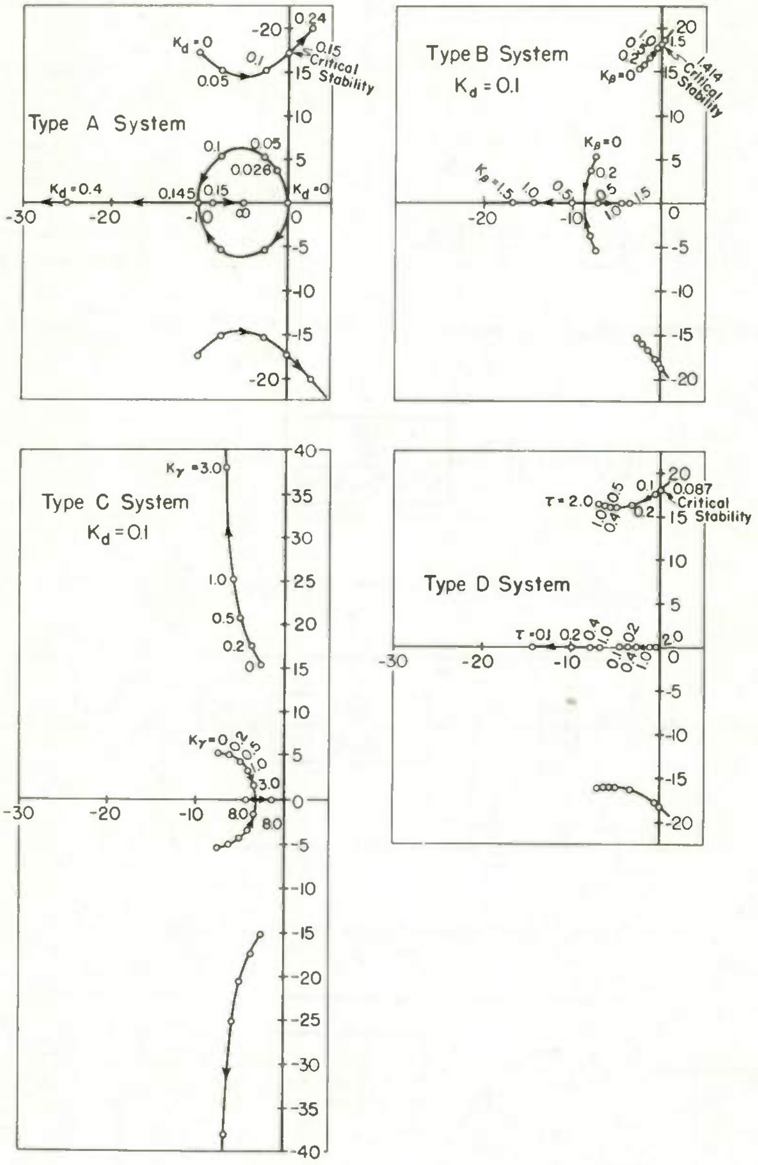


Fig. 15. Root positions vs gain constants.

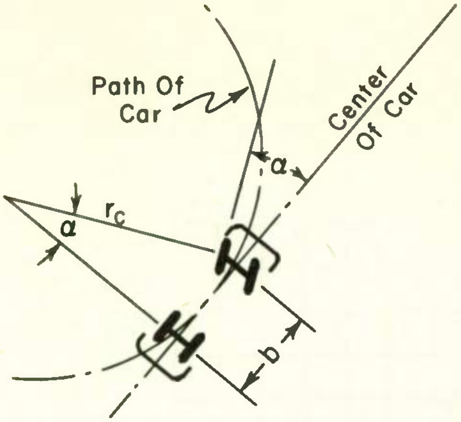


Fig. 16. Relationship between angle of front wheels and radius of path.

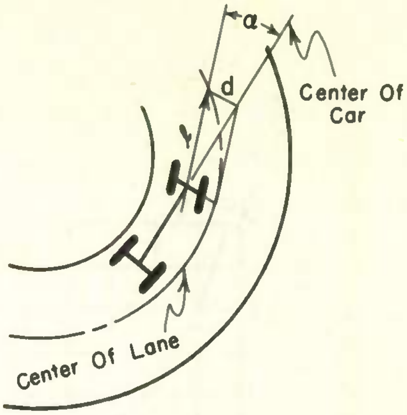


Fig. 17. Displacement from circular track for Type D system.

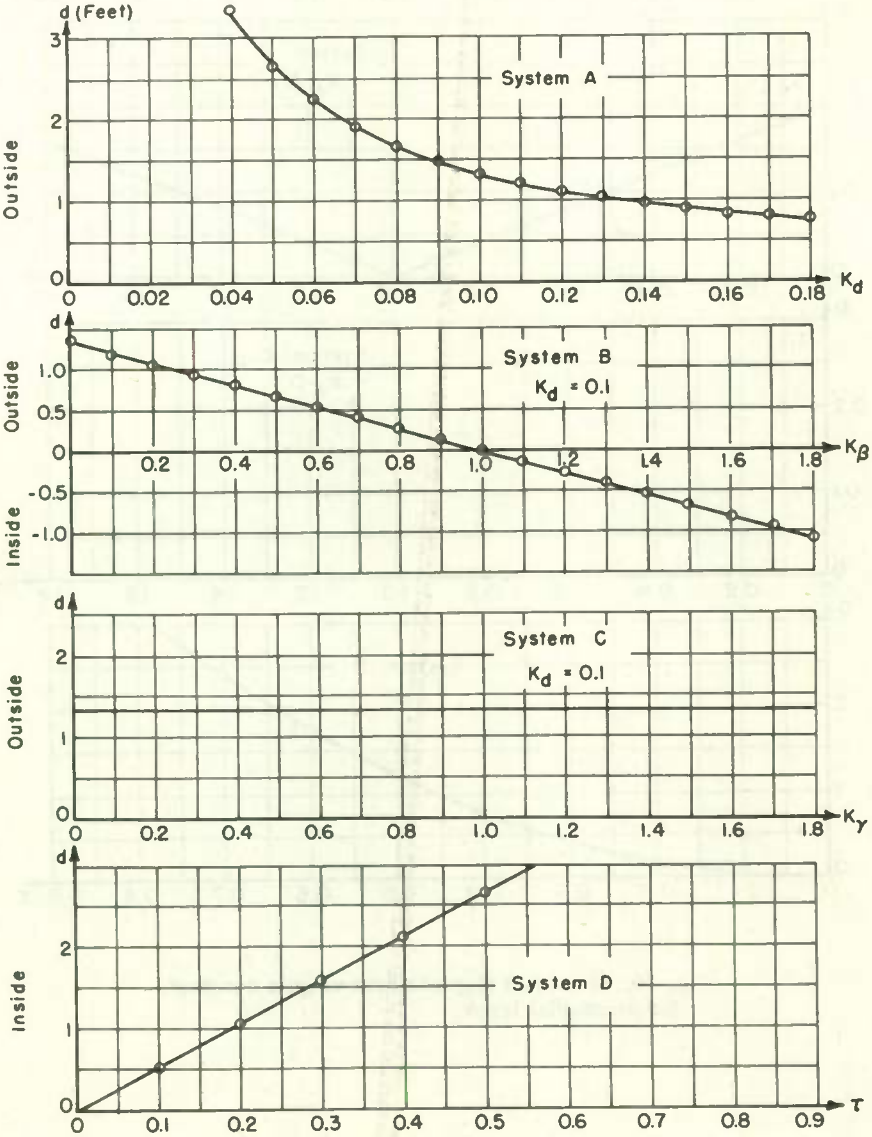


Fig. 18. Displacement vs gain constants for circular track.

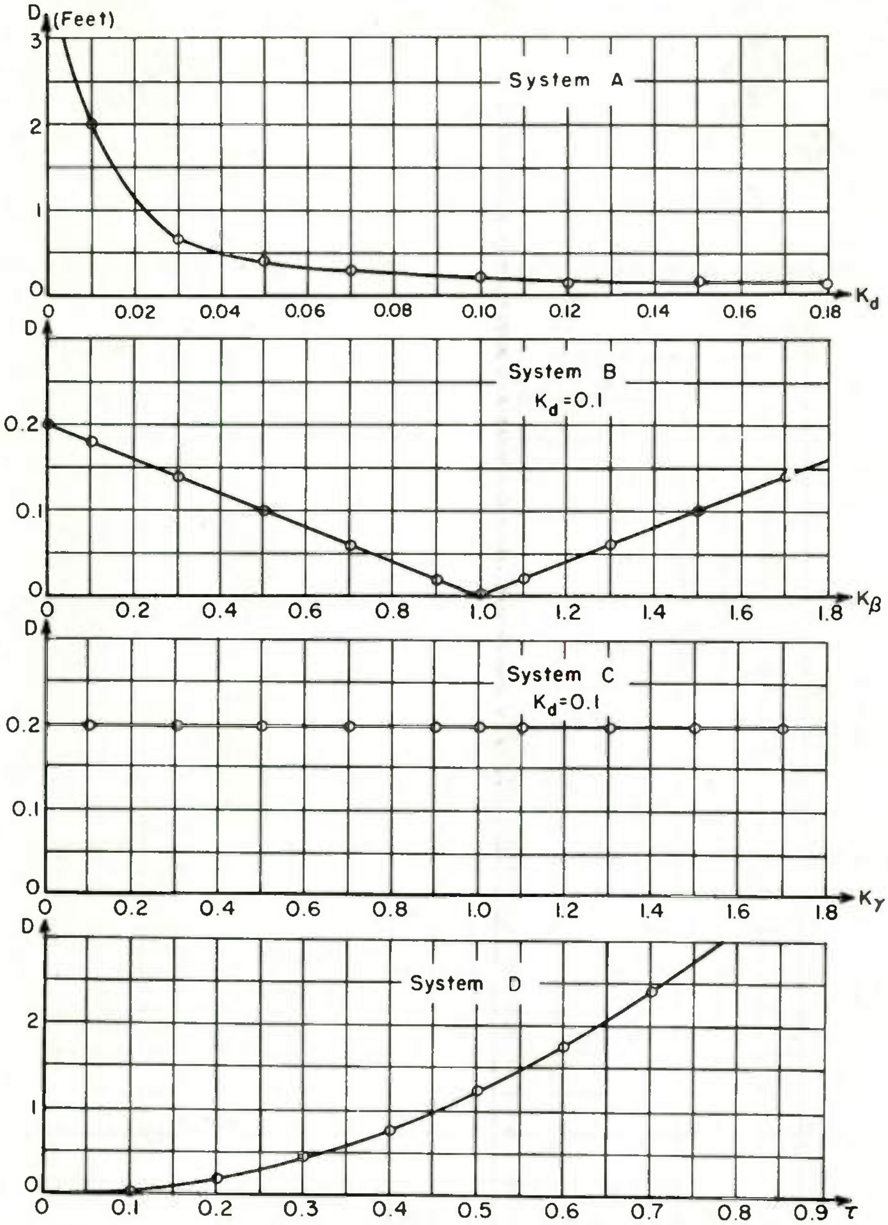
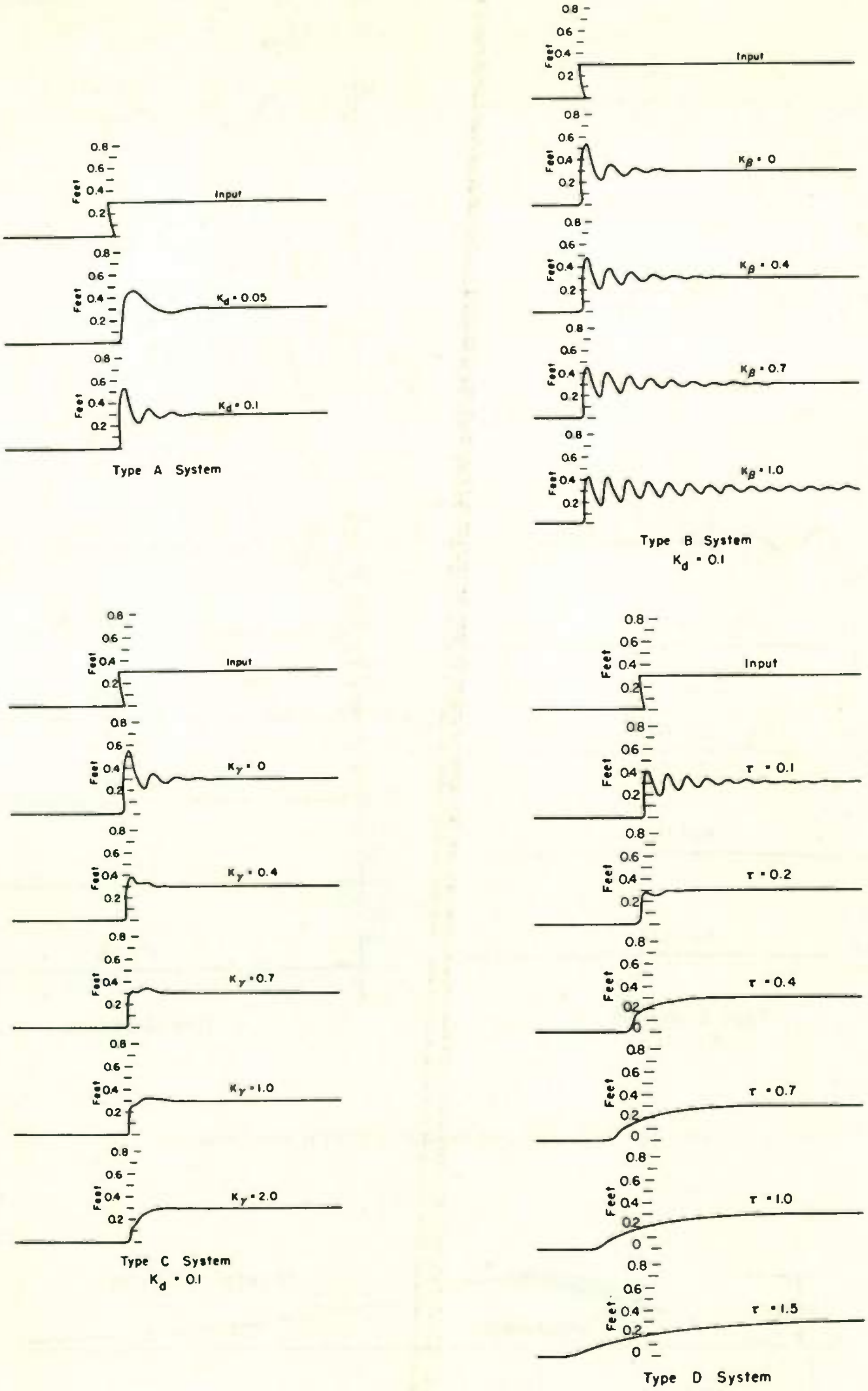


Fig. 19. Maximum displacement vs gain constants for sinusoidal track.



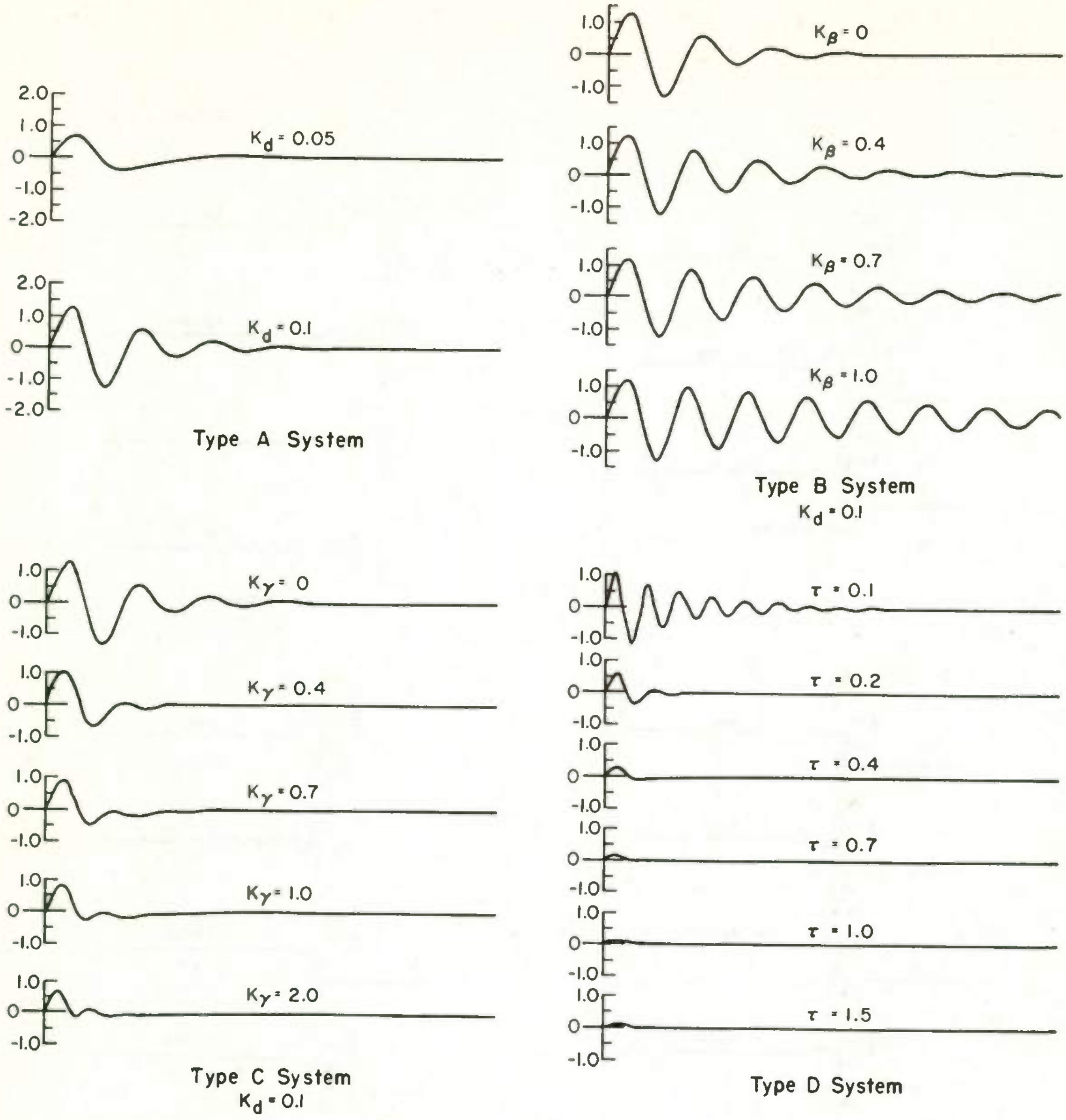
Type A System

Type B System
$K_d$ = 0.1

Type C System
$K_d$ = 0.1

Type D System

Fig. 20. Step input and output response.



Type A System

Type B System
$K_d$ = 0.1

Type C System
$K_d$ = 0.1

Type D System

Fig. 21. Step input and resulting lateral acceleration.

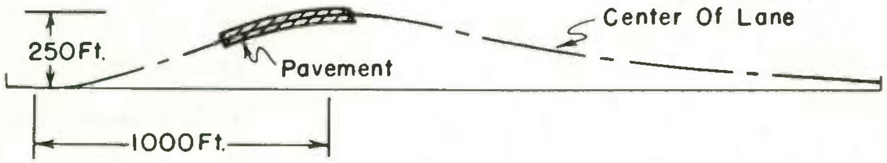


Fig. 22. Random road section used as input.

# SEARCH AND LOCK RECEIVER

by

F. O. Gray

Airborne Instruments Laboratory
A Division of Cutler-Hammer, Inc.
Deer Park, Long Island, New York

## Summary

The Search and Lock Receiver system developed at AIL is capable of searching across a broad frequency band and accurately locking on a signal intercepted within that band. The frequency limits over which the search mode takes place can be varied by external digital commands.

To achieve high scanning rates along with accurate lock-on, a discriminator-controlled, bang-bang type servo is used.

The logical-image-rejection system is used to prevent the receiver from attempting to lock-on when a superheterodyne image frequency is intercepted.

## I. Introduction

The rapid and automatic search of wide frequency bands for the location and identification of radio frequency interference sources is becoming increasingly important.

Equipment capable of conducting rapid surveillance of the radio frequency environment is required at air traffic control centers, monitoring stations, and missile launch and tracking stations.

The receiver described in this paper is designed to search for and then tune to pulsed RF signals. In addition, this receiver can accommodate continuous-wave or high-duty-cycle signals by local gating modulation that effectively converts them to pulsed signals in one channel for control purposes. This technique can be applied within any RF band. Obviously, it is important to search the selected frequency band rapidly. However, if the scanning speed is too fast, the receiver band pass may pass through the interfering signal frequency between RF pulses. Based purely on system considerations, therefore, the maximum scan speed must be determined by the maximum interval between pulses, by the bandwidth of the receiver, and by the maximum number of pulses that will be required to initiate the lock-on command.

When a signal is intercepted, the receiver switches from the search mode to the lock mode. The receiver must tune within 0.1 Δ of the signal frequency, where Δ represents the IF bandwidth. However, it must not lock on an image frequency. In addition to the above requirements, the receiver must respond to digital commands that determine scan limits, and when commanded to do so, must stop at a specified frequency.

## II. Search Mode

During the search mode, the receiver scans continuously between two preset frequency limits. The binary output of an analog to digital encoder, which is coupled to the tuning shaft of the receiver, is compared with binary words representing the high and low limits of the desired frequency sector by two digital matcher circuits. When the encoder word matches one of the frequency limit words, a matching pulse is produced that changes the state of the direction flip-flop, and thus reverses the scanning direction of the receiver. These limits can be reset to permit scanning any sector within the band.

As previously mentioned, the receiver can be stopped at any desired frequency. This is done by first setting the limit at the high frequency end of the scan sector to correspond to the desired frequency and then by giving the stop command. The receiver will scan the band until the match occurs, and then stop on the preprogrammed word representing the high frequency limit.

If an RF signal is received as the band is scanned, the control circuits must either command lock-on or inhibit it until the receiver scans past the signal. For the purpose of this paper, the signal response is obtained when the local-oscillator frequency is above that of the incoming signal, and the image response is obtained when the local-oscillator frequency is below that of the incoming signal. Later on I will describe the method used to inhibit lock-on for image frequencies.

## III. Lock Mode

When the receiver intercepts a valid signal, the lock mode is initiated. In this mode, tuning error information is derived from a discriminator that is fed from a limiting IF amplifier stage. As seen in Figure 1, the discriminator and detector outputs go to the lock control. The lock control performs the logic and timing functions that generate the direction and speed commands for the scanning motor. Figure 2 is a logic diagram of the lock control.

Frequency versus amplitude characteristics of the discriminator and detector are shown in Figures 3A and 3B. Superimposed on these characteristics are the outputs of the 10-microsecond monostable flip-flops (A and B) and the 5-microsecond monostable flip-flop (C). Variable thresholds on the input to each monostable flip-flop are adjusted to give exactly the characteristics in Figures 3A and 3B.

If the inhibit flip-flop has been set, then the discriminator, through monostable flip-flops A and B, will control the scan direction. An output from A causes the scan motor to move up in frequency whereas an output from B causes the motor to move down in frequency. An output from either A or B will cause the scan motor to remain on, whereas a pulse from C without a pulse from A or B will turn the motor off.

It can also be seen from Figure 2 that when the inhibit flip-flop is set, the scan motor will revert to slow speed.

The monostable flip-flop D is unlike ordinary ones because, after each set pulse, it will remain in the "1" state for 8 milliseconds regardless of its initial condition. Its purpose is to reset the inhibit flip-flop when no pulse is received within 8 milliseconds.

The resulting control characteristic of these circuits, as seen in Figure 3C, is typical of a relay-operated or bang-bang servo. This type of servo has a fast response time because maximum and minimum errors produce the same correction rate.[1]

Most AFC systems that are designed to operate on pulsed RF signals integrate the output of the discriminator to produce a DC level proportional to tuning error. The method that I have just described produces error information on a pulse-to-pulse basis, and unlike the integration method it does not limit response speed.

Phase-plane methods of analysis are applicable to this type of system since all of the nonlinearities are signal-dependent.[2] Figure 4 is a phase-plane diagram wherein the x coordinate is scanning position and the y coordinate is scanning speed. The case of only one initial condition is shown. This diagram shows the point where the servo parameters are switched and the results of changing system parameters such as scan speed, slow or fast, or servo amplifier gain. You will note that Figure 4 shows that there are limits between which commands can be given. This is caused by variations in time when pulses will be received. The variations are minimized by overshooting the signal once and then approaching the off zone at slow speed.

With the bang-bang type of servo, there is an optimum condition where the off command can be given, and the scan velocity and tuning error will immediately drop to 0. The threshold on the direction command monostable flip-flops was selected so that noise would not cause false triggers. Because these thresholds are determined directly by RF noise and indirectly by system parameters--such as the receiver noise figure, IF bandwidth, etc.,--the frequency where the off command is given must be fixed. Another parameter easily varied is the scan speed. On this receiver, adjusting the slow scan speed to about 25 percent of the fast scan speed will give minimum scanning error.

A new type of servo amplifier was designed for this system that permits very high gain without any loop stability problems. The measured turn-around time and stop time for slow-speed scan are 10 milliseconds and 3.5 milliseconds respectively.

## IV. Image Rejection

Some method had to be found to prevent the receiver from trying to lock on image frequencies. It is well known that the sense of control commands for AFC is reversed for image frequencies of superheterodyne receivers.[3]

For example, if the receiver is scanning up in frequency and intercepts an image frequency, as seen from Figure 3A, the first pulse from the discriminator will be negative. This pulse will change the state of the direction flip-flop so that a "down" command will be given. The motor will scan down in frequency until the low frequency limit is reached, and then the scan direction will change to up. This cycle will continue to repeat itself. Thus, the receiver scan frequency is constrained below the incoming image frequency.

To eliminate this problem, several logic gates and a flip-flop were added to the lock control. These inhibit the discriminator direction commands from changing the scan direction until after certain conditions occur. If the receiver is scanning up and an"up"command is given by monostable flip-flop A, or if the receiver is scanning down and a "down" command is given by monostable flip-flop B, then the inhibit flip-flop will be set. The discriminator will control the scan direction until the inhibit flip-flop is reset. These conditions are always met when the receiver frequency approaches the signal frequency; therefore, lock-on is initiated upon receiving the first pulse from the discriminator.

However, in the case of an image frequency, these conditions are not met until after the receiver has scanned past the center of the image frequency. Even then, when the inhibit flip-flop is set and the discriminator commands control direction, the receiver continues to scan in the same direction. Monostable flip-flop D will reset the inhibit flip-flop after 8 milliseconds.

## V. Conclusion

All of the desired design objectives for the Search and Lock Receiver have been achieved. In addition to the achievement of system performance goals, the design of the Search and Lock Receiver offers the advantages discussed below. Using a method of logical image-rejection eliminated the preselector. The use of several logic gates and a flip-flop eliminated the problems associated with preselector tracking, its size and weight, and the extra load on the scanning motor. The servo loop used for scanning the receiver proved very effective and reliable over extreme temperature variations. The digital-type circuits used throughout the lock control, the modulator, and the servo amplifier have simplified the electronic circuitry and have made the servo loop more stable against variations in temperature and changes in component values.

## VI. References

1. R. E. Kopp, "On Bang-Bang Adaptive Control Systems," IRE International Convention Record, vol 9, Part 4, p 3-17, 1961.

2. J. G. Truxal, "Control System Synthesis," McGraw-Hill, New York, 1955.

3. J. G. Stephenson, "Combined Search and Automatic Frequency Control of Mechanically Tuned Oscillators," Proc IRE, vol 38, No. 11, p 1314-1317, November 1950.

## Acknowledgment

The author wishes to thank R. E. Leslie for his invaluable help, and M. R. VanDusen and B. A. Woodward for many helpful discussions.

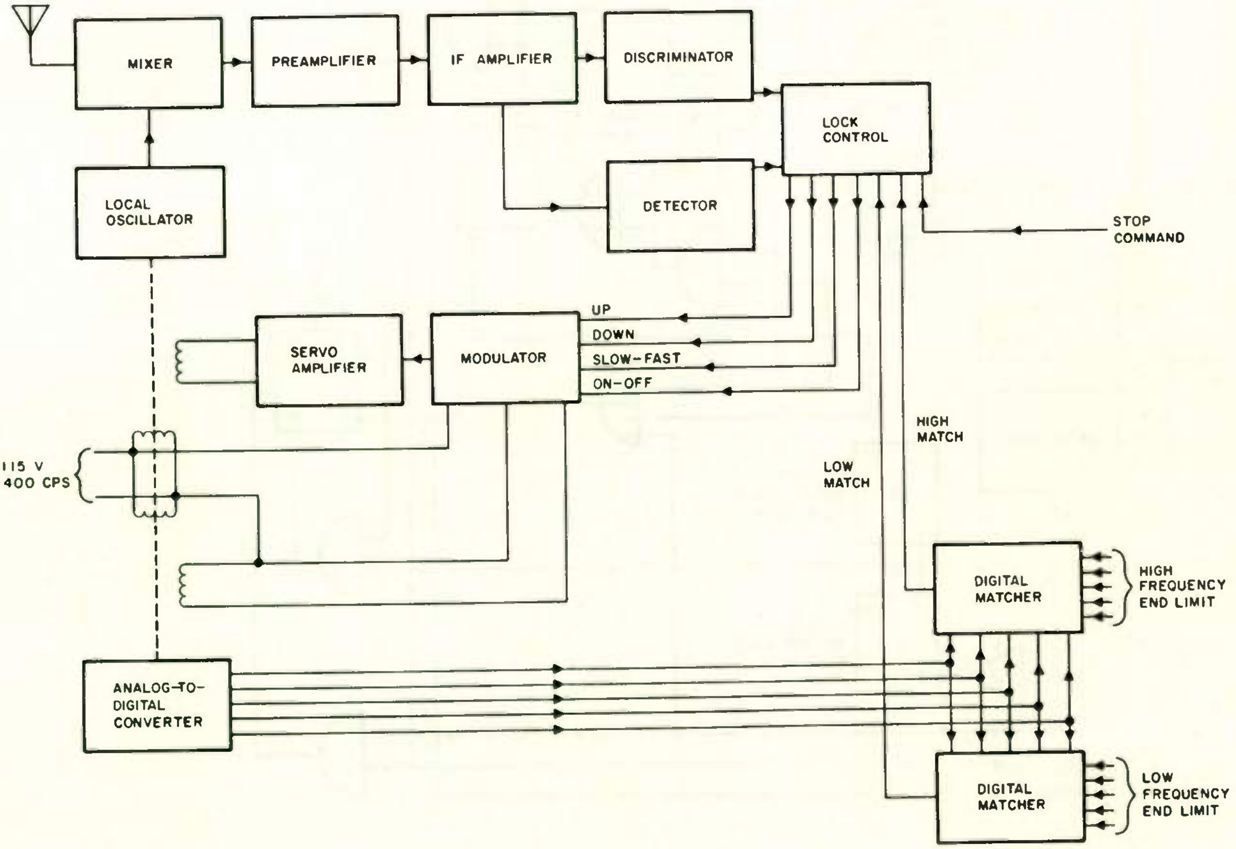


Fig. 1. Block diagram of search and lock receiver.

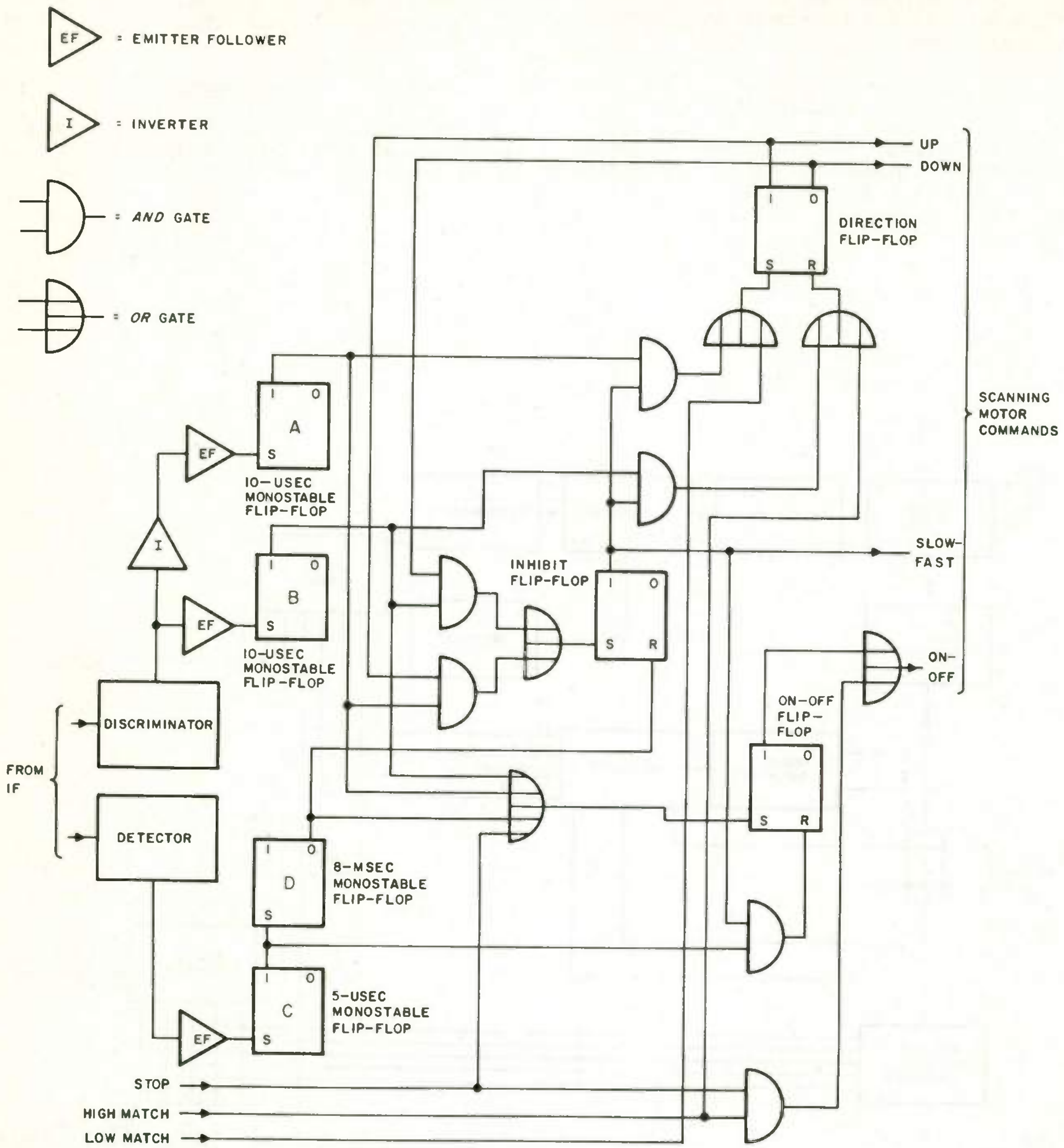


Fig. 2. Logic diagram of lock control.

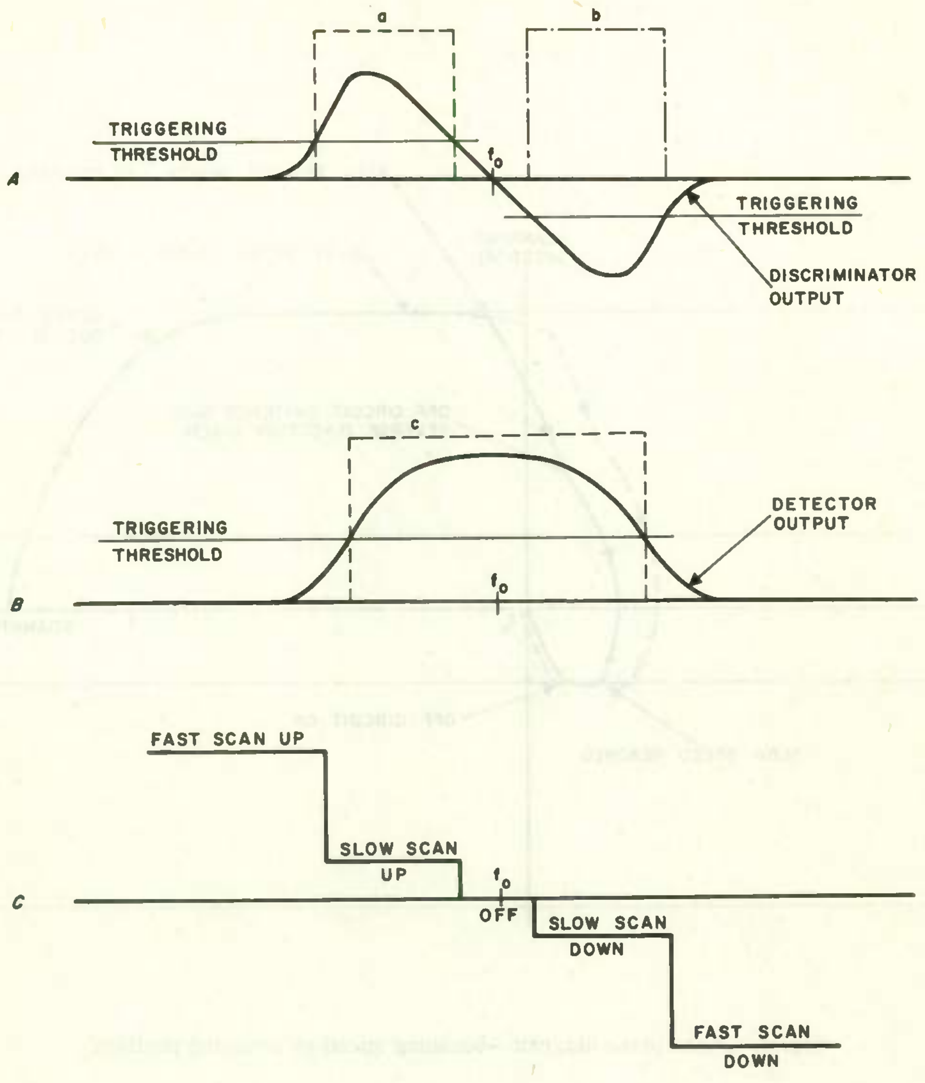


Fig. 3. Lock control characteristics.

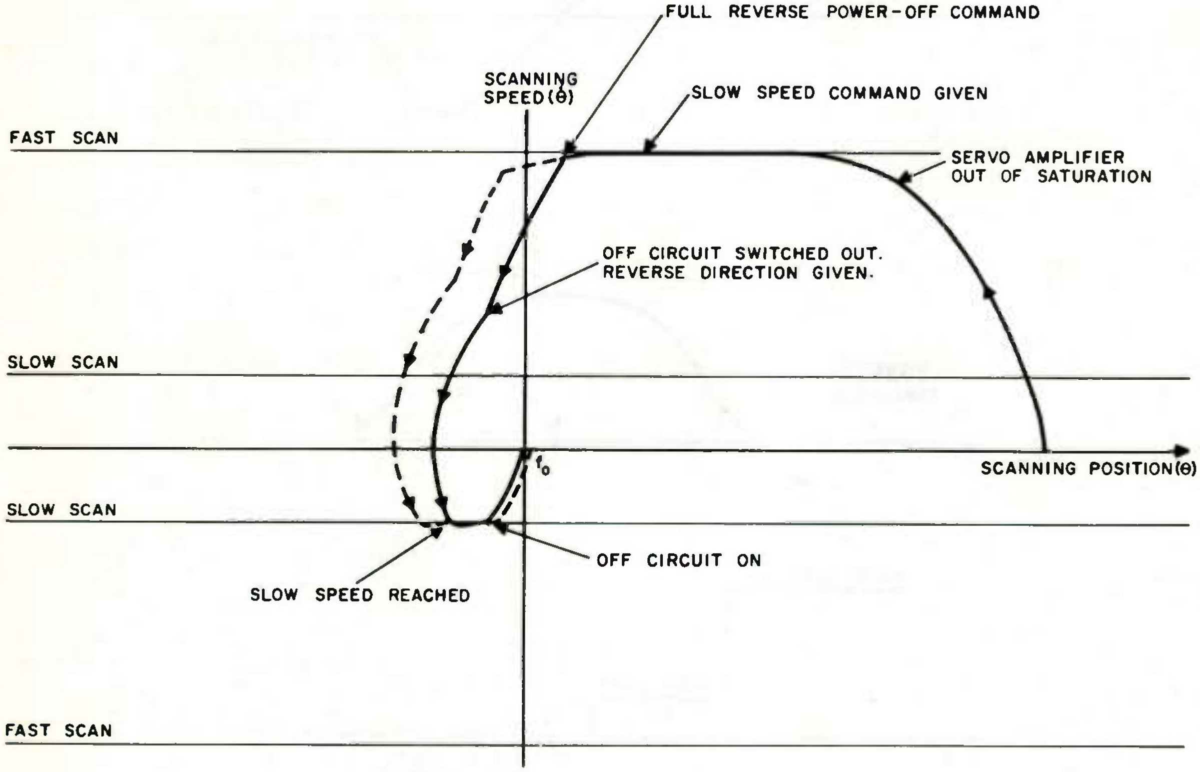

FULL REVERSE POWER-OFF COMMAND
SCANNING SPEED($\dot{\theta}$)
SLOW SPEED COMMAND GIVEN
FAST SCAN
SERVO AMPLIFIER OUT OF SATURATION
OFF CIRCUIT SWITCHED OUT. REVERSE DIRECTION GIVEN.
SLOW SCAN
$t_0$
SCANNING POSITION($\theta$)
SLOW SCAN
OFF CIRCUIT ON
SLOW SPEED REACHED
FAST SCAN

Fig. 4. Phase plane diagram—scanning speed vs scanning position.

# DESIGN OF THE SATURN S-IV STAGE
# PROPELLANT UTILIZATION SYSTEM

D. J. Allen
and
L. G. Bekemeyer
Douglas Aircraft Company, Inc.
Santa Monica, California

## Summary

This paper indicates the requirements for closed-loop propellant utilization control on the S-IV stage of the Saturn launch vehicle. An analysis of the system is presented. The design of the capacitance sensors and the electronics assembly is described.

## Introduction

The S-IV vehicle, which is presently being designed and manufactured by Douglas Aircraft, is the second stage of the initial, or C-1, configuration of the Saturn booster. It is powered by six 15,000 pound Pratt and Whitney rocket engines which use liquid hydrogen and liquid oxygen as propellants. The total propellant load is 100,000 pounds divided in the ratio of 5 pounds of oxygen to 1 of hydrogen. With this load the nominal burning time is 467 seconds. The C-1 launch vehicle is designed to be capable of orbiting a satellite weighing more than 20,000 pounds.

If this vehicle is to reach its maximum capability it must be able to burn almost all of the propellant which has been loaded. This essentially means that when one propellant has been depleted the amount of the other propellant remaining, which is an unusable residual, must be small. A typical open-loop engine mixture ratio history is shown in Figure 6. This curve shows the ratio of the propellants being burned (lbs. of oxygen per lb. of hydrogen) as a function of flight time with the assumption that all the parameters that influence this ratio are at their predicted values. Also shown is a band over which this mixture ratio can vary if these influencing parameters vary to their limits. It will be noted that during most of the flight the nominal mixture ratio is near the engine manufacturers design value of 5:1. The slow increase in ratio during flight is due primarily to the gradual warming up of the hydrogen, thus reducing both its density and the mass being pumped into the engine.

Previous propellant utilization practice, on kerosene fueled ballistic missiles, has been to attempt to determine this nominal curve by analysis and monitoring of numerous test flights. The operational vehicles would then be loaded so that if this nominal curve were followed, simultaneous depletion would result. The possible open-loop errors were small enough to be accepted.

There are several reasons why this method would not be satisfactory on the S-IV vehicle. First, because of the cost of the vehicle the number of test flights will be limited to a number insufficient to predict an accurate nominal curve. Second, as a result of the use of hydrogen as a fuel the open-loop mixture ratio band is significatly wider. In fact, these engine mixture ratio errors, combined with a reasonable loading error, could result in as much as 3,000 lbs. residual propellant.

On the S-IV there is a loss of about 1.1 pound of payload for each pound of unexpended propellant. Therefore, a 3,000 lb. propellant residual would result in a loss of 3,300 lbs. of payload. For this reason a requirement for closed-loop control of propellant utilization was established. For this purpose a system would be installed in the vehicle which would continuously senser the amount of each propellant remaining in the tanks and regulate the engine mixture ratio to insure near-simultaneous depletion of both propellants.

## System Requirements

The requirements and functions of the Propellant Utilization (PU) system on the Saturn S-IV vehicle are as follows:

1. To provide sufficient flow control to deplete both propellants to 500 pounds or less while maintaining the Engine Mixture Ratio (EMR) to 5 pounds Lox per pound $LH_2$ ± 10% ("Lox" and "$LH_2$" are terms used repeatedly in this paper for Liquid Oxygen and Liquid Hydrogen).

2. To control the loading of propellants by providing the ground support equipment with an accurate indication of the propellant masses.

3. To provide propellant mass information during flight for telemetry.

4. To provide signals for propellant depletion logic, and for the fuel tank pressurization system.

## System Analysis

### System Operational Outline

The PU system as shown in Figure 1 consists of capacitance sensors for measuring propellant masses, a summing device for comparing mass signals, a shaping network for periodic disturbance attenuation, and six valve assemblies for changing Lox flow as a function of mass error. Since the sensor measures fluid mass, it can be represented by a gain C. (see Figure 2) It is in one leg of a servo balanced bridge. When unbalanced by a propellant mass change, it is rebalanced by the servomotor and feedback potentiometer. In rebalancing the bridge, the motor position also provides mass signals for PU valve actuation, loading, mass telemetry, and switching. The generalized bridge assembly closed loop transfer function is:

$$\theta_C \longrightarrow \boxed{\frac{K\ K_M G_B}{S(T_M S + 1 + K_M K_V) + K\ K_M G_B K_P}} \longrightarrow \theta_O$$

Its damped frequency is 42 rad. per sec. with a damping ratio of .7. Velocity feedback was necessary to prevent potentiometer damage from limit cycling. Its effect is shown analytically in Figure 3. Without velocity feedback, limit cycling occurs at a frequency of 6 cps. With velocity feedback it does not exist. A bridge of this type is used with each sensor, and the difference of their output signals is the mass error.

The valve positioning loop for each of the six engines is also described in Figure 2. The closed loop transfer function for this loop is:

$$\theta_C \longrightarrow \boxed{\frac{K_{MV} K_V G_V}{S(T_C S + 1)(T_M S + 1) + K_{MV}\ K_V G_V K_P}} \longrightarrow \theta_V$$

Its damped frequency is 25 rad. per sec. with a damping ratio of ≈ .8. A nyquist stability plot of the valve loop is also shown in Figure 3. It can be seen that limit cycling is no problem in this loop.

The valve lag shown in Figure 2 has a bandpass of 10 cps. Since the engine mixture ratio is defined as Lox flow divided by $LH_2$ flow, valve gain can also be expressed in terms of percent EMR change per degree of valve. Total valve travel is mechanically limited to ± 60°. Total EMR change for ± 60° valve is 4.5 to 5.5 or ±10%.

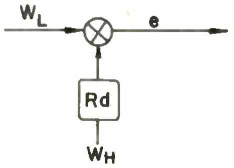
The shaping network transfer function is also shown in Figure 2. Although primarily designed to attenuate slosh, it also is used to provide desired system performance and stability characteristics. The band pass of this network (.01 rad/sec) is much lower than that of any other element in the major loop.

An open loop frequency response plot for the PU system is given in Figure 4 for normal (.2 #/sec per pound Lox error) and initial (.04 #/sec/#) system steady state gain. (The gain change will be explained later in detail.) In obtaining this frequency response information, it was found that no measurable change occurred in the area of interest when the bridge loop, valve loop, and valve dynamics were replaced with their respective steady state gain. This conclusion is substantiated by the relative "Root to Origin" distances observed on the Root Locus Plot in Figure 5.

System Flow Equations

As was stated in the previous section, the $LH_2$ bridge output signal which represents $LH_2$ mass times the desired tank mixture ratio is subtracted from the Lox bridge output signal. This difference is defined as mass error.

Symbolically, mass error $e = W_L - RdW_H$



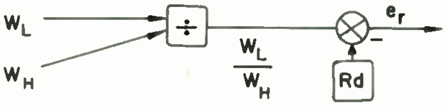
where $W_L$ = Lox mass

$W_H$ = $LH_2$ mass

Rd = 5 = desired tank mass ratio

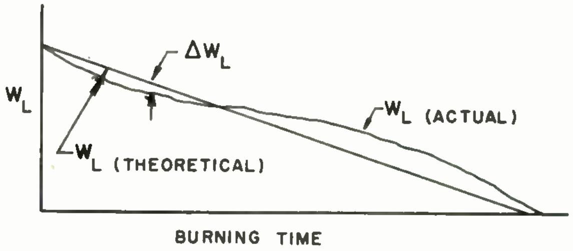
Another method considered for comparing propellant masses is a ratio error signal.

Ratio error $e_r = \frac{W_L}{W_H} - Rd = \frac{W_L - RdW_H}{W_H}$



The gain of this type of system increases to infinity as $W_H \rightarrow 0$. Primarily because of this gain change characteristic, the ratio error system was not used.

The actual propellant weights $W_L$ and $W_H$ can be divided into a theoretical weight plus a weight pertubation.



If $W_L = W_{LT} + \Delta W_L$

and $W_H = W_{HT} + \Delta W_H$

then $e = W_{LT} + \Delta W_L - Rd\ (W_{HT} + \Delta W_H)$

Where $W_{LT}$ and $W_{HT}$ represent the theoretically correct values of propellant mass. $W_L$ and $W_H$ represent mass pertubations due to tolerances and distrubances.

Since $W_{LT} - Rd\ W_{HT} = 0$, $e = \Delta W_L - Rd\Delta W_H$. This equation can be expanded to include those disturbances which make up $\Delta W_L$ and $\Delta W_H$.

$$e = \Delta W_{Li} + \Delta W_{LS} \sin \omega t + \Delta W_{LA} + \int (\Delta \dot{W}_{LU} - \Delta \dot{W}_{LC})\ dt$$

$$- 5 \left[ \Delta W_{Hi} + \Delta W_{HS} \sin \omega t + \Delta W_{HA} + \int (\Delta \dot{W}_{HU} - \Delta \dot{W}_{HC})\ dt \right]$$

$\Delta W_{LC}$ and $\Delta W_{LU}$ are flow changes from the propellant utilization flow control valve.

$\Delta W_{Li}$ and $\Delta W_{Hi}$ are initial loading errors. These errors result from loading inaccuracies, boiloff variations during boost phase, and variations in the amount of engine prestart cooldown propellants used. Since the same capacitance sensors are used for loading and propellant utilization, the loading errors as seen by the PU system will probably be less than the actual loading errors. This difference would be in the sensor inaccuracy. A maximum loading error equivalent to 750# Lox was used to study system performance.

$\Delta \dot{W}_{LU}$ and $\Delta \dot{W}_{HU}$ represent uncontrolled flow rate errors. These errors are caused principally by propellant temperature and tank ullage pressure variations. The total effect of these variations on engine mixture ratio is shown in Figure 6.

$\Delta W_{LS}$ and $\Delta W_{HS}$ are sloshing disturbances. The sensor will give erroneous propellant mass information if the fluid surface is not perpendicular to the vehicle centerline. Sloshing is a fluid surface tilt which is periodic and can be attenuated in the system electronics. The equations used in this analysis describing slosh are $\Delta W_{LS} = 400 \sin 2.5t$

$\Delta W_{HS} = 450 \sin 2.0t$

Thus $e$ (slosh only) $= 400 \sin 2.5t - 5 \times$

$450 \sin 2t$

$= 2650$ maximum

These sloshing errors, being primarily from very low damped $LH_2$ are expected to exist to some extent for the total burning period. The shaping network shown in Figure 2 is designed primarily to reduce this disturbance by 53 db at the $LH_2$ slosh frequency. Thus valve movement is reduced to ± 3° or ± .025 EMR.

$\Delta W_{LA}$ and $\Delta W_{HA}$ represent non-periodic errors from fluid surface tilt. Any time the resultant total thrust vector is not parallel to the vehicle centerline, a fluid surface tilt condition results. If, for example, the vehicle C.G. through which the thrust vector must act were not on the vehicle centerline, a surface tilt condition would occur.

## Gain Change Requirement

For total system response analysis, the open loop transfer function can be stated:

$$KGH = \frac{K_L\ (T_2\ S + 1)}{S(T_3\ S + 1)\ (T_1\ S + 1)}$$

Where $K_L$ + Total Loop Gain

From this, the velocity constant $C_v$ can be obtained.

$$C_v = \lim_{S \to o} S(KGH)$$
$$= K_L = .2$$

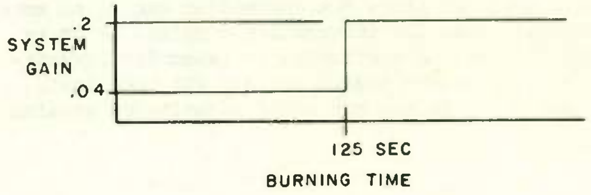
The theoretical residual from uncontrolled rate errors $(\Delta W_{LU})$ then is:

$$\frac{\Delta \dot{W}_{LU}}{C_v} \quad \text{or} \quad \frac{\Delta \dot{W}_{LU}}{.2}$$

The maximum EMR variation in Figure 6 can be approximated with a value of 5 ± .3. In units of $\Delta \dot{W}_{LU}$, +.3 EMR = 10.8#/sec. Therefore, $\frac{\Delta W_{LU}}{.2} = \frac{10.2}{.2} = 51.6$ pounds theoretical residual. The total residual will include the effects of sensor errors and system hardware errors.

The engine manufacturer has stated that the EMR remain within 4.5 to 5.5 during burning. If valve flow pertubations $(\Delta \dot{W}_{LC})$ are added to the uncontrolled deviations $(\Delta \dot{W}_{LU})$ the EMR can be 5.8. This condition can occur if the polarity of $\Delta W_{Li}$ is opposite that of $\Delta \dot{W}_{LU}$. A condition of this type was used as one parameter in evaluating system transient performance. To keep the engine mixture ratio within the prescribed limit, it was necessary to use reduced system gain until a definite relationship between $\Delta W_{Li}$ and $\Delta \dot{W}_{LU}$ was established. To satisfy this EMR excursion limit requirement as well as keep residuals from uncontrolled rate errors to a minimum, the following gain change program was developed.

SYSTEM GAIN VS BURNING TIME



System Performance

System response data has been obtained for various system disturbances. Because of the gain change characteristic, the analog computer was primarily used. The disturbances and associated response curves shown in Figure 7 are representative of the studies made. It was assumed in obtaining these curves that $\Delta W_H = 0$, and all disturbances were introduced as lox disturbances. The relatively high bandpass of the servo balanced bridges make this a valid simplification. Lox rate errors are introduced as shown in Figure 7. It is assumed that these errors are not erratic and will generally be of one sign. The system is not expected to correct for the sudden rise in EMR late in flight. It has been established however that a positive EMR deviation indicates excess Lox flow, and the resulting mass at burnout will be $LH_2$. This mass is not expected to exceed 75 lbs.

The $LH_2$ sensor will be biased to read empty with a remaining mass of $LH_2$ in the tank. The effect of this bias can best be explained with an example.

Assume

Lox sensor accuracy = .2% of total mass

= .002 (83,333)= 167#

$LH_2$ sensor accuracy = .002 (16,667) = 33#

Condition A: If the Lox sensor read low, and the $LH_2$ sensor high, residuals from sensor error would be $W_L + W_H (5)$

$= 167 + 33 (5) = 332\#$

Condition B: If the Lox sensor read high and the $LH_2$ sensor low, residuals from sensor error would be $33 + \frac{167}{5} = 66\#$.

If the $LH_2$ sensor were biased to read 45 # low; Condition (A) above would result in 332 - 5 (45) = 107 # residuals, and Condition (B) above would result in 66 + 45 = 111 # residuals. Thus residual propellants at burnout due to sensing errors can be reduced by this $LH_2$ bias.

## Capacitance Sensors

### Sensor Requirements

The selection of a sensing system is probably the most important decision to be made in the design of a Propellant Utilization System. Ultimately, the success of the system will turn on this decision since the controller can be no more accurate than the information supplied to it by the sensors. Accordingly, an extensive investigation of liquid gaging methods was undertaken, culminating in the choice of capacitance sensing for this system. Although the capacitance gage has found extensive use in industrial and aircraft fuel gaging applications this is the first time it has been applied to a propellant utilization system, and it has been found necessary to make several "state-of-the-art" improvements in its design.

The prime requirement of sensing system is, of course, accuracy. This requirement can be conveniently subdivided:

1. Zero stability: The sensor shall indicate zero propellant at the exact time the tank is emptied. This indication must not be degraded over the expected range of vehicle environmental conditions.

2. Linearity: When there is liquid in the tank the sensor shall accurately indicate its mass.

The reason for the first requirement is clear, for a sensor zero shift will cause an error which will directly add to propellant residual. The requirement that the measured parameter be propellant mass is not as obvious. There are several reasons: first, the engines are most efficient when they are run at their calibrated mass mixture ratio; and second, mass measurements are required both on the ground, for propellant loading, and in flight for vehicle performance evaluation.

One way of determining the mass of propellant in a vehicle tank would be to determine the position of the liquid-gas interface and from this compute the liquid volume. The propellant mass could then be obtained by use of a calculated density or one which had been obtained by use of a measuring device located near the bottom of the tank. With cryogenic propellants there are a number of objections to a technique of this type. First, the interface may be fairly undefined and perturbed by boiling or sloshing; and second, there is apt to be considerable density stratification throughout the length of the tank so that a density sample taken at one place in the tank would not be representative.

These disadvantages would be overcome by a gaging system which measured mass by integrating a fluid property related to density over the length of the tank. The capacitance sensor is such a system. In this system a long capacitor is placed in the tank, parallel or nearly parallel to the tank axis. When the tank is empty the capacitance of the unit will be proportional to the dielectric constant of the gas between the plates, which we will call $E_G$. If the tank is filled and the liquid allowed to flow between the plates, the observed capacitance will increase because some of the gas has been replaced by a liquid with a higher dielectric constant. If the capacitor is of uniform cross-section, its capacitance increase will be proportional to both the immersed length and the liquid dielectric constant minus the gas dielectric constant ($E_1 - E_s$). Since the dielectric constant of all gases is very nearly one, this last proportionality constant can be regarded as ($E_L$ - 1).

If we know $E_L$, we now have a system which will tell us the level of liquid in the tank. Since the capacitance of the unit is inversely proportional to the distance between its electrodes, we can build a sensor which will give a direct volume readout in a non-uniform cross-section tank by making the spacing between the electrodes at any point a function of the tank cross-section at that point. If ($E_L$ - 1) of the liquid being gaged is proportional to density, as it very nearly is for our propellants, we will have a system which will give a direct mass readout.

The dielectric constant of liquid hydrogen and oxygen is quite closely described by the Clausius-Mosotti equation.

$$\frac{E - 1}{E + 2} = KP$$

Where E = dielectric constant

P = density

K = constant dependent on the polarizability of the material involved.

It can be seen that if E is near 1 that E - 1 is almost a direct function of P. Since the dielectric constant of liquid hydrogen is about 1.22 and that of liquid oxygen about 1.48 this condition is fulfilled.

## Sensor Design

The sensors used in Saturn S-IV are cylindrical with an outer diameter of two inches. Correction for tank geometry is made by varying the diameter of the inner electrode. The mounting method is shown in Figure 1. It can be seen that because of shape of the tank it has been necessary to tilt the hydrogen sensor 18° from the vehicle axis. This sensor is 260 inches long and is, as far as the author knows, the longest capacitance sensor which has been built to date.

Several second order error sources exist in a practical capacitance gaging system. One of these arises from making the electrodes of materials which expand with temperature. In any system in which the sensors operate in a varying temperature environment the resulting expansion will cause a capacitance change which must be taken into account. The capacitance of a cylindrical capacitor is given by the formula:

$$C = KE \frac{L}{\text{Log } \frac{R_o}{R_1}}$$

Where L = length of the capacitor

$R_o$ = radius of outer electrode

$R_1$ = radius of inner electrode

E = dielectric of media between electrodes

K = constant

For a capacitor with both electrodes of the same material Log $R_o/R_1$ will not change as the capacitor expands and contracts, so the capacitance change will be proportional to the change in L. By making the electrodes out of materials with different expansivities, it is possible to make Log $R_o/R_1$ vary proportionally to L and thereby make a unit whose capacitance is very nearly constant over a wide temperature range. This is illustrated in Figure 8.

The sensor zero capacitance, or its capacitance when all the propellant in the tank has been expended is influenced by the temperature and pressure of the residual gas. Of course, this effect will be predicted and allowed for in calibrating the system. However, any error in this prediction will be seen by the system as a zero shift. In the liquid oxygen tank, where the pressurent is helium, which has a very low dielectric constant, this effect is negligible. In the hydrogen tank the residual is hydrogen gas under several atmospheres pressure and its effect can be quite appreciable. Even if we are able to make a reasonably accurate prediction of this effect, the resulting zero shift can be as much as .3% to .4% of full scale. Fortunately it is possible to compensate for the major portion of this error, the part that is due to the uncertainty in the temperature of the residual gas. This is done by designing the sensor so that its change in capacitance with temperature, due to thermal expansivity, balances out the change in ullage gas dielectric with temperature. An example is shown in Figure 9.

It can now be seen that the design of these sensors was a complex process. Since the prime requirement was zero stability this was attacked first. A computer program was written which would design a sensor which had a capacitance vs. liquid level height function which matched the tank volume vs. height function. For simplicity, it was assumed that the gas above the liquid was at a fixed nominal temperature and that the sensor electrodes were at the temperature of the fluid medium in which they were immersed. This program was then used to design a number of sensors with various electrode material combinations and different ratios of the inner and outer electrode diameters. The response of each of these designs when immersed in gas at various temperatures and pressures was determined with the aid of the computer. Based on this information, the material combination and diameter ratio which gave the smallest capacitance change over the range of gas temperature and pressures expected in the tank at the end of flight was selected for the final sensor design. The optimum material combination was an aluminum outer electrode and a stainless steel inner electrode for both the hydrogen and oxygen sensors. However, the diameter ratios were different.

Having determined these parameters, a more refined computer program was developed for designing the actual vehicle sensors. This program was similar to the first one, with provision for a variable temperature distribution in the gas and in each of the electrodes. In addition, the computer was enabled, by means of an iterative process, to vary the shape of the inner electrode in order to bring the sensor capacitance vs. liquid level curve, considered under the predicted vehicle environment, into correspondence with the tank mass vs. liquid level curve. Further, the program was arranged so the final print out of a capacitance vs. sensor length function would assume room temperature electrodes and dielectrics to simplify fabrication and calibration.

## Electronics Assembly

The electronics assembly incorporates the circuitry which supplies propellant mass signals to the ground loading computer and the vehicle telemetry systems. It also creates and shapes the system error signal and generates the electrical commands for positioning the engine mixture ratio valves.

Since this assembly had to be located at some distance from the sensors, a three-wire bridge was chosen for sensor readout. Further, to obtain the necessary accuracy and stability, a balanced bridge incorporating a servo rebalance loop was used. This circuit is shown in Figure 10 and is typical for both the hydrogen and oxygen sensors.

In this circuit the sensor forms one leg of a bridge, the opposite leg is a fixed reference capacitor. The other two legs of the bridge are voltage sources supplied by the secondary of the reference transformer. The output of the bridge is the input to a servo amplifier. If the bridge is initially in balance and the sensor capacitance is increased or decreased by adding or withdrawing propellant, the bridge will be unbalanced and an input supplied to the amplifier. The amplifier in turn will drive a servomotor which repositions the rebalance potentiometer to return the bridge to null. The rebalance potentiometer voltage, or shaft position, change is proportional to sensor capacitance change. This shaft position is the output of the bridge.

It will be noted that the capacitance to ground associated with cable connecting the sensor to the electronics assembly does not add to the stray capacitance of the sensor. This is a particular advantage of this form of bridge since this capacitance can easily be larger than the sensor capacitance. However, it is important to avoid capacitor coupling between the two sensor leads since this cannot be distinguished from sensor capacitance. This requirement is met by using shielded cable and coaxial connectors for the high impedance lead.

The rebalance pot is one gang of a four-gang ten-turn potentiometer. The other three gangs are used for loading and course telemetry, fine telemetry, and forming the system error signal. The loading potentiometer is excited with 28 volts DC and the voltage ratio output taken to a digital ratiometer in the ground loading computer. The ratiometer is calibrated to give a readout of the pounds of propellant in the tank and to provide signals for operating the loading valves.

The fine telemetry potentiometer has been incorporated to enable an accurate inflight determination of propellant mass. The accuracy of this measurement would normally be limited by inherent telemetry inaccuracies amounting to about 2% of full scale. To overcome this, the potentiometer is divided into 20 equal segments by tapping; alternate taps are excited by 5 volts DC, with the remainder grounded. This effectively provides a 20 times expanded scale and reduces the telemetry errors to .1%.

Figure 11 shows the method of forming the error signal. The two bridge output potentiometers are excited in parallel from a 100 volt DC source. The bridges have been calibrated so that any time the propellant masses in the tank are at the desired 5:1 ratio the potentiometer wiper positions and therefore wiper voltages are equal. Accordingly, in operation the voltage difference between the wipers is the system error signal. In order to have this signal referred to ground the potentiometer excitation supply is floating and the wiper of the hydrogen output potentiometer is grounded.

The error voltage is first taken to a switched voltage divider which mechanizes the gain change. The gain change switch is driven by the oxygen bridge servo. A RC network is used to shape the error signal. The high attenuation desired in this filter to reject disturbances from propellant slosh and the need to keep the filter capacitors to reasonable size has resulted in the filter presenting a high source impedance to the amplifier which follows it. Efforts to design a sufficiently stable DC amplifier to work from this source were unsuccessful, so a modulator-AC amplifier-demodulator combination was used.

The demodulator output is the command input to six parallel position servos which are used to control the engine mixture ratio valves. DC is used for the command and feedback signals to avoid quadrature problems which would arise from summing two slightly out-of-phase signals. The command and feedback signals for each loop are summed at the input of a magnetic modulator. This modulator is followed by an AC amplifier which drives a servomotor located on the engine.

SATURN S-IV PROPELLANT UTILIZATION SYSTEM

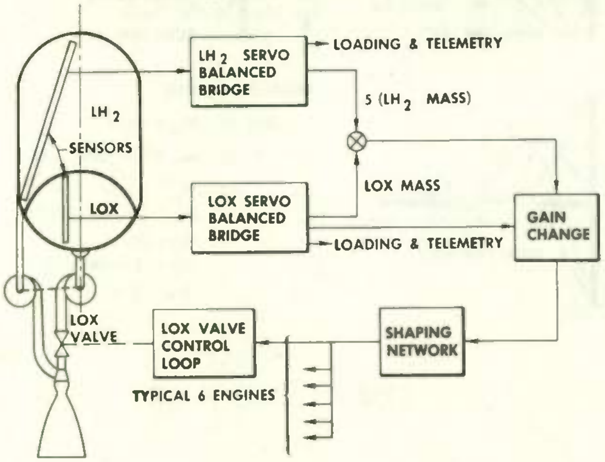


Fig. 1.

PROPELLANT UTILIZATION SYSTEM BLOCK DIAGRAM

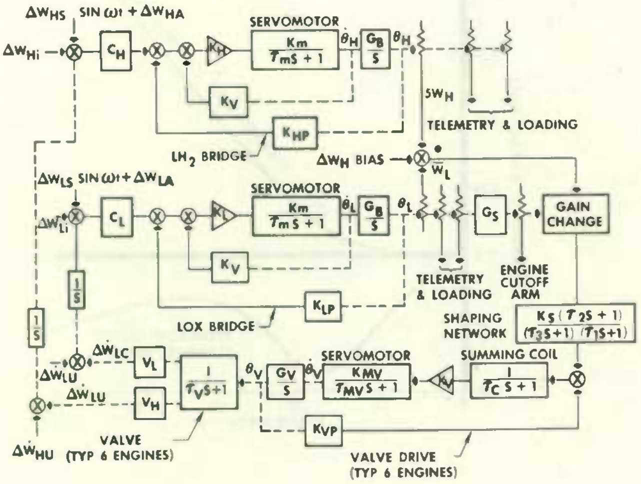


Fig. 2.

BRIDGE & VALVE DRIVE OPEN LOOP FREQUENCY RESPONSE

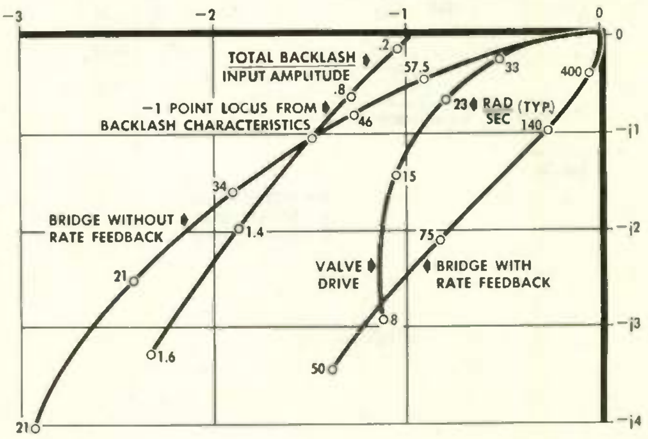


Fig. 3.

PROPELLANT UTILIZATION SYSTEM OPEN LOOP FREQ RESPONSE

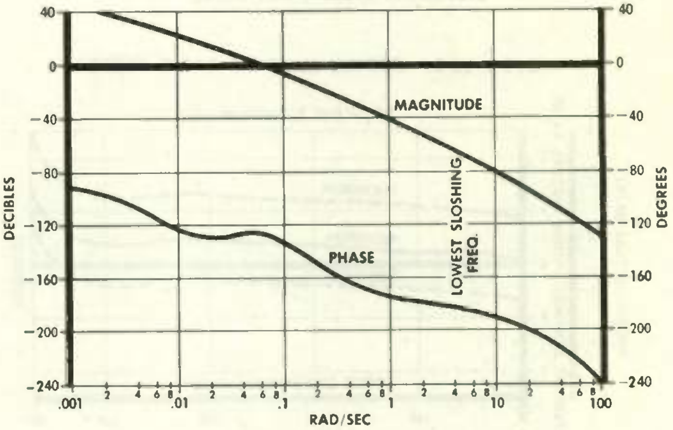


Fig. 4.

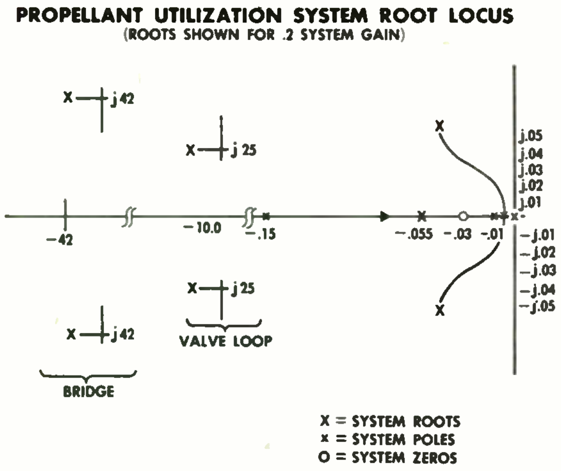


Fig. 5.

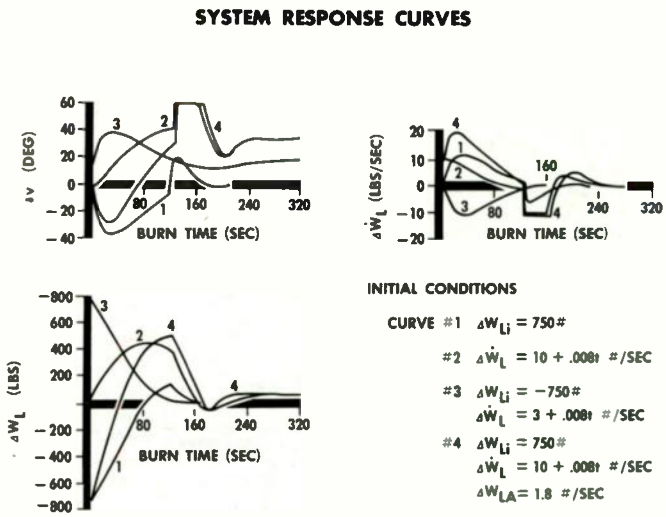


Fig. 7.

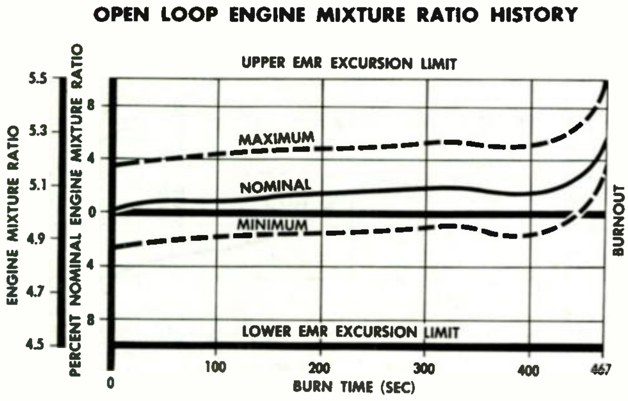


Fig. 6.

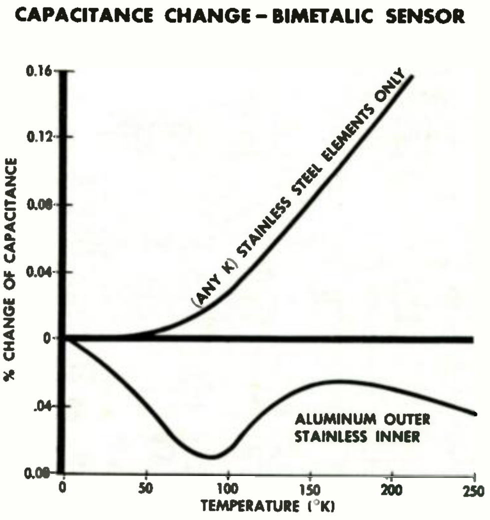


Fig. 8.

TEMPERATURE STABILIZED SENSOR

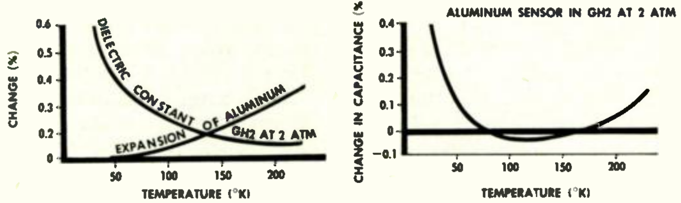


Fig. 9.

SERVO-BALANCED CAPACITANCE BRIDGE

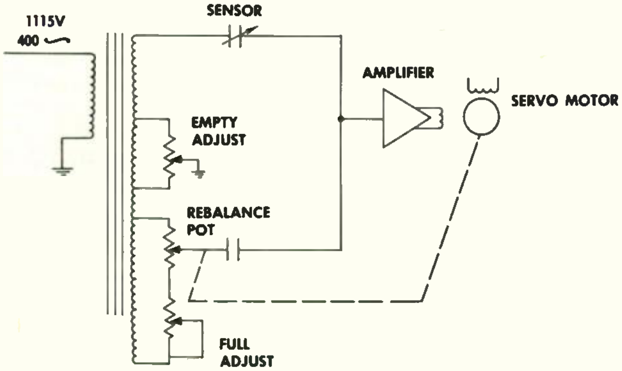


Fig. 10.

PROPELLANT UTILIZATION SYSTEM

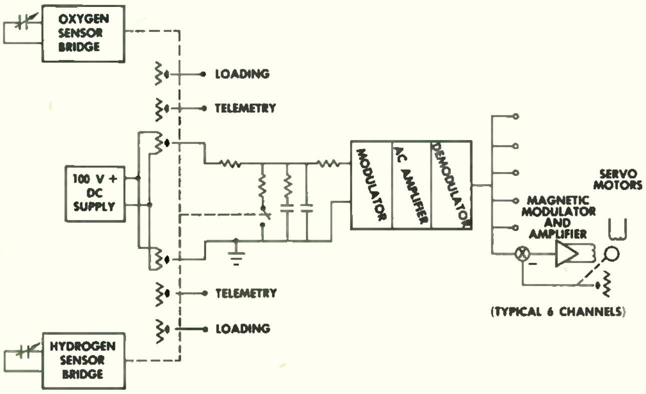


Fig. 11.

# VAPOR JET CONTROL OF SPACE VEHICLES

James E. Vaeth
Martin Marietta Corporation
Baltimore 3, Maryland

## Summary

This paper describes a reaction jet attitude control technique which affords significant advantages in terms of accuracy, reliability, fuel economy and operational flexibility. These advantages are realized by the use, in combination, of low-thrust vapor jets and time-dependent on-off switching circuits. An accuracy potential comparable to inertia wheel control is thus provided, while the proverbial wheel problems of speed saturation, bearing life, threshold nonlinearities, gyroscopic coupling and vibration excitation are avoided.

Very low thrust magnitudes are attained by simply opening a small orifice to allow fuel to vaporize into the surrounding vacuum. Fuel storage, pressurization, circulation and mixing requirements are thus minimized. By augmenting conventional on-off valve switching circuitry with electronic networks that generate thrust pulses of small but constant time duration, vehicle angular rate can be controlled to a very low threshold. This minimizes fuel consumption and valve cycling frequency.

The capabilities and limitations of this design approach were substantiated by an analog computer program incorporating breadboard switching circuits, and by vacuum chamber testing of critical components. These technique and component developments are applicable to such space missions as astronomical observation, earth reconnaissance and stellar navigation. Design guides are presented for synthesizing a reaction jet system to meet any particular set of performance specifications.

## Introduction

The analytic studies, system design approach and test programs to be described were initiated during the design competition for the NASA Orbiting Astronomical Observatory (OAO)[1, 2]. The very stringent criteria for vehicle attitude control, in terms of pointing accuracy, reliability and operational flexibility, dictated significant improvements over existing techniques and equipments.

Preliminary studies verified that the required accuracy of 0. 1 arc seconds could be attained by means of proportional inertia wheel control[3]. However, the questionable reliability associated with wheel bearings operating continuously for a year[4], plus the definite need for preventing wheel rate saturation (such as by firing auxiliary jets once each orbit) were the major factors leading to an extensive investigation of the capabilities and limitations of reaction jets for fine pointing control.

Although on-off reaction jets had been successfully employed in prior missile and space programs, their applicability for precision attitude control of the OAO vehicle called for much lower angular momentum impulses than had ever been utilized. The low thrust vapor jets, on-off switching techniques and system design approach developed to meet these specific requirements, together with pertinent test results and growth potential, are presented in the pages that follow.

## System Design Requirements

The significant design requirements[1] for fine pointing control of the OAO vehicle may be summarized as follows:

Continuous control of vehicle orientation to within the sensing accuracy of the primary optical telescope, which was specified to be 0. 1 arc seconds, or better, about two axes;

Capability of reducing an initial pointing error of 120 to less than 0. 1 arc seconds within a 3 minute time duration;

Operational compatibility with auxiliary star trackers, possessing resolution limitations of 10 arc seconds or worse, during periods when the primary telescope is occulted by the earth or moon;

Ability to cope with external disturbing torques exceeding 100 dyne centimeters;

Sufficient flexibility to cope with optical or instrument noise associated with the stellar trackers, and with saturation of the stellar detector signal*;

Operational flexibility to perform all required functions with a minimum of mode changing; and

Operating duration of at least one year.

Another implicit design consideration was to limit system size and weight, including fuel, to a few pounds per axis. The reliability concern stemming from the one-year operating life emphasized the need to minimize moving parts. Accuracy considerations also dictated minimizing internal motions and associated vibration excitations. An important design criteria, therefore, was to keep the jet valve cycling rate as low as possible--consistent with the performance requirements.

In attempting to meet the combined design requirements of precision performance, relia-

*This would ease sensor design requirements.

bility and early availability, the obvious approach was to utilize proven techniques and components, wherever applicable, and to substantiate the performance capabilities of any novel or critical items by system analysis and component testing.

The major problem areas were the required low thrust jet units and the electronic switching circuits needed to minimize thrust on-time. Although this paper is primarily concerned with the switching techniques, some discussion of vapor jet thrust generation is warranted because the two are closely interdependent.

## Vapor Jet Thrust Generation

The jet control moment should ideally be no greater than that needed to satisfactorily accomplish the initial damping function (120 arc seconds within 3 minutes). For the OAO vehicle, the desired control torque about each axis is subsequently shown to be between 1,000 and 20,000 dyne centimeters, or less than 0.00015 foot-pound.

The desired low thrust levels made the use of vapor jets very attractive. Accordingly, a vacuum chamber test program was undertaken to determine the thrust and specific impulse of various vapor fuels as functions of orifice size and shape, pressure and temperature differential, etc. Detailed test procedures and results are presented in Ref. 2. Pertinent conclusions are as follows:

A specific impulse of 50 to 100 seconds was measured for such fuels as water and methyl alcohol;

The desired low thrust levels were attainable by using the proper orifice diameter;

Because of the small orifice size (0.02-inch diameter), a single-level on-off system appeared necessary, as opposed to proportional control of thrust; and

Thrust variations with internal temperature were such that system operation should be compatible with a 20% uncertainty in absolute thrust level.

## On-Off Switching Technique

The electronic circuits required for each axis of control must position a solenoid valve in accordance with the optical error signal. The three valve positions are closed, open-left and open-right.

Analytic studies to determine the required switching circuitry began with an evaluation of the conventional and proven technique[5] for accomplishing all the functions outlined under "System Design Requirements." This technique keeps the jet valve open whenever the sum of the measured attitude displacement and rate signals exceeds a preset voltage (equivalent to an error voltage of 0.1 arc second).

Analysis procedures were initiated making use of phase plane techniques[5] and culminated in an analog computer simulation to investigate the effects of disturbing torques, optical noise and switching hysteresis inherent in the breadboard circuits. In the analog program, additional breadboard circuits were provided for generating a small but constant thrust impulse, in case such a refinement proved necessary.

Figure 1 is a functional block diagram of the analog computer simulation (see List of Symbols). The jet valve opens for a fixed time duration ($T_p$) when the applied switching signal ($\theta_s$) reaches a preset value ($\theta_d$ equivalent to 0.1 arc seconds) and keeps the valve open whenever the error exceeds twice the pulsing level ($2\,\theta_d$). The valve closes when $\theta_s$ reduces below $2\,\theta_d$, and opens in identical fashion--but in the opposite direction--for $\theta_s$ signals of reverse polarity. When $T_p$ is set at zero and the second switching level is reduced to 0.1 arc seconds, operation of a conventional-type system is simulated.

As noted in Fig. 1, a derivative or lead circuit ($\alpha\, T_L$) is employed to generate attitude rate signals. Rate gyros were avoided for reasons of reliability and the required threshold level. The filter is incorporated to attenuate optical noise.

The capabilities and limitations of the system shown in Fig. 1, with $T_p$ finite and zero, are presented in subsequent pages for the limit cycle phase of operation and for the initial damping phase. Both phases strongly influence overall system design.

## Initial Damping Phase

The initial damping requirement of 120 arc seconds within 3 minutes can be met by various combinations of control acceleration ($\ddot{\theta}_c$) and rate-to-displacement gain ratio ($\alpha\, T_L$ in Fig. 1). A critical factor in selecting $\ddot{\theta}_c$ and $\alpha T_L$ is the linear range of the optical error detector. If the linear range is ±60 arc seconds or greater, the recommended scheme is to use the minimum $\ddot{\theta}_c$ which satisfies the following two criteria:

$\ddot{\theta}_c$ is sufficient to rotate the vehicle 120 arc seconds ($\theta_{mi}$) in 3 minutes ($t_m$) by applying positive $\ddot{\theta}_c$ for 1.5 minutes and negative $\ddot{\theta}_c$ for 1.5 minutes,or

$$\ddot{\theta}_c \geqq 4\,\theta_{mi}/t_m^{\,2} \qquad (1)$$

$\ddot{\theta}_c$ is an order of magnitude greater than any disturbing moment.

Having thus chosen $\ddot{\theta}_c$, $\alpha\tau_L$ is selected so that

$$\alpha\tau_L = (\theta_{mi}/4\ddot{\theta}_c)^{1/2} \qquad (2)$$

The above technique for selecting $\ddot{\theta}_c$ and $\alpha\tau_L$ minimizes thrust magnitude, assures compliance with the $t_m$ damping requirement for any initial attitude error less than $\theta_{mi}$--and requires no circuit complication.

When $\theta_i < \theta_{mi}$, decreased fuel consumption can be realized by computing an optimum switching function from measured rate, displacement and an assumed value of $\ddot{\theta}_c$. Thrust polarity would be switched when

$$\theta = -\dot{\theta}^2/2\ddot{\theta}_c \qquad (3)$$

However, this scheme was impractical because of $\ddot{\theta}_c$ uncertainty.

Fuel usage, which is proportional to the product of $\ddot{\theta}_c$ and jet on-time, may be reduced by means of rate limiting ($\dot{\theta}_m$), but again, at the expense of circuit complication. Using rate limiting (such as 1 arc sec/sec) the criterion for selecting $\alpha\tau_L$ becomes

$$\alpha\tau_L = \dot{\theta}_m/2\ddot{\theta}_c \qquad (4)$$

For the practical case, with error detector saturation ($\theta_m$), angular rate measurement by a derivative circuit is limited to within the detector linear range. Thus the use of rate limiting would require a low threshold rate gyro, with associated reliability degradation. A more attractive solution would increase $\ddot{\theta}_c$ and decrease $\alpha\tau_L$, such that

$$\alpha\tau_L \cong (\theta_m/2\ddot{\theta}_c)^{1/2} \qquad (5)$$

However, $\ddot{\theta}_c$ must now be chosen sufficiently large to attenuate $\theta_{mi}$ within the specified $t_m$. This is illustrated in the phase plane plot of Fig. 2 for $\theta_m$ of 20 arc sec, $\theta_{mi}$ of 120 arc sec, $\ddot{\theta}_c$ of 0.227 arc sec/sec$^2$ and $\alpha\tau_L = 6.7$. The required time duration is approximately 3.0 minutes. It is noteworthy that, for the saturation case, the necessary $\ddot{\theta}_c$ is primarily dependent upon $\theta_m$, whereas the use of Eq (5) in selecting $\alpha\tau_L$ contributes much less toward minimizing $\ddot{\theta}_c$ than does Eq (2) without detector saturation.

Should the detector saturation level be less than ± 5 arc sec, the required large $\ddot{\theta}_c$ significantly increases fuel usage and causes a severe limit cycle complication, as will be shown. Alternative solutions are as follows:

Incorporation of low threshold rate gyros, together with rate limiting and the use of Eq (4).

Use of time-counting circuits to switch the jets on and off during the initial damping phase.

The first has already been discussed and is the more desirable in terms of fuel consumption, circuit simplicity and low $\ddot{\theta}_c$. The second circumvents the need for extremely reliable and precise rate gyros, by incorporating time counting circuits whose function is illustrated in Fig. 3. Note that an essentially steady-state limit cycle (denoted by trace ABCD) will result with the switching levels set at ±0.1 arc sec and without the lead circuit.

The counting circuit simply measures jet on-time ($t_o$) for one half cycle (point A to B) and then during the next half cycle reverses thrust after 85.35% of $t_o$. This occurs at point E of Fig. 3, after which the reversed thrust is maintained for 35.35% of $t_o$. At this time, angular rate and displacement reach zero simultaneously (point F), and the normal jet switching circuits (including lead) are reactivated to maintain precise accuracy.

This technique does not require knowledge of vehicle attitude, attitude rate or jet thrust magnitude. It requires only that the positive and negative $\ddot{\theta}_c$ be equal to within about 2%. To alleviate this contingency, $\ddot{\theta}_c$ could be increased so that the thrust reversal periods of 85.35% of $t_o$ and 35.35% of $t_o$ would not be initiated at point C (Fig. 3), but delayed one half cycle until point A is again reached.

Although detector saturation was not defined[1], it appeared judicious to use a $\ddot{\theta}_c$ magnitude sufficiently high to be compatible with any of the above damping schemes, including $\theta_m$ = 20 arc seconds. A value of 0.226 arc sec/sec$^2$ appeared to be a good compromise, with $\alpha\tau_L$ ranging from 2.5 to 12.

### Limit Cycle Operation

Continuous tracking to within the specified accuracy of ±0.1 arc sec can be accomplished as illustrated in the phase plane plot of Fig. 4. The vehicle ideally rotates at a constant angular rate through the ± $\theta_d$ dead band (point F to A) until the switching signal ($\theta_s$ in Fig. 1) exceeds a preset level. This occurs at point B with the increment from A to B resulting from the combined system lag ($\tau_T = \tau_\theta + \tau_L + \tau_c$ in Fig. 1, assuming $\tau_c << \tau_\theta + \tau_L$). Control torque is applied at point B and--if $\tau_T$ were zero--thrust would cease at point C. However, $\tau_T$ delays thrust termination until point D.

The obvious criterion for fuel economy is to minimize angular rate ($\dot\theta_L$) through the dead band. For the more practical case, it will be shown that $\dot\theta_L$ should be no less than ± 0.01 arc sec/sec because of expected disturbing moments. Attainment of such a low limit cycle rate with a conventional system is possible, but only if switching hysteresis (H) and $\tau_T$ are very small. Because of lead circuit limitations and the possible need to smooth optical noise, compatibility with a minimum $\tau_T$ of 0.5 sec seemed very desirable.

The simplified equations derived in Ref. 5 for defining the pull-in and drop-out lines by phase plane techniques imply that the minimum jet on-time ($\tau_{on}$, from point B to D in Fig. 4) is approximately equal to $\tau_T$. This is an excellent approximation if the hysteresis delay (H) is predominant. However, when the required $t_{on}$ is less than $\tau_T$, the exponential transient cannot be neglected.

For the case in which a near constant $\dot\theta_L$ is maintained for a time interval equivalent to $3\tau_T$, thrust on-time can be determined as follows. From Fig. 4, the required incremental rate ($\Delta \dot\theta$) that must be sensed in order to signal drop-out (point B to C) is

$$\Delta \dot\theta_R \cong \frac{1}{\alpha \tau_L} (H + \dot\theta_L \tau_T) \qquad (6)$$

The measured $\Delta \dot\theta$ after thrust initiation at point B is

$$\Delta \dot\theta_M \cong \ddot\theta_c \; t_{on} \Big/ (1 + \tau_T S) \qquad (7)$$

Thrust termination requires that

$$\Delta \dot\theta_M \geqq \Delta \dot\theta_R \qquad (8)$$

Substituting the Laplace transform $\frac{1}{S^2}$ for $t_{on}$ in Eq (7) and writing the inverse transform,

$$\Delta \dot\theta_M \cong \ddot\theta_c \left[ t_{on} - \tau_T \left(1 - e^{-t_{on}/\tau_T}\right)\right] \qquad (9)$$

Note that if $t_{on} >> \tau_T$, the thrust will cease when $\Delta \dot\theta$, from point C toward point D, equals $\ddot\theta_c \tau_T$. This approximation is quite incorrect for $t_{on} \leqq \tau_T$.

The usefulness of Eq (8) as a design guide is augmented by combining the limit cycle relationship

$$2 \dot\theta_L = \ddot\theta_c t_{on} \qquad (10)$$

with Eqs (6) and (9), which gives

$$1 - \frac{\tau_T}{t_{on}} \left(1 - e^{-t_{on}/\tau_T}\right) \geqq \frac{1}{2\alpha \tau_L} \left(\tau_T + \frac{H}{\dot\theta_L}\right) \qquad (11)$$

This minimum on-time criterion for a conventional system is plotted in Fig. 5. As an example, the desired $\dot\theta_L$ of 0.01 arc sec/sec and $\ddot\theta_c$ of about 0.2 arc sec/sec$^2$ give a desired value for $t_{on}$ of 0.1 sec. Using $t_{on}$ = 0.1 sec and $\tau_T$ = 0.5 sec, Fig. 5 indicates that the right-hand side of Eq (11) must not exceed 0.095. This, in turn, is realized if $\alpha\tau_L$ is set at 5 and H does not exceed .0 045 arc sec. The obvious method of decreasing $t_{on}$ is to increase $\alpha\tau_L$; however, this accentuates the importance of the initial assumption, in the derivation of Eq (11), that $\dot\theta_L$ had been nearly constant for a short time prior to thrust initiation.

Equation (11) is not applicable if the pull-in line in Fig. 4 (point A) is crossed while torque is being applied. This situation does exist during the latter part of the initial damping maneuver. Until the system damps to within the two pull-in lines, a good approximation for minimum jet on-time is simply $\tau_T$. Thus a second criterion for satisfactory limit cycle operation is

$$\ddot\theta_c \tau_T < 2 \theta_d / \alpha\tau_L \qquad (12)$$

Consequently, with $\ddot\theta_c$ = 0.2 arc sec/sec$^2$, $\tau_T$ = 0.5 sec and $\theta_d$ = 0.1 arc sec, it follows that $\alpha\tau_L$ must be less than 2.0. Since this low value of $\alpha\tau_L$ would preclude satisfying Eq (11), $\tau_T$ must be significantly decreased.

To avoid the squeeze imposed by Eqs (11) and (12), to significantly improve the ability to smooth optical noise and to make system nonlinearities and uncertainties (such a switching hysteresis) much less critical, breadboard circuits for generating the timed pulse ($\tau_p$ in Fig. 1) were designed, tested and incorporated into the analog program.

The performance improvements attributable to the timed pulse were determined by analog computation. Response plots of a conventional system ($\tau_p = 0.$), shown in Fig. 6, substantiate the criterion of Eq (12) in that $\dot{\theta}_L$ never settles within the two pull-in lines. Note that thrust on-time is very high. Significant parameters were: $\ddot{\theta}_c = 0.226$ arc sec/sec$^2$, $\theta_d = \pm 0.1$ arc sec, $\alpha\tau_L = 2.5$ sec and $\tau_T = 0.67$ sec. By adding the $\tau_p$ pulse of 0.1 sec when $\theta_s$ reaches ±0.05 sec, the system not only damps to within the two switching lines, but settles onto a limit cycle with an accuracy better than 0.05 arc sec and a jet on-time to off-time ratio of only 1/60 (Fig. 7).

An excellent demonstration of the significance of $\tau_T$ and $\alpha\tau_L$ is obtained by comparing Figs. 7 and 8. Doubling $\alpha\tau_L$ and $\tau_T$ (Fig. 8) results in a longer settling time and the higher $\alpha\tau_L$ causes greater fuel consumption during limit cycle. The minimum thrust on-times ($\ddot{\theta}_c$ traces) achieved during initial damping, excluding the constant 0.1-sec pulses, agree quite well with the criterion of Eq (11) and Fig. 5 in that $t_{on} \leqq 0.3$ sec is realized, provided thrust has been off for a time interval of at least $\tau_T$ prior to thrust initiation. Note the pulse at t = 40 in Fig. 8. This also indicates that the hysteresis in the breadboard switching circuits amounted to less than 0.005 arc sec.

Further increase of $\alpha\tau_L$ and $\tau_T$ (above the values of 5.0 and 1.12, Fig. 8) prevented the system from settling within the two pull-in lines, despite the $\tau_p$ pulse. As suggested by Eq (12), this situation was alleviated by doubling the system dead zone--which also halved fuel consumption, while maintaining the specified accuracy.

Based on analog simulation results, the system parameters recommended to comply with the requirements outlined in "System Design Requirements" were $\alpha\tau_L = 5$ sec, $\tau_L = 0.5$ sec, $\tau_\theta = 0.5$ sec, $\tau_p = 0.1$ sec when $\theta_\epsilon = 0.1$ arc sec, $\tau_c \leqq 0.03$ sec and $\ddot{\theta}_c = 0.226$ arc sec/sec$^2$ (by using an $M_c$ of 12,000 dyne-cm with $I_v$ = 800 slug-ft$^2$).

A phase plane response plot of the recommended configuration with a constant $M_E$ of 100 dyne-cm is presented in Fig. 9. The jet on-time to off-time ratio is 1/120 or $M_E/M_C$. This ratio was verified for $M_E$ values from 50 to 300 dyne-cm. For $M_E = 0$, the on-time to off-time ratio was less than 1/100. Note that if a constant $M_E$ of 300 dyne-cm or greater should be specified as always being present, it would be desirable to increase $T_p$ to 0.3 second. This would minimize valve cycling and wear, but assure the minimum attainable (1/40) on-time to off-time ratio.

Analog traces verified that, for $M_E > 22$ dyne-cm, the limit cycle will become symmetrical about the θ axis (Fig. 8) when $T_p$ is zero or finite. With such values of $M_E$, $\theta_d$ can be increased to $[\theta_d + \dot{\theta}_L \alpha \tau_L - \dot{\theta}_L \tau_T]$. However, with lower $M_E$ values, the limit cycle will not be symmetrical--so that $\theta_d$ cannot be decreased.

The principal effect of optical noise is to increase valve cycling and fuel consumption. If the filter networks can sufficiently attenuate the noise level, the $\theta_s$ signals will not excessively energize the valve. For the recommended system parameters, response to optical noise is shown in Fig. 10. Note the additional jet pulses in the $\dot{\theta}$ trace. Analog studies demonstrated that valve response to any noise spectrum (magnitude and frequency range) can be predicted as a function of the system parameters--in particular, $\tau_T$, $\alpha\tau_L$ and $\theta_d$.

When control is switched to an auxiliary star tracker, functional operation remains the same, except that the voltage level of the detector error signal ($\theta_\epsilon$ in Fig. 1) is made compatible with the desired switching levels.

To summarize briefly, the $T_p$ pulsing technique significantly improves system design flexibility. For example, should $\ddot{\theta}_c$ have to be increased by a factor of 5 to comply with initial damping requirements or with limitations on minimum thrust level, the desired limit cycle ($\dot{\theta}_L = \pm 0.01$ arc sec/sec) could still be attained by reducing $T_p$ to 0.02 sec.

## Fuel Considerations

Required fuel weight per axis is simply

$$W_f = Ft_T/I_{sp} \qquad (13)$$

where total jet on-time ($t_T$) includes both initial damping maneuvers and limit cycle operation. $I_{sp}$ was measured as 50 sec minimum, and the jet thrust (F) required to produce 12,000 dyne-cm is less than 3 x 10$^{-5}$ pound.

For a one-year operating life, a conservative estimate of 10,000 initial damping maneuvers, each of 160-seconds duration, results in a $W_f$ of less than one pound.

For a limit cycle on-time to off-time ratio of 1/120 (assuming $M_E$ = 100 dyne-cm), continuous operation for one year would require a $W_f$ of 0.16 pound. Required fuel weight would increase directly with $M_E$.

## Component Design and Testing

The feasibility of the vapor jet design approach was substantiated by engineering design, fabrication and testing of critical components.[2] Detailed design and test data are beyond the scope of this paper, but a brief description seems pertinent.

### Jet Switching Circuit with Timed Pulse

This circuit functions as shown in Fig. 11. The detector d-c error signal, after compensation, is modulated and fed to both an amplitude-sensing and a phase-sensing channel. The latter closes either transistor switch A or B, permitting the proper jet to be energized when the error amplitude reaches one of the switching levels.

In the amplitude channel, the error signal drives two separate level sensors (diode clamps). When either of the switching levels is exceeded, the associated Schmitt trigger changes state and drives a C-R circuit so that a pulse is generated. The level 1 monostable flip-flop applies a pulse of fixed duration to the transistor switch, whereas the level 2 bistable flip-flop keeps the switch continuously energized whenever level 2 is exceeded.

A breadboard was designed, tested and operated for 100 hours (with no failures) during the analog simulation program. A development model, weighing approximately 1 pound (shown in Fig. 12), was fabricated and successfully cycled 5 million times during vacuum chamber testing of the system.

### Digital Damping Circuit

The initial damping technique of reversing thrust at 85.35 and 35.35% of measured time ($t_o$) was mechanized digitally, using appropriate counting registers and logic. The procedure is as follows: a number of pulses proportional to $t_o$ are registered during the first half cycle, with subsequent switching when 6/7 of the $t_o$ pulses (85.71%) are counted and again when 6/17 (35.29%) are counted. A developmental model weighing 2.5 pounds was fabricated and successfully tested.

### Vapor Jet Propulsion System

The vapor jet test apparatus used for experimental determination of vacuum thrust and $I_{sp}$ is shown in Fig. 13. An experimental propulsion system consisting of a jet valve and actuator, a nozzle assembly and a fuel system was used in conjunction with the switching electronics to conduct vacuum tests of thrust buildup time, valve leakage and component life.

By photographing electrode arcing response with a high speed camera, thrust risetime of 0.034 sec was measured--and most of this time was attributable to the relay lag in generating the timing light. The system was successfully cycled 3.5 million times, simulating in-orbit operation of more than one year. Subsequent leakage tests and careful inspection substantiated system feasibility.

## Potential Applications

The switching techniques and component developments described above can be applied to many space missions that require precision control of vehicle orientation.

In synthesizing a reaction jet system to meet a specified set of performance requirements, the system evolution and design guides outlined in Eqs (1) through (13) should prove very useful. In particular, the need for introducing a timed pulse circuit can be effectively evaluated from Eqs (11) and (12).

An attractive sophistication of the $t_p$ pulsing technique provides an effective range of linear control between two switching levels--one being the dead band level desired for limit cycle operation and the other, perhaps, as large as the detector saturation level.

Between these levels, thrust impulse is made proportional to the input error signal by combining pulse frequency and pulse width modulation[6]. This affords the capability for simultaneous control (using common thrust units) of vehicle translation and orientation as required for space rendezvous. This technique can provide limit cycle operation equivalent to the $T_p$ pulse system, but it possesses the potential disadvantage of increased valve cycling (and wear) during initial damping maneuvers and because of noise inputs.

## Acknowledgments

Verification and implementation of the system design approach and switching techniques described in this paper represent combined effort. In particular, R. Mayer, J. Higginbotham and L. Miklaszewicz were responsible for the electronic circuit design,

fabrication and testing. H. Burke and W. Miessner were instrumental in the analog computer simulation program. J. Miller conducted the vacuum chamber test programs and the vapor jet propulsion system design, with invaluable consultation provided by T. Hill and Dr. J. Vandrey.

## References

1. National Aeronautics and Space Administration Request for Proposal, GS-390, Orbiting Astronomical Observatory, 5 May 1960.

2. "Orbiting Astronomical Observatory," ER 11208P, The Martin Company, Baltimore, June 1960.

3. Vaeth, James E., "Flywheel Control of Space Vehicles," 1960 IRE International Convention Record and IRE PGAC Transactions, August 1960.

4. DeLisle, J. E., Hildebrand, B. M., and Petranic, I. D., "Attitude Control of Spacecraft," Astronautics, November 1961.

5. Pistiner, Josef S., "On-Off Control System for Attitude Stabilization of a Space Vehicle," ARS Journai, Vol. 29, pp 283-289, April 1959.

6. Schaefer, Richard A., "New Pulse Modulator for Accurate DC Amplification with Linear or Non-Linear Devices," IRE Transactions on Instrumentation, March, 1962.

## List of Symbols

| | |
|---|---|
| $\theta$, $\dot{\theta}$, $\ddot{\theta}$ | Vehicle angular displacement, rate and acceleration |
| $\theta_\epsilon$ | Attitude sensor error signal |
| $\theta_d$ | Attitude displacement dead band |
| $\theta_s$ | Signal applied to jet switching circuitry |
| $\theta_{mi}$ | Maximum initial error |
| $\theta_i$ | Initial condition error |
| $\theta_m$ | Attitude sensor saturation level |
| $\ddot{\theta}_c$ | Jet control acceleration |
| $\dot{\theta}_L$ | Attitude rate during limit cycle |
| $\Delta\dot{\theta}_M$ | Measured incremental rate after thrust initiation |
| $\Delta\dot{\theta}_R$ | Incremental rate required to signal thrust drop-out |
| $\dot{\theta}_m$ | Rate limiting level |
| $M_C$ | Jet control moment |
| $M_E$ | External disturbing moment |
| $\alpha T_L$ | Attitude rate to displacement gain ratio |
| $T_L$ | Lead circuit time constant (denominator) |
| $T_\theta$ | Noise filter time constant |
| $T_c$ | Thrust build-up or decay time constant |
| $T_T$ | Total time constant or lag ($T_T = T_L + T_\theta + T_c$) |
| $T_p$ | Time duration of constant jet pulse |
| $t_{on}$ | Jet on-time increments with conventional switching |
| $t_o$ | Jet on-time measured by digital counting circuit |
| $t_m$ | Time duration allowed for initial damping |
| $t_T$ | Total (accumulated) jet on-time |
| H | Equivalent switching hysteresis |
| $I_v$ | Vehicle moment of inertia |
| $I_{sp}$ | Fuel specific impulse |
| $W_f$ | Required fuel weight per axis |
| F | Jet thrust magnitude |
| S | LaPlace operator |

NOTE: $\tau$ and T are equivalent symbols.

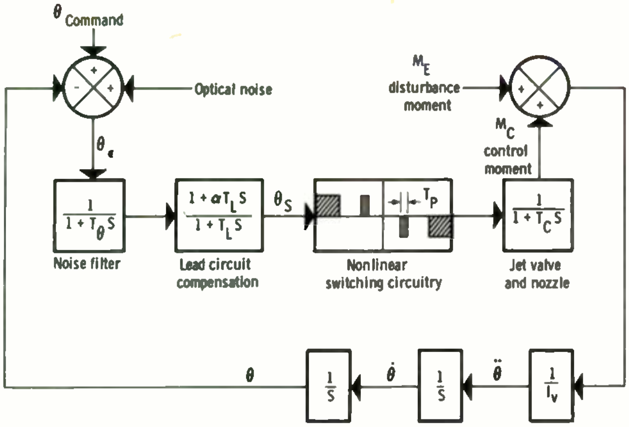


Fig. 1. Functional block diagram, jet control loop.

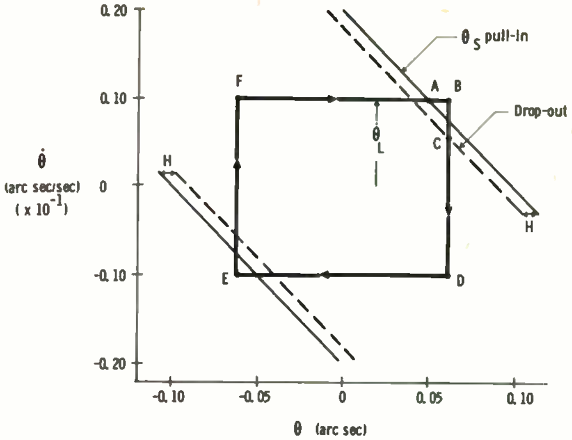


Fig. 4. Limit cycle operation.

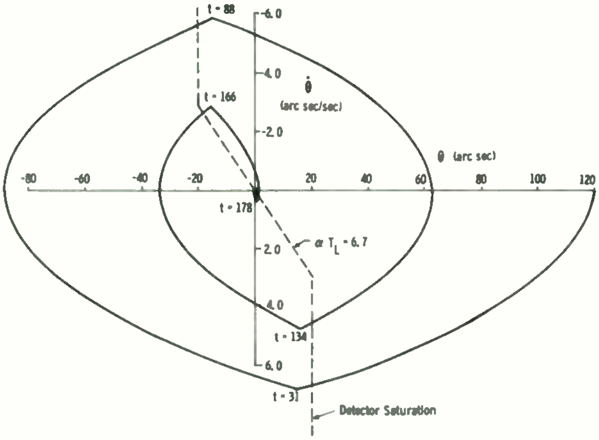


Fig. 2. Initial damping with detector saturation.

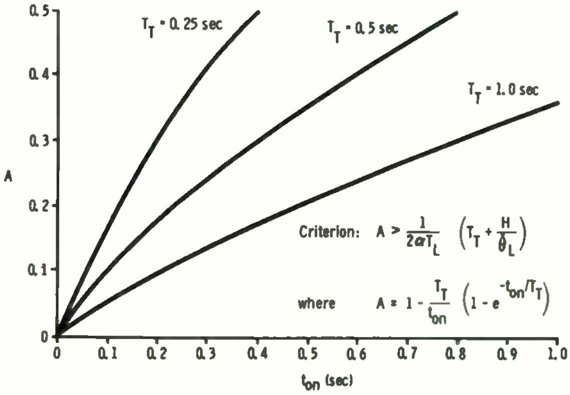


Fig. 5. Minimum pulse on-time criterion.

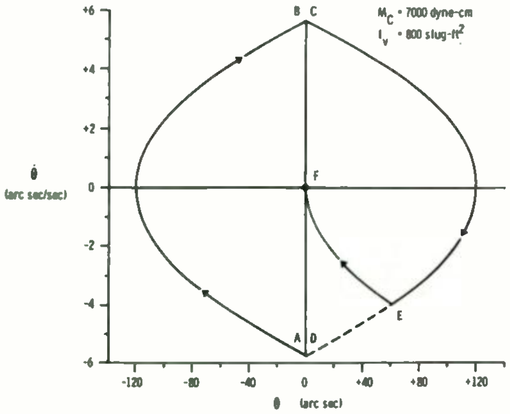


Fig. 3. Initial damping using time measurement.

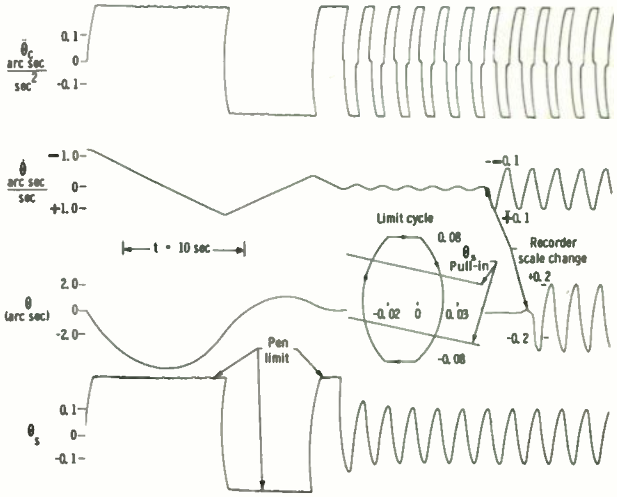


Fig. 6. System response with $\tau_p = 0$.

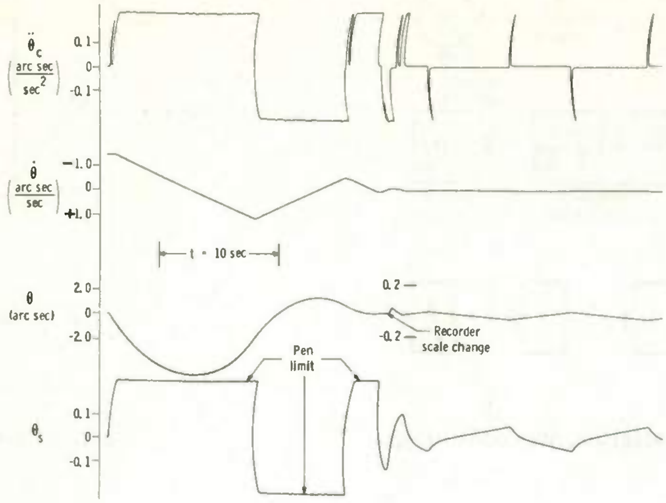


Fig. 7. System response with $\tau_p$ = 0.1 sec.

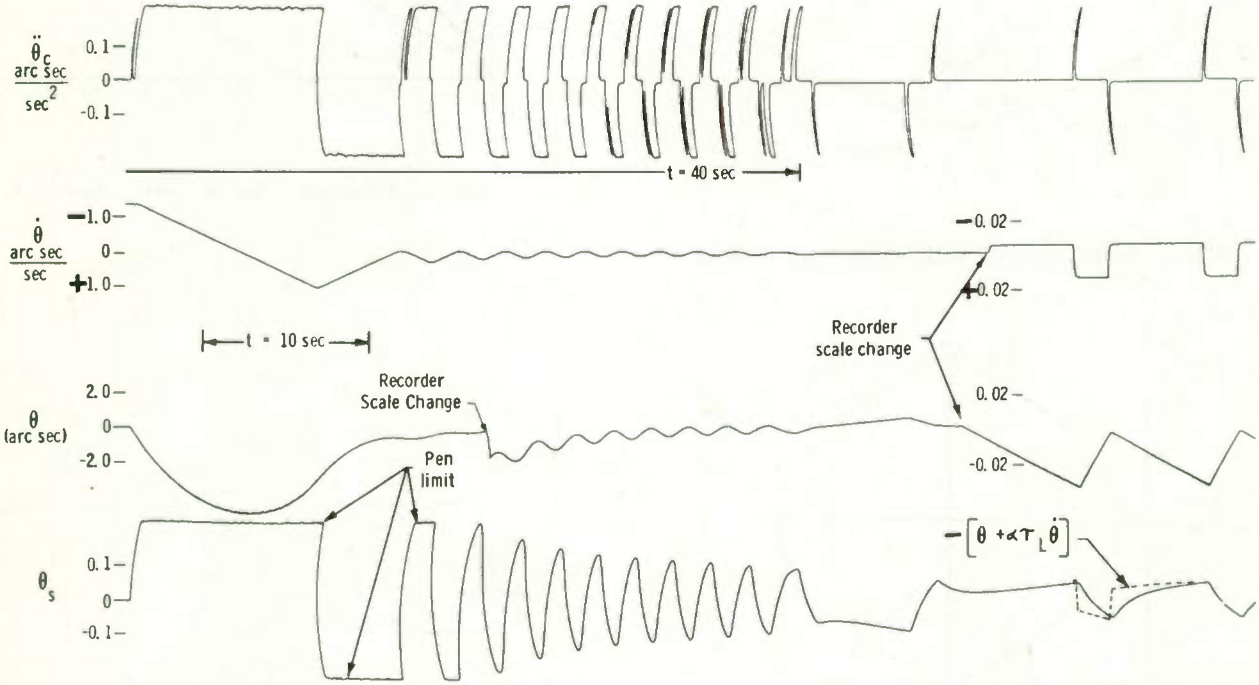


Fig. 8. System response with $\tau_p$ = 0.1 and increased $\tau_T$ and $\alpha\tau_L$.

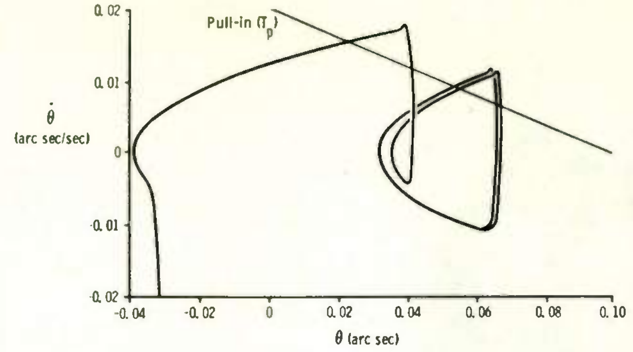


Fig. 9. System response with $M_E$ = 100 Dyne-cm.

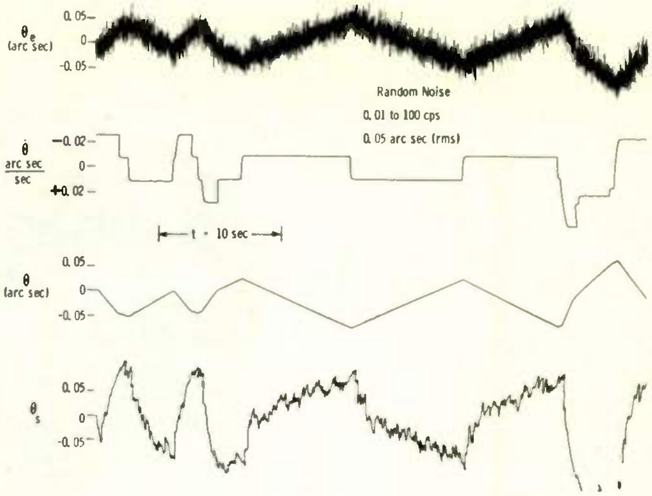


Fig. 10. System response with noise input.

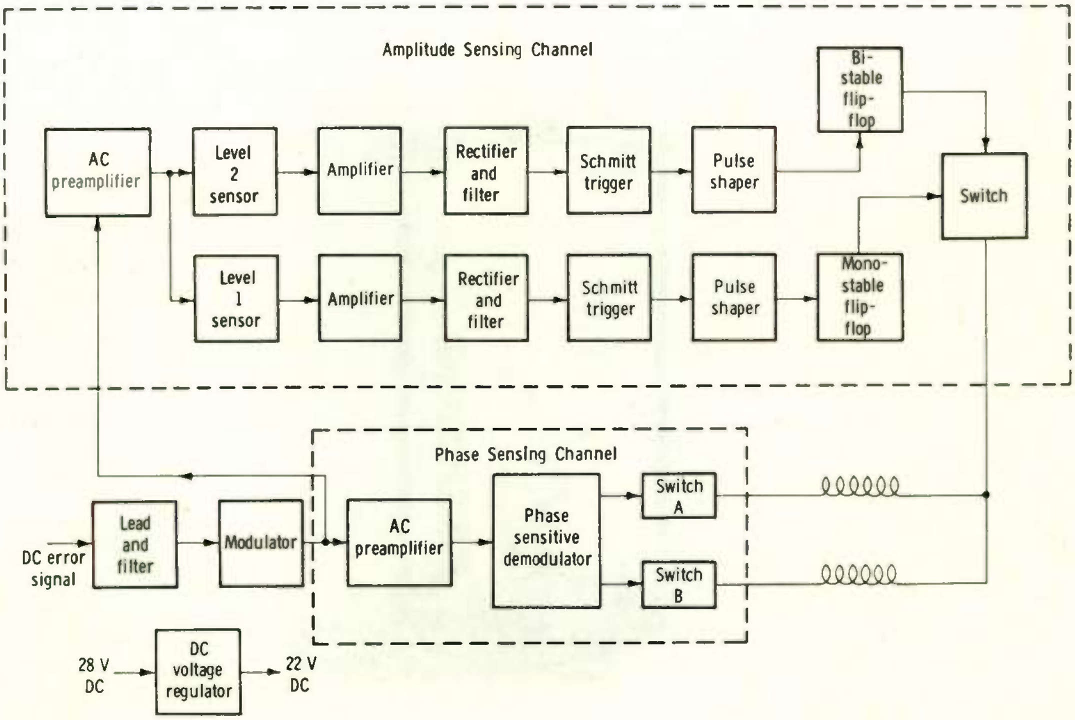


Fig. 11. Functional block diagram of jet switching circuitry.

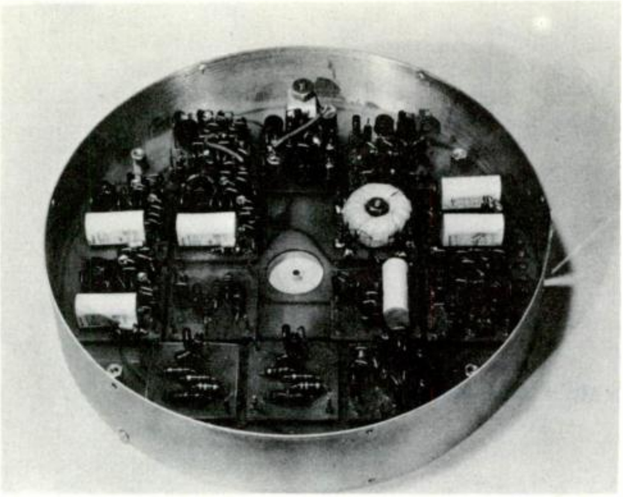


Fig. 12. Engineering model, jet switching circuitry.

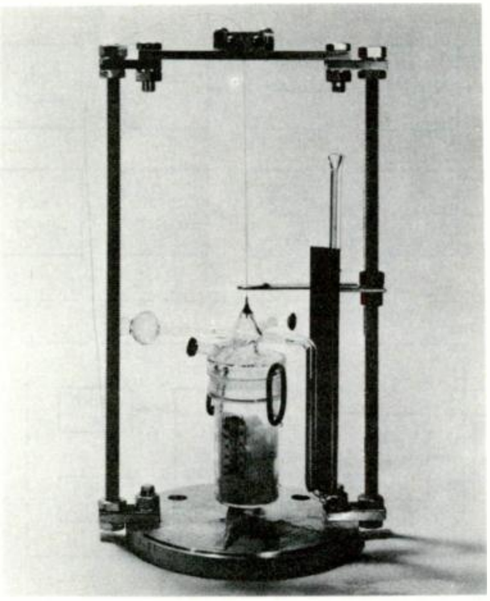


Fig 13. Vapor jet test apparatus.

# CIRCUIT REVOLUTION: A TUTORIAL INTRODUCTION TO THE SYMPOSIUM ON THE DESIGN OF NETWORKS WITH A DIGITAL COMPUTER

Philip R. Geffe, Axel Electronics,
Inc., Jamaica, N. Y.

## Abstract

Computers are revolutionizing circuit theory because any well-defined calculative procedure is now a practical procedure with a digital computer: i.e., computation costs are reduced by several orders of magnitude. Network applications yield theoretical studies as well as designs for hardware production. The first part of this paper surveys these usages to date.

In the second part of the paper, the computer art from a beginner's point of view is discussed. Programming is discussed in terms of object language, interpretive language, and problem-oriented symbolic coding systems. Some practical advice is offered for novices.

# EXPERIENCE WITH A STEEPEST DESCENT COMPUTER PROGRAM FOR DESIGNING DELAY NETWORKS

C. L. Semmelman
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey

## Summary

A computer program has been written for the purpose of designing networks by a successive approximation technique. The program performs two operations alternately: calculating the improvement made possible by a small change in each network design parameter and making larger changes in all parameters to reduce the error in performance. The program is written in Fortran-II for IBM-704 or 7090 computers with 16 or 32 K storage. It has been used mainly to design delay networks; however by rewriting one subroutine it may be used to design other types of networks or solve sets of equations.

## Operating Features

The steepest descent computer program was written in order to provide a general purpose tool for designing a wide variety of large and very precise networks and solving complicated sets of equations. It makes possible the rapid design of networks for which analysis methods are available but synthesis techniques are not. It allows the engineer to impose practical design objectives, such as nonuniform dissipation and the range of available element values. Further, it makes possible unusual design objectives, such as simultaneous loss and phase or combined frequency- and time-domain requirements. It does not restrict the designer to equal-ripple approximations or infinite-Q elements.

The program is written in Fortran-II language for the IBM-704 and 7090 computers and requires 16 K registers of core storage. In order to use the program an engineer must make a first estimate of the value of each design parameter, determine the maximum and minimum values permitted for each, and prescribe the number and location of the requirement points, e.g., $R_j$ and $f_j$ in Eq. (3). The program allows 128 match points and 64 parameters. It improves the original estimates by a successive approximation procedure so that the actual network behavior approximates the requirement in a least squares manner subject to the imposed constraints.

## Mathematical Formulation

Whereas a human designer could look at the shape of an error curve and decide whether or not a change in some parameter had made an improvement, a computer finds this a very difficult task - akin to pattern recognition. To ease this burden, the steepest descent program examines the value of one variable, y, which is defined for delay networks in the following manner. Let the delay requirement for the network be given by the J values $R_j$ at the frequency points $f_j$, and the actual delay of some n-section network be $T_j$, at those same points.

$$T_j = \sum_{i=1}^{n} \frac{b_i}{2\pi f_{ci}} \times \frac{1 + \left(\frac{f_{ci}}{f_j}\right)^2}{1 + \left(\frac{b_i}{2}\right)^2 \left(\frac{f_j}{f_{ci}} - \frac{f_{ci}}{f_j}\right)^2} \qquad (1)$$

where $f_{ci}$ is the frequency at which the phase shift of the ith section is 180° and $b_i$ is its stiffness parameter. These delay section parameters are defined in Fig. 1.

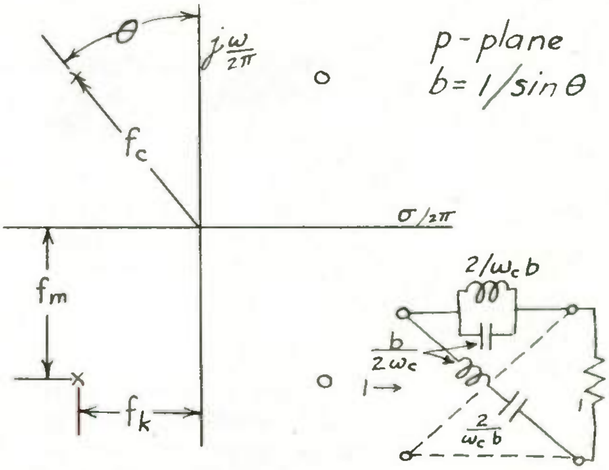


Fig. 1 - Definition of Delay Section Parameters

Equation (1) states that the delay provided by the network is a function of the frequency and the network parameters, or

$$T_j = T(x_i, f_j) \qquad (2)$$

where $x_i$ includes $f_{ci}$ and $b_i$ and $1 \leq i \leq 2n$. The match value, y, may then be defined as:

$$y = \sum_{j=1}^{J} (R_j - T(x_i, f_j))^2 = y(x_i) \qquad (3)$$

The problem, then is to minimuze y by adjusting the values $x_i$ subject to their constraints, where the expression $y(x_i)$ is an extremely messy function of, possible, several dozen variables.

## Description of the Program

### Direction of Changes

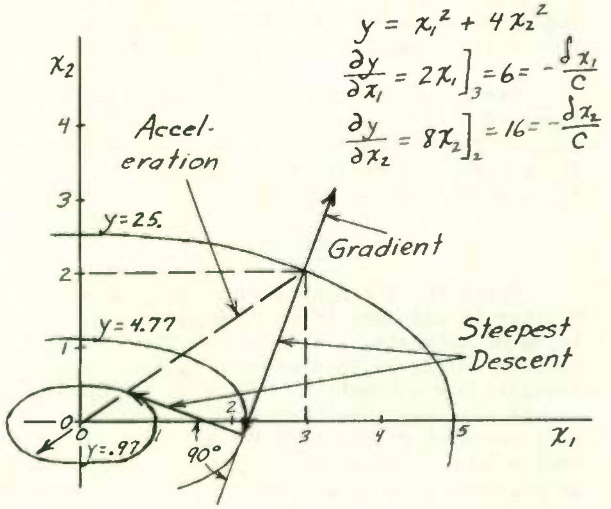
The method of attack on this problem comes from the fact that the direction in x-space in which y is increasing most rapidly is the direction defined by the gradient.

$$\text{grad } y = \frac{\partial y}{\partial x_1}\underline{x_1} + \frac{\partial y}{\partial x_2}\underline{x_2} + \ldots + \frac{\partial y}{\partial x_{2n}}\underline{x_{2n}} \quad (4)$$

Consequently the increments for the $x_i$ required to decrease y as rapidly as possible are given by:

$$\delta x_i = - C\frac{\partial y}{\partial x_i} \approx - C\frac{\Delta y}{\Delta x_i} \quad (5)$$

where C is a constant. This is shown graphically in Fig. 2, where x has been restricted to two dimensions and constant values of y are shown by contours. As the computer program must be able to evaluate the match value, y, it was decided to use the first difference ratio in place of the partial derivative. This was entirely a matter of convenience, as an additional evaluation of y was faster than a calculation of the derivative. The $\Delta x$ used in the first difference ratio is 0.0001 x.



$y = x_1^2 + 4x_2^2$
$\frac{\partial y}{\partial x_1} = 2x_1\Big]_3 = 6 = -\frac{\delta x_1}{C}$
$\frac{\partial y}{\partial x_2} = 8x_2\Big]_2 = 16 = -\frac{\delta x_2}{C}$
$x_2$
4
3
2
1
0
Accel-
eration
y=25.
Gradient
y=4.77
Steepest
Descent
2
y=.97
90°
0
1
3
4
5
$x_1$

Fig. 2 - Illustrating Gradient and Direction of Steepest Descent

## Magnitude of Changes

Having established the direction in which to change the x's, i.e., the signs and relative sizes of their increments, it is now necessary to determine their actual sizes. In the vicinity of a minimum it may be expected that contours will be ellipsoids in many dimensions and y will be approximately parabolic. If the value of y is calculated at three points the location and value at the parabola minimum may be determined. The points $x_i$, $x_i$+1F$\delta x_i$ and $x_i$+2F$\delta x_i$ are used for this purpose and values $y_0$, $y_1$ and $y_2$ respectively are found. F is initially an arbitrary constant but is adjusted as the calculation progresses.

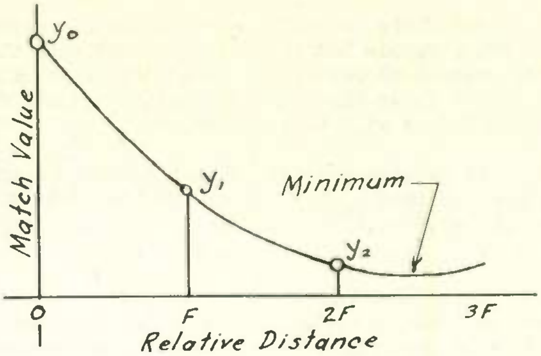


Fig. 3 - Parabola Along Negative Gradient

The location of the parabola minimum is given by

$$x_{im} = x_i + MF\delta x_i \quad (6)$$

where

$$M = \frac{3y_0 - 4y_1 + y_2}{2y_0 - 4y_1 + 2y_2} \quad (7)$$

The parabola minimum will in general not be the ultimate goal, as the negative gradient will not usually point toward the absolute minimum. This is shown in Fig. 2, where the descent direction does not pass through the origin. The parabola minimum will, however, be the point at which the steepest descent direction is tangent to a contour. Having reached the parabola minimum, a repetition of the above calculations will bring further improvement. Successive gradients, as shown in Fig. 2, will be perpendicular; however, in more than two dimensions the direction of the next path cannot be predicted by this means.

## Safety Features

Although the basic operating features of steepest descent approximation are covered in the preceding paragraphs, many devices to protect against unforeseen occurrences have been incorporated in the program.

It is conceivable that the points $y_0$, $y_1$, and $y_2$ could determine a parabola which was concave downward or could lie on a straight line. The former would cause the program to locate a maximum and the latter would result in a division by zero. To avoid such occurrences, zero or negative curvature causes the middle point $y_1$ to be dropped and a new one, $y_4$, calculated at $x_i$+4F$\delta x_i$. The points $y_0$, $y_2$ and $y_4$ are then checked for curvature and the process is repeated using $y_0$, $y_4$ and $y_8$; $y_0$, $y_8$ and $y_{16}$, etc., until positive curvature is found.

The curvature, although positive, may be so small that the location calculated for the minimum is unreliable because of rounding error. The process described above is used in this case also, to obtain more reliable information.

When this procedure of doubling the spread of the parabola has been executed too many times, the program increases the numerical value of the constant F, so that with the next gradient, fewer calculations will be required.

It is also possible that $F\delta x_i$ will be too large, so that $y_1 > y_0$, i.e., on the other side of the minimum. In this case F is reduced by a factor of ten and a new value is found for $y_1$.

After the location of the parabola minimum has been found, it is possible that the match value at that point may be larger than one or more of the values $y_0$, $y_1$ or $y_2$. The program selects the lowest as a point from which to continue and thereby assures that the approximation never becomes worse.

In case the adjustment of parameters by adding $KF\delta x_i$ results in one or more of them assuming values outside their constraint range, the corresponding limiting value is immediately substituted. Although this results in a departure from the steepest descent direction, it will produce results in keeping with the designer's wishes.

## Time-Saving Features

Two features have been incorporated in the program solely in the interest of speed. One of these is made possible by the fact that each term in the summation of Eq. (1) is a function of only two of the totality of network parameters. For this reason, only that one term needs to be recomputed when the first difference ratios are being evaluated. The second feature is the acceleration step suggested by A. I. and G. E. Forsythe.[1] The operation of this mechanism is shown in Fig. 4.

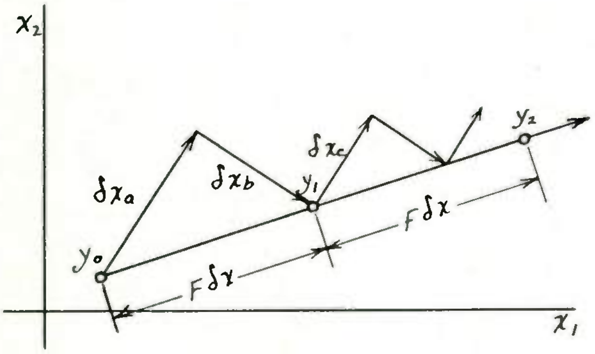


Fig. 4 - Acceleration Step

Frequently, successive changes in a set of parameters will appear as shown by $\delta x_a, \delta x_b, \delta x_c$, etc. When this happens, time can be saved by forming the vector sum of two successive changes, say $\delta x_a$ and $\delta x_b$, and using this in place of $F\delta x_i$ in Fig. 3. The corresponding values $y_0$ and $y_1$ have already been evaluated and stored so that only one new evaluation, $y_2$, is needed to prepare for a major stride forward. By comparison, the equivalent of at least four evaluations would be needed to prepare for an additional step of the $\delta x_c$ type.

## Other Features

Because the evaluation of delay from Eq. (1) forms a very large part of the computer program, the time required may be estimated on this basis. The IBM-7090 requires about 0.7 millisecond to evaluate each term of this summation.

As the designer will not, in general, know the optimum amount of flat delay to include in his requirement, the program selects this for him. This feature may be suspended when necessary.

The program stops when either of two conditions is met: either the number of trials prescribed for the machine run has been completed or the program has failed to make improvements totaling one per cent in the last three trials.

## Results

Since the steepest descent program was written it has been tried on a wide variety of types of networks. In adjusting three to six element values in constant-R bridged-T loss equalizers it has achieved uniformly excellent results. It has also been used to adjust element values in a five-branch finite-Q filter to meet a loss requirement while maintaining as low phase shift as possible at a given frequency. In this case one degree phase was weighted as heavily as 1 db loss, and good results were obtained. The program has also worked well in adjusting the $f_c$ values of four phase sections whose b values were fixed at 2.0, in order to produce a constant phase difference across a given band.

## Delay Networks

When applied to the design of delay networks, however, the results have been only partly satisfactory. Figure 5 shows typical error curves at the beginning and end of a run in which a parabola minimum was found six times. This network contained 20 sections and the requirement was

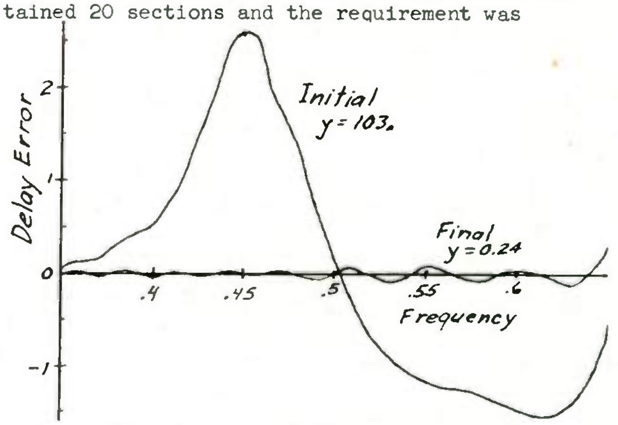


Fig. 5 - Typical Improvement

specified at 61 frequencies. For most uses this design would be entirely satisfactory; however, the ripple is not as small as is theoretically possible with 20 sections.

A good measure of the performance of the program may be obtained by comparing the number of crossings of the calculated delay characteristic with the required curve. Optimum results would be obtained when this number equals the number of delay section parameters plus one. The added crossing comes about because the program is allowed to select its own flat delay level, making an additional variable. Table 1 shows the best results that have been obtained for networks of various sizes.

Table 1 - Results of Steepest Descent in Designing Delay Networks

| Total Number of Parameters | Final Number of Crossings |
|---|---|
| 7 | 4 |
| 21 | 17 |
| 33 | 14 |
| 41 | 17 |

Several conjectures have been advanced to account for the final numbers of crossings being less than the theoretical values. One possible explanation is that the program becomes trapped in a local minimum and cannot get out of it to find a better match elsewhere. On an intuitive basis, it is difficult to believe this. There is certainly some combination of parameter values which produces 2n+1 crossings. It seems inconceivable that the match can be made worse instead of better when changes are made which are each a very small fraction of the changes required to reach the optimum. We have been unable to demonstrate rigorously that local minima either can or cannot exist. A second explanation is that the 27 bit fractions carried in the computer do not have sufficient numerical accuracy to carry through to the final solution. This explanation appears to be much more reasonable, as the 7-parameter case has been run until the computer began to repeat the calculations with identical numbers. It achieved only the four crossings shown in Table 1. The use of a first difference ratio instead of the true partial derivatives has also been suggested to be the cause of the difficulties.

One characteristic which is common to all the delay section computations is that very small changes are made in the b parameters. In a 3-section delay equalizer for a filter pass band the initial and final values and the per cent changes in the parameters are listed in Table 2.

Table 2 - Results Using Steepest Descent Direction

| | Initial | Final | Change | Av. Change |
|---|---|---|---|---|
| $f_{c1}$ | 2.60 | 2.58974 | .394% | .272% |
| $f_{c2}$ | 2.65 | 2.64627 | .014 | |
| $f_{c3}$ | 2.7 | 2.71103 | .409 | |
| $b_1$ | 19.0 | 19.00089 | .000473% | .00112 |
| $b_2$ | 20.0 | 19.99471 | .002650 | |
| $b_3$ | 21.0 | 20.99952 | .000225 | |
| y | .237 | .080 | 66.3% | |
| Crossings | 2 | 4 | +2 | |

If it is considered that only the $f_c$'s were adjustable, the results in Table 1 appear to be quite near the theoretical limits.

## Scale Factor Effect

It is believed that this avoidance of b changes is attributable to the scale factor effect. Because of it, the gradient direction is not invariant to a change in the units in which the parameters are expressed. This may be demonstrated readily by consideration of the equation and elliptic contours shown in Fig. 2. The optimum direction in which to change the parameters is toward the origin; however, only the gradients at points on the semimajor and semiminor axes will result in such a change. If the transformation

$$x_1 = 2w \qquad (8)$$

is made, the equation becomes

$$y = 4w^2+4x_2^2 \ , \qquad (9)$$

which plots as concentric circles. Now the negative gradient at every point will point directly to the minimum at the origin. In such a situation the program might, for example, converge quickly if frequencies were expressed in megacycles per second but not if the designer preferred kilocycles.

## Least Squares Direction

In order to test this hypothesis, the method of determining the improvement direction was modified by calculating the change in each parameter needed to make a least squares match to the requirement curve. This is the Taylor Series method described by M. R. Aaron.[2] These coefficients determined the direction of change and the distance was determined by putting a parabola through three equally spaced points, as before. The results of this modification are shown in Table 3 and are plotted in Fig. 6.

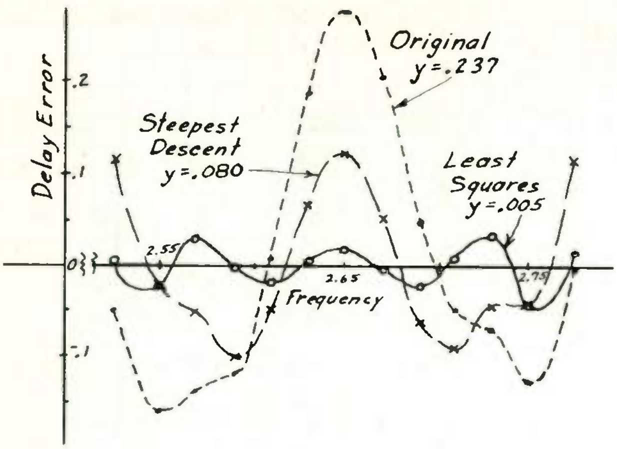


Fig. 6 - Comparison of Steepest Descent and Least Squares Results

Table 3 - Results Using Least Squares Direction

| | Initial | Final | Change | Av. Change |
|---|---|---|---|---|
| $f_{c1}$ | 2.6 | 2.59601 | .15% | 0.19% |
| $f_{c2}$ | 2.65 | 2.64531 | .18% | |
| $f_{c3}$ | 2.7 | 2.70664 | .25% | |
| $b_1$ | 19.0 | 25.23380 | 32.8% | 31.4% |
| $b_2$ | 20.0 | 12.92902 | 35.3% | |
| $b_3$ | 21.0 | 26.52252 | 26.3% | |
| y | .2370 | .00546 | 97.7% | |
| Crossings | 2 | 8 | +6 | |

These results show that large changes in b values are needed and are not being made with the negative gradient program. To obtain more positive evidence, the final parameter values from the steepest descent run (Table 2) were used as initial values for a least squares run. The results do not differ to any practical extent from the final results in Table 3. This shows that sufficient numerical accuracy was available and that a local minimum did not stop the improvement.

Although it worked very well in this example, the least squares procedure does not appear to be the ultimate answer. Cases have been found where two parameters produce rather similar changes in delay. The simultaneous equations which result will then be ill-conditioned and may require more precision than is available. Further, the least squares procedure may then require an exhorbitant positive change in one parameter and an equally undesirable negative change in the other. When these changes are tried, the nonlinearity of the functions prevents the expected improvements.

## Conclusion

The steepest descent procedure appears to be adequate for a wide range of engineering applications. In cases where results near the theoretical limits are required. a better process for determining the direction of change would be desirable. Several ideas are being investigated but none has yet proven itself capable of consistently producing the optimum theoretical results.

## References

1. A. I. and G. E. Forsythe, "Punched Card Experiments with Accelerated Gradient Methods for Linear Equations," Nat. Bu. Stds., Applied Math Series, Vol. 39 (1954) pp. 55-70.

2. M. R. Aaron, "The Use of Least Squares in System Design," IRE Trans. on Circuit Theory. Vol. CT-3, pp. 224-231.

3. C. A. DeSoer and S. K. Mitra, "Design of Lossy Ladder Filters by Digital Computers," IRE Trans. on Circuit Theory, Vol. CT-9, pp. 192-201.

# FILTER SYNTHESIS USING A DIGITAL COMPUTER

G.C. Temes
Northern Electric Research and Development Laboratories
Ottawa, Canada

## Summary

The design of filters generally involves several successive steps: the derivation of a realizable approximation to the specified transfer characteristics; the calculation of the immittance parameters; and finally, the establishment of the circuit diagram and the element values. The synthesis is usually followed by an analysis of (and, if necessary, compensation for) the effects of parasitic phenomena and element variations. By combining analytical, numerical and graphical procedures, a great variety of specifications can be satisfied by networks realizable in practical configurations and with feasible element values. This paper classifies and gives a brief description of the digital computer programs that were found to be most useful in these calculations. Several representative examples are included.

## I. Introduction

The purpose of this paper is to give a classification and brief description of the digital computer programs that have been - or are expected to be - useful in the synthesis of filters and related networks at the Communication Networks Laboratory of Northern Electric R & D Labs. This laboratory carries out research and development to design filters and other networks used in telephone and radio transmission systems. All computational design work is done on a digital computer. The computer currently being used is an IBM 1620.

The programs referred to in this paper are subject to the following restrictions:

a, The design methods are based on the insertion loss theory.[1]

b, The filters are resistively terminated lumped LC ladder networks without mutual inductances.

c, The networks are sharply frequency-selective, with clearly distinguished pass and stop bands.

d, All specifications are in a frequency-domain.

The synthesis of such networks is carried out in the following steps:

1, On the basis of the specifications, a rational function satisfying the realizability conditions, and meeting the specifications is derived. This procedure is called approximation. The derived function is usually the insertion voltage ratio $\Lambda$ or the insertion characteristic function $\Phi$ of the desired filter.[2]

2, $|\Lambda|^2$ and $|\Phi|^2$ are related through energy considerations, and one can be obtained from the other. The immittance parameters or driving-point immittances of the ladder are then obtained from both $\Lambda$ and $\Phi$ .

3, The filter element values can be found using one or more of these immittances. If necessary, network equivalences and transformations can be utilized to modify the configuration or element values. Steps 2 and 3 embrace the synthesis portion of the network design.

4, In addition to the general realizability conditions on passive, lumped-element, resistance-terminated LC - fourpoles,[1-3] the requirements for ladder realizability[4] must be satisfied. For some special attenuation characteristics a check can be made after Step 1; for others, after Step 2.

5, The performance of the synthesized network is usually examined immediately after the design has been completed. The attenuation, phase, driving-point immittance etc., can be checked.

6, The effects of parasitic phenomena (dissipation, aging, tolerances etc.) are analyzed. Some preliminary estimate can sometimes be obtained after Step 1, but usually the full investigation can be performed only after the network has been obtained. Steps 4, - 6, are concerned with the analysis of the filter and its parameters.

7, Depending on the results of Step 6, it may be necessary to repeat the synthesis procedure including a precorrection for the effects of parasitic elements. Alternatively, the filter may be cascaded with correcting networks (equalizers) to improve its performance.

A schematic illustration of the complete design procedure is shown in Fig. 1.

Although it is theoretically possible to construct programs that carry out all necessary calculations in one run, it was found that the computer storage capacity does not permit this for major design problems. For this reason, and also for flexibility in application, the more involved synthesis programs are built up from several (2 - 4) sub-programs, from which a large number of combinations can be obtained. Corresponding to the design steps described above, these sub-programs can be divided into the following groups:

A, Approximation programs;

B, Synthesis programs;

C, Analysis programs;

D, Programs used in the estimation and precorrection for parasitic effects.

These subprograms will now be briefly described.

## II. Computer Programs Used in the Approximation Procedure

The purpose is usually to derive the function $\Phi$ . $\Phi$ is generally chosen because:

a, The realizability restrictions on $\Phi$ are not as severe as those on $\Lambda$ .

b, In the pass band $\Phi$ approximates zero rather than a constant as does $\Lambda$ .

c, Generally the zeros as well as the poles of $\Phi$ lie on the real frequency axis.

All of these factors tend to make it easier to solve the approximation problem in terms of $\Phi$

rather than in terms of $\Lambda$.

### A, Approximation of the Ideal Low-Pass Attenuation Characteristics

In practice, the most common filter specification is one requiring a low-pass attenuation characteristic. The following programs dealing with this approximation problem have been used.

1, Butterworth polynomial approximation. The attenuation response approaches zero in the pass band and infinity in the stop band in a maximally flat manner. The degree of the polynomial, the parameters of the characteristic function and the element values of the filter can be obtained directly from formulae[5] expressing them in terms of the maximum pass band attenuation $A_p$, the minimum stop band loss $A_s$, the band limits $\omega_p$ and $\omega_s$, and the terminations $R_1$ and $R_2$.

2, Chebyshev polynomial approximations. Explicit formulae are also available[5] for the case of a Chebyshev pass band, maximally flat stop band response. The complementary case (flat pass band, Chebyshev stop band) can be treated by replacing $|\Phi|$ by $1/\Phi$ and $\omega$ by $1/\omega$.

3, Darlington-Cauer (elliptic) approximation. Both the pass band and the stop band loss are approximated in a Chebyshev manner. The design tables for one to four section symmetrical filters, given by Grossman[6], were programmed. These tables yield $\Phi$, $\Lambda$ and the element values. For other networks the roots and poles of $\Phi$ (the so-called Cauer parameters) are obtained using a $\vartheta$-series approximation.[6-7]

Although other approximations of the low-pass characteristics exist, the frequency of applications did not warrant their programming.

Using reactance transformations[3], these results can be applied to the design of high-pass, frequency-symmetrical band-pass, band-stop and multiband filters (cf. Para. III. D.1.)

### B, Chebyshev Pass Band, Arbitrary Stop-Band Approximation

Filters approximating zero loss in the Chebyshev sense in their pass bands and satisfying arbitrary frequency-dependent specifications in the stop bands are frequently used. Usually the design procedure begins with the location of the attenuation poles to satisfy the stop band requirements. This can be done by a graphical procedure employing special templates and calibrated graph-paper[8-9]. The process can also be programmed on a computer. Different templates or programs are used for low-pass (high-pass, frequency symmetrical band-pass and band-stop) filters and for frequency unsymmetrical band-pass networks. The following programs are available for the design of such networks:

1, Chebyshev pass band low-pass approximation. After the attenuation poles have been established, a step-by-step design scheme[10], based upon Darlington's reference filter method[1], is followed to synthesize networks with up to four sections.

2, Chebyshev stop band, arbitrary pass band approximation can be obtained in some cases by combining the Chebyshev pass band synthesis procedure with the $|\Phi| \to 1/|\Phi|$, $\omega \to 1/\omega$ transformation.

3, For unsymmetrical band-pass networks the design procedure described by Saal and Ulbrich[5] was programmed. The network thus obtained could be transformed into a band-stop filter with different ripples in the two pass bands.

4, Recurrence relations yielding $\Phi$ for Chebyshev pass band filters are given by Fettweiss[11].

### C, Filters with Arbitrary Pass Band and Stop Band Characteristics

Several design methods are also available for filters satisfying frequency-dependent specifications in both their pass and stop bands. Some of these were programmed on the computer.

1, Chebyshev pass band filters cascaded with constant-resistance equalizers. The programs of Para. II. B. are used to achieve the required stop band characteristics and constant resistance ladders[12] or bridged-T equalizers are designed* to bring the pass band response within specifications.

2, Approximation methods such as least-squares polynomial approximation[13], curve matching at selected points[3], and the steepest descent method[14] are used.

3, The design method of Linke[15] has been programmed. This approximation procedure makes use of a group of curves showing the effects of the individual complex natural frequencies and imaginary attenuation poles. The use of these curves is demonstrated in Fig. 2. The critical frequencies thus obtained are fed into the machine one by one and the resulting error typed out for further approximation. After a close approximation has been obtained, the computer improves upon the results to give the least-squares error.

4, Darlington's Chebyshev polynomial series approximation[16] has been used for the synthesis of filters with preassigned attenuation poles and prescribed pass band characteristics.

5, The iterative method described by Shenitzer[17] can be used to obtain equiripple approximation of a prescribed pass band response and specified attenuation poles.

### D, Filters with Prescribed Attenuation and Phase Response

The following approximating programs are useful in the design of filters satisfying both attenuation and phase requirements:

1, Synthesis of a filter exhibiting the required attenuation response, cascaded with an additional all-pass phase corrector. The latter can be designed by Darlington's method[16, 18-20] or by trial and error[22].

2, Filters with Chebyshev pass band and prescribed phase characteristics can be designed using Bennett's method[21]. This involves, esentially, the construction of a numerator for $\Lambda$, which gives the proper phase response, and then the selection of a constant-phase denominator which secures the Chebyshev pass band behaviour.

3, A somewhat similar approach has been suggested by Skwirzynski and Zdunek[23]. The attenuation associated with the minimum phase network required to satisfy the phase characteristic is computed

*Several programs and graphical aids were created for the design of these equalizers. These are, however, beyond the scope of this paper.

using Bode's relationship[25]. This loss response is approximated by a Chebyshev polynomial series and equalized by including a proper constant-phase denominator, with $j\omega$-axis roots[24].

4, Darlington's Chebychev polynomial series approximation procedure[16] has been used successfully to satisfy simultaneous loss and phase specifications[5].

5, Low-pass response combined with maximally flat group delay characteristics can be obtained by using Bessel-polynomials[26] to construct $\Lambda$.

6, Beletskiy describes a method[27] suitable for the design of filters with maximally flat pass band delay and Chebyshev stop band attenuation characteristics.

## III. Filter Synthesis Programs

The approximation procedure yields one of the functions $\Phi, |\Phi|^2, \Lambda$ or $|\Lambda|^2$. The network element values can then be computed using Darlington's method[1].

### A, Calculation of $\Phi$ and $\Lambda$

In general the network immittances are functions of $\Phi$ and $\Lambda$. The calculation of $\Phi$ and $\Lambda$ constitutes the most laborious part of the design. The usefulness of the digital computer, at this stage, cannot be overemphasized.

1, Computation of $\Lambda$ from $|\Lambda|^2$. The denominator of a ladder-realizable $\Lambda$ has only conjugate $j\omega$-axis roots, while its numerator is a Hurwitz-polynomial. Hence, a straight-forward root-solving routine is used to find the left half-plane root factors of the numerator of $|\Lambda|^2$ and then to multiply them together to obtain the numerator of $\Lambda$. The denominator of $\Lambda$ is simply obtained from the square-root of the denominator of $|\Lambda|^2$.

The computation of the roots of the numerator polynomial of $|\Lambda|^2$ can be performed by any one of several published methods[28-30]. A program, using 18-digit precision subroutines and the Newton-Bairstow method[30], has yielded roots correct to an accuracy which has been found adequate for all practical networks.

Also available is a direct method for the Hurwitz-factorization of polynomials without solving for the zeros[31].

2, The computation of $\Phi$ from $|\Phi|^2$ can be carried out in a similar way. The denominator of $\Phi$ also has $j\omega$-axis roots only; however, the numerator need not be Hurwitz. Normally, a Hurwitz-type numerator for $\Phi$ is derived first, and then the formulae[1] giving the reflection coefficients of the network in terms of $\Phi$ and $\Lambda$ are utilized to find the signs of the even and odd parts of the final numerator, corresponding to the required termination ratio and configuration[1,2].

3, The calculation of $|\Phi|^2$ from $|\Lambda|^2$ is done using Eq. (17) of Darlington's paper[1]. For single-loaded networks these two functions are equal.

4, $|\Phi|^2$ can be obtained from $\Phi$, or $|\Lambda|^2$ from $\Lambda$ using the equality of $|f(p)|^2$ and $f(p)\ f(-p)$ on the $j\omega$-axis.

5, The synthesis of Butterworth or Chebyshev polynomial filters can be effected by using simple formulae[5] expressing the element values directly in terms of the attenuation requirements and the ratio of the terminations.

6, For elliptic filters, approximating formulae are used to find the critical frequencies[5] and also the element values of symmetrical filters with 1 - 4 sections[6].

### B, The Derivation of the Filter Element Values

The element values can be obtained by developing a suitable driving-point immittance function. For large networks it is advisable to develop the circuit from both an input and output immittance, to increase the accuracy of the realization, and to provide a check.

1, The design immittances can be found from Darlington's Eqs. (14) and (18)[1]. In order to extend the validity of these formulae to $\Lambda$ and $\Phi$ with odd denominators, however, the subsitution of $A \to \Lambda_e$, $pB \to \Lambda_o$, $A' \to \Phi_e$, $pB \to \Phi_o$ (suffixes e and o denoting the even and odd parts, respectively) must be made. Equation (14) can also be used, after these substitutions are made, to check that the signs of $\Phi_e$ and $\Phi_o$ conform to the desired configuration and terminations.

2, For constant-k type configurations the realization can be performed through the continued fraction expansion[3] of the design immittances or equivalently, by the removal of poles at zero and infinity from this immittance.

3, For general mid-series or mid-shunt low-pass filters the design process described in Darlington's Table I[1] was programmed. For symmetrical low-pass filters Darlington's Table II[1] gives the element values of networks with up to three sections. The formulae for four-section networks can be found in Grossman's article[6].

4, For more general ladder networks the pole-removing immittance expansion method due to Bader[32] was programmed. It can be used to expand a driving-point, as well as an open or short circuit immittance. The design formulae were tabulated by Saal and Ulbrich[5]. When this process is used, the order of removal of finite poles must be properly chosen to satisfy the ladder realizability conditions[4,8]. Also, it is expedient to check the accuracy by ladder development from each end of the network[3], since the number of significant figures in the element values decrease rapidly with increasing network complexity.

### C, Programs for the Synthesis of Band Separating Filters

The synthesis of filter groups performing the separation or combination of frequency bands is somewhat different from the design of other filters. Since these networks are either parallel or series connected at one end, constraints are placed upon their driving point immittances. As a consequence, the voltage or current transfer ratio N or M and the voltage or current characteristic function $\Phi_N$ or $\Phi_M$ is the most useful design parameter. The programs used in the design of these networks are based on the following design approaches:

1, The exact design of band separating filter groups is carried out by approximating the specified responses with the $\Phi_{N_i}$ or $\Phi_{M_i}$ of the individual networks[33]. In addition, the $\Phi_{N_i}$ ($\Phi_{M_i}$) have to satisfy a constraint originating from a required constant driving-point immittance at the parallel (series) connected terminals.

2, A much more economical synthesis procedure can be used if only an approximately constant driving point immittance is permissible[8,34].

3, Auxiliary networks can be used to absorb power if the pass bands do not overlap[33].

4, For filters with stringent stop-band requirements, it is expedient to cascade a band separating filter group designed for low stop-band loss with individual filters yielding the required stop-band attenuation[3].

5, The design of a harmonic separating filter set, i.e. of filters than can be used to suppress or select the harmonic content of signals within a prescribed frequency range, involves the calculations of the optimal selectivity of the set[35]. This calculation can be programmed for various approximations and filter types.

## D, Network Transformations

Sometimes it is advantageous to synthesize a simplified model of the required networks and then to convert this "prototype" into its final form using network transformations. Transformations are also used to achieve desirable configurations and element values. The following transformations were programmed:

1, Reactance transformations[3] used to convert a normalized low-pass prototype network into denormalized low-pass, high-pass, band-pass, band-stop, or multi-band filters.

2, The approximating procedures described by Atiya[36] and Cohn[37] transforming constant-k type low-pass prototypes into band-pass filters built up from capacitively coupled tuned circuits. The latter is a most useful configuration for high-frequency, narrow band filters.

3, The "zig-zag" transformation converting low-pass prototypes into band-pass filters having a minimum number of coils has also been programmed, using Figure 12 of Saal and Ulbrich[5]. Recently, recurrence formulae were also published[38].

4, Another transformation[5] converts a single section low-pass prototype into a band-pass circuit in which all nodes are capacitively loaded.

5, The impedance level in some parts of the circuit can be changed using transformations due to Norton[3].

6, The configuration can be changed using the familiar T-$\pi$ tranformation and other specialized equivalences[3].

## IV. Analysis Programs

An important part of the design process is the analysis of the feasibility of the requirements and of the performance of the resulting network.

## A, Realizability Analysis

The criteria necessary for ladder realization without mutual inductance can be programmed into various parts of the synthesis routine. However, for simple networks it is normally just as easy to carry out the design and check for negative elements. Also, realizability nomographs and curves can be developed for convenient use. The following is a summary of the routines used to test the realizability of networks[4].

1, For symmetrical filters without finite attenuation poles (Butterworth, Chebyshev pass band filters) there are no restrictions on the ladder realizability. For Chebyshev stop band filters the stop band loss must be larger than a minimum value which is a function of the degree.

2, Symmetrically terminated antimetric filters with Butterworth stop band characteristics are always realizable; for Chebyshev stop band filters there is a lower limit on the stop band loss.

3, For elliptic filters with 2 or more sections, the stop band loss, the pass band ripple and the selectivity all have lower limits that is a function of the other parameters and the degree.

4, For more general mid-series and mid-shunt networks, the criteria of Meinguet[4] apply.

## B, Network Analysis Programs

The following programs are normally used immediately after realization to test the response of the network:

1, The transfer and driving-point characteristics are plotted as functions of frequency[39]. The loss and phase, the delay, the voltage and current ratios and the driving-point immittances can be calculated with or without parasitic elements.

2, The transfer and driving point immittances, as well as the soldering iron and plier type immittances of ladders can be computed from the schematics as rational functions of the frequency[40].

3, The transient response of the network, for a step and various pulse inputs or modulated sine-wave excitation can be computed from the schematic or transfer function[41].

## C, Mathematical Analysis Programs

Some general purpose analysis programs that are used in various applications are:

1, Calculation of attenuation and phase from the zeros and poles of transmission or reflection.

2, The calculation of the reflection factor, the driving-point immittance, $\Lambda$ or $\Phi$ from each other.

3, The calculation of the real and imaginary components of a minimum phase network function from each other using Bode's formulae[25].

4, Evaluation of a rational function to find its real and imaginary parts, phase, and absolute value. Also, finding the $j\omega$-axis minimum of the absolute value or the real part.

5, Conversion among rational, continued and partial fractions.

6, Calculations of various driving-point and transfer quantities from the immittance or chain matrix. Conversion of various parameter matrices into each other.

## V. Programs for the Estimation of and Precorrection for Parasitic Effects

Even for moderately difficult specifications, the effects of parasitic elements, temperature, tolerances, aging etc. become significant. Some of these (dissipation, stray elements) can be taken into account in the synthesis procedure*; others (aging, tolerances) can be limited only by specifying high quality components. In any case, the anticipated effects must be estimated.

### A, The Estimation of Parasitic Effects

The following programs are used:

1, The estimation of parasitic effects on the synthesized network can be carried out utilizing the analysis program of Para. IV, B.1. Using a first-order perturbation method, the distortion can also be calculated as a mathematical function of the parasitics and the frequency[42].

2, The effects of uniform dissipation on the attenuation (phase) response can be estimated in advance by performing a shift on the frequency-variable[1] in $\Lambda$ or from the derivative of the phase (attenuation) function[25]. If the Q's are high and the frequency origin is not contained in the passband, the response thus derived will also give a good approximation for semiuniform loss distribution[3].

3, For general networks with semiuniform losses a straightforward procedure was developed[42], which can be used to estimate the distortion before the network is realized.

### B, Precorrection for the Effects of Dissipation

If the estimation process indicates that a compensation for the incidental losses is needed, the following programs can be utilized for this purpose.

1, The precorrection for uniform dissipation is carried out by a simple frequency-shift[1] in the independent variable of $\Lambda$ . The termination ratio is normally also modified to preserve the realizability of $\Lambda$ .

2, For semiuniform loss distributions, Darlington's procedure can be used[1]. Although this procedure is straightforward for single-loaded fourpoles[43], it is quite complicated for double-loaded networks and a simple process based on a perturbation approach may be preferable[42].

3, For special network configurations the methods described by Geffe[44] and Dishal[45] are applicable. Geffe gives explicit formulae for the predistorted voltage ratio of a single-loaded constant-k type low-pass filter with uniform or semiuniform losses, while Dishal describes a design method for band-pass filters with Butterworth or Chebyshev passband response.

4, For networks with arbitrary loss distributions and without finite attenuation poles the predistortion method of Desoer[46] can be used. For the precorrection of arbitrary ladder networks, a more general and easily applicable precorrection method was found to be effective[42]. A recently published correction technique[47] based upon steepest descent type distortion minimization seems to demand an unduly high cost in terms of stop band discrimination.

5, The design of a predistorted double-loaded filter is laborious; furthermore, it always introduces some flat loss and thus decreases the return loss. Also, the sensitivity to element variations is increased. For these reasons, it is sometimes more convenient to use constant-resistance equalizers—usually bridged-T networks, with one resistor and a few tuned circuits in each bridge arm—to correct for the effects of dissipation.

## VI. Accuracy Considerations

A very important aspect of the programs is the number of significant figures used in the input and output operations, and in the internal computations. All programs were written either in FORTRAN or in the Symbolic Programming System (SPS). Since FORTRAN programs require more storage, are less flexible and are limited to 8-digit precision*, such programs were only used when higher accuracy was not required. In SPS the limiting factor on the number of significant digits used was mainly the storage capacity of the memory in use. It is most convenient to be able to store the program and all partial results simultaneously without intermediate input-output operations, but this was not possible for major programs using more than 12-digit precision in computation.

Experience showed that by establishing the number of significant figures according to the following lists, satisfactory (4 - 5 digit) accuracy results both in the element values and the transfer and driving-point parameters of filters with up to 6 sections:

a, Approximation programs: 8 digit. Exceptions: Shenitzer's method and the Chebyshev passband, arbitrary stop band approximation (12 digits).

b, Synthesis and precorrection programs: 12 digits. However, the programs computing the roots of $\Lambda$ and $\Phi$ use 18-digit subroutines in order to give sufficiently accurate values for the roots. Although the ladder-expansion program (using Bader's method[32]) uses only 12 digits, this precision has proven to be barely adequate, so that the more complex networks had to be developed from both ends.

c, Network transformations: 8 digits.

d, Programs used for analysis and for the estimation of parasitic effects: 6 digits for input and output operations, 12 digits for internal computations.

Numerical checks were incorporated in all programs wherever possible. Also, test cases were worked out and attached to the program description to test the reliability of the program and the computer.

Although these precautions seem to be exaggerated, predistortion calculations showed that a correction in the third or fourth digit of the coefficients of $\Lambda$ may cause changes in the order of 0.5 db in the pass band of sharply selective filters. Similarly, bitter experience showed that an error in the eighth digit of a coefficient of $Z_{11}$ may make it unrealizable. Even if 12 significant figures are used, care must be taken to detect round-off errors in singular cases.

* A first-order precorrection for temperature effects can be obtained by choosing the temperature coefficients according to the relation $\delta_R = \delta_L = -\delta_C$.

* An improved version, FORTRAN II, which will soon be available, will not have the latter limitation.

The 6-digit type-out of results in the network analysis programs proved adequate for all practical purposes. A larger number of figures would have unduly slowed down the output operation. With 6-digit output, the time required for the calculations at one frequency varies between 5 and 25 seconds, depending on the number of transfer quantities and immittances listed, and on the network complexity.

Finally, the reliability of this method of filter synthesis must be emphasized. It was found to be practically impossible to design a filter with 3 or more sections on a desk calculator using the insertion loss method, due to unavoidable human errors and fatigue. With the numerical checks used in the computer programs, no undetected error has ever been discovered in the realized networks during several years of operation, in spite of the fact that the computer has been run on the "open-shop" basis, by a number of design engineers and technicians.

## VII. Examples

The uses of the computer programs listed above will now be illustrated by describing some practical network designs.

### A, An Elliptic Low-Pass Filter Corrected for the Effects of Dissipation

A low-pass filter was designed to satisfy the following requirements*:

Pass band limit: 85.68 kc,
Stop band limit: 94.54 kc,
Pass band ripple: 0.1 db.
Stop band discrimination: 40 db
Generator and load impedances: 600 ohms.

Since these requirements are fairly typical, the design procedure will be described in some detail.

1, The order of the network was found from a design chart to be 7. This corresponds to a 3-section network.

2, The normalized characteristic function was obtained using $\vartheta$ -functions (Para. II. A.3):

$$\Phi = \frac{\begin{aligned} &p^7 + 1.915116384p^5 + 1.1341150241p^3 \\ &+ 0.197196423351p \end{aligned}}{\begin{aligned} &0.0508885542p^6 + 0.292669982919p^4 \\ &+ 0.494215372808p^2 + 0.25806022913 \end{aligned}}$$

3, Using the programs of Para. III. A.3 and III. A.1., $\Lambda$ was obtained

$$\Lambda = \frac{\begin{aligned} &p^7 + 1.57974755008p^6 + 3.16162272252p^5 \\ &+ 3.1967976697p^4 + 3.0052610299p^3 \\ &+ 1.87144975298p^2 + 0.865855223020p \\ &+ 0.258060229130 \end{aligned}}{\begin{aligned} &0.050885542p^6 + 0.292669982919p^4 \\ &+ 0.4942153728080p^2 + 0.25806022913 \end{aligned}}$$

4, With the aid of filter tables, it was predicted that all inductance values would fall between 0.5 mH and 2 mH. Coils in this inductance and frequency range can be designed to have Q's in excess of 300. The estimation procedure of Para. V. A.3. was used, to determine whether or not precorrection would be necessary. The estimated response is shown in Fig. 3. It is apparent that predistortion was necessary to satisfy the specifications.

5, The program referred to in Para. V. B.4 was used to derive a precorrected $\Lambda$ . To achieve good selectivity, the Q of the coil producing the lowest attenuation pole was chosen to be 400, the other Q's to be 250. Then

$$\Lambda_{pred} = \frac{\begin{aligned} &p^7 + 1.56924755006p^6 + 3.14503537333p^5 \\ &+ 3.17338738793p^4 + 2.98267450900p^3 \\ &+ 1.85870601248p^2 + 0.85907832590 1p \\ &+ 0.257183016349 \end{aligned}}{\begin{aligned} &0.0507155709771p^6 + 0.291675122725p^4 \\ &0.492555408239p^2 + 0.257183016349 \end{aligned}}$$

7, The performance of the network was obtained using the analysis programs of Para. IV. B.1. The lossy response is shown in Fig. 3, the locus of the driving-point impedance in Fig. 5. The deviation of the response from the specifications is less than 0.002 db in the pass band, less than 0.2 db in the stop band.

8, The tolerances were established with the aid of the programs discussed in V. A.1., by plotting the distortion introduced by a small change in each element value and Q in turn and then using statistical considerations or counting on the most unfortunate distribution of element variations. These calculations will not be reproduced here.

9, The measured response of the filter built with the element values shown in Fig. 4 agreed, within the accuracy of measurement, with the characteristics shown in Fig. 3. A photo of the pass band response, as displayed on a visual analyzer, is shown in Fig. 6.

The computer time to realize the network and obtain its response was approximately one hour.

### B, Equalized Minimum-Inductance Band-Pass Filter

A band-pass filter was to be built to satisfy the following specifications:

Pass band: 80 - 88 kc
Stop bands: 0 - 76 kc. and 92 kc -
Pass band ripple: 0.1 db
Stop band discrimination: 70 db
Generator and load impedance: 135 ohms

To allow for anticipated parasitic effects, the synthesis was based on an 11 kc pass band width, and a ripple of 0.05 db.

The design was carried out in the following steps:

1, The order of the circuit had to be even to allow the use of a "zig-zag" transformation. An eighth-degree prototype network satisfied the specifications.

2, The characteristic function resulted from a $\vartheta$ -function approximation (II. A.3). $\Lambda$ and $Y_{11}$ were obtained using programs described in III. A.3., III. A.1. and III. B.1.

3, The low-pass prototype was found by expanding $Y_{11}$ into a mid-shunt ladder (III. B.4).

4, The band-pass network was derived by a minimum-inductance transformation (III. D.3). To obtain more convenient element values, two Norton-transformations were performed in the first section, and the first coil was tapped. The final network is shown in Fig. 7.

5, The resulting lossy network did not meet the pass band specifications and some correction was

* Some safety margin is already incorporated in these specifications.

necessary. Because of the high selectivity, however, a predistortion would have greatly increased the sensitivity to parameter variations, and hence was not attempted. The correction was achieved by cascading a constant-resistance equalizer. The design of this equalizer was carried out using a computer program that matches the equalizer loss to the required response at pre-assigned points. The resulting circuit is shown in Fig. 7, the corrected response in Fig. 8.

Computer time needed to complete this design was about 90 minutes.

### C, Chebyshev Pass Band Filters

The design process for filters with Chebyshev pass band and arbitrary stop band response is similar to that followed in the previous examples, but is preceded by:

1, the location of the attenuation poles, using special programs or templates (II. B.);

2, the derivation of $\Lambda$ from the attenuation poles and the pass band specifications (III. B.1-2).

The specifications, the circuits and the responses of a low-pass and a band pass filter are illustrated in Figs. 9-10. It took approximately one hour of computer time to carry out the design.

### D, Band Separating Filter Pair

The approximating synthesis of a front-parallel connected low-pass/high-pass filter pair will now be briefly described. The design stages are as follows:

1, The pass and stop band loss requirements are recalculated in terms of voltage ratio, assuming a constant driving-point impedance (III. C.2).

2, The voltage characteristic functions are calculated using $\vartheta$ functions (II. A.3).

3, The voltage ratios and $Y_{22}$'s of the filters are calculated.

4, The networks are obtained by ladder expansions (III. B.4). The frequency denormalization of the two circuits is carried out in such a way as to ensure an approximately constant driving-point impedance throughout the transition region.

The circuit, the measured response and the driving-point impedance of a filter-group designed by this procedure is shown in Figs. 11 - 13. The network used coils with a Q of 130.

Two hours machine time was required for this design.

### E, Filter with Special Pass and Stop Band Response

The design of filters with prescribed pass band behaviour will be illustrated through a network designed to equalize in its pass band the loss due to 20 miles of open wire and to suppress signalling tones in the stop band. The design was carried out using Shenitzer's method (II. C.5). The specified and actual responses are compared in Fig. 14. The error in the pass band was less than 0.15 db. The circuit is shown in Fig. 15.

The design (including response plotting) required two hours of computer time.

### Acknowledgement

The program library described in this paper was created by the Scientific Staff members of the Northern Electric R & D Labs., with the support of IBM scientists. Most programs were prepared by J.D. MacDonald and M. Sablatash.

### References

1, S. Darlington, "Synthesis of Reactance 4-Poles which Produce Prescribed Insertion Loss Characteristics", Jour. Math. and Phys., vol. 18, pp. 257 - 353; Sept. 1939

2, J. Zdunek, "The Network Synthesis on the Insertion Loss Basis", Proc. IEE Monograph No. 278R; Jan. 1958

3, W. Cauer, "Synthesis of Linear Communication Networks", McGraw-Hill Co., Inc., New York,1958

4, J. Meinguet, "Réalisabilité des Filtres Électriques sans Inductance Mutuelle", D.Sc. thesis, University of Louvain, Belgium, 1960

5, R. Saal and E. Ulbrich, "On the Design of Filters by Synthesis", IRE Trans. on Circuit Theory, vol. CT-5, pp. 284 - 327; Dec. 1958

6, A.J. Grossman, "Synthesis of Tchebycheff Parameter Symmetrical Filters", Proc. IRE, vol. 45, pp. 454 - 473, April 1957

7, H.J. Orchard, "Computation of Elliptic Functions of Rational Fractions of a Quarterperiod", IRE Trans. on Circuit Theory, vol. CT-5, pp. 352 - 355; Dec. 1958

8, E. Rumpelt, "Uber den Entwurf elektrischer Wellenfilter mit vorgeschriebenem Betriebsverhalten", Dr. Ing. - thesis, Technische Hochschule, Munich, Ger., 1947

9, R. Rubini, "Procedimente grafici per la risoluzione del problema di approssimazione dei filtri elettrici", Alta Frequenza, vol. 30, pp. 135 - 155; Feb. 1961

10, G.C. Temes, "The Synthesis of General Parameter Insertion Loss Filters Using a Digital Computer", AIEE Transactions, Part 1, Communication and Electronics, vol. 80, pp. 181-186

11, A. Fettweiss, "Recurrence Formulae for the Calculation of the Characteristic Function of Filters with Tchebycheff Pass-Band Behaviour", Revue H.F., pp. 230 - 239, Apr. 1960

12, A.D. Bresler, "On the Approximation Problem in Network Synthesis", Proc. IRE, vol. 40, pp. 1724 - 1728, Dec. 1952

13, F.B. Hildebrand, "Introduction to Numerical Analysis", McGraw-Hill Co., Inc., New York, 1956, pp. 258 - 312

14, J.B. Dennis, "Mathematical Programming and Electrical Networks", John Wiley and Sons, Inc. and the Technology Press, New York, 1959

15, J.H. Linke, "A Graphical Approach to the Synthesis of General Insertion Attenuation Functions", Proc. IEE, vol. 97, Part III, No. 47, pp. 179 - 187, May 1950

16, S. Darlington, "Network Synthesis Using Tchebycheff Polynomial Series", Bell System Tech. J., vol. 31, pp. 613 - 665; July 1952

17, A. Shenitzer, "Chebychev Approximation of a Continuous Function by a Class of Functions", J. Assoc. Comp. Mach., vol. 4, pp. 30 - 35; Jan. 1957

18, G. Szentirmai, "The Problem of Phase

Equalization", IRE Trans. on Circuit Theory, vol. CT-6, pp. 272 - 277; Sept. 1959

19, S. Hellerstein, "Synthesis of All-Pass Delay Equalizers" IRE Trans. on Circuit Theory, vol. CT-9, pp. 215 - 222; Sept. 1961

20, J.V. Fall, "A Digital Computer Program for the Design of Phase Correctors", ibid., pp. 223 - 236

21, B.J. Bennett, "Synthesis of Electric Filters with Arbitrary Phase Characteristics", 1953 IRE Convention Record, pt. 5, pp. 19 - 26

22, D.J. Brockington, "An Application of the "Deuce" Computer to Network Design", Marconi Rev., vol. 23, pp. 140 - 148, 1960

23, J.K. Skwirzynski and J. Zdunek, "Design of Networks with Prescribed Delay and Amplitude Characteristics", ibid, pp. 115 - 139

24, D.J. Hull, "Insertion-Loss Equalization with a Digital Computer, ibid, pp. 149 - 152

25, H.W. Bode, "Network Analysis and Feedback Amplifier Design", D. Van Nostrand Co., Inc., New York; 1945

26, L. Storch, "Synthesis of Constant-Time Delay Ladder Networks Using Bessel Polynomials", Proc. IRE, vol. 42, pp. 1666 - 1675; Nov. 1954

27, A.F. Beletskiy, "Synthesis of Filters with Linear Phase Characteristics", Telecommunications (English edition of Elektrosvaz), pp. 38 - 48; April 1961

28, D.E. Muller, "A Method for Solving Algebraic Equations Using an Automatic Computer", MTAC, vol. 10, pp. 208 - 215; Oct. 1956

29, W.E. Milne, "Numerical Calculus", Princeton U. Press, Princeton; 1949, pp. 36 - 63

30, F.B. Hildebrand, op. cit., pp. 424 - 477

31, F.L. Bauer, "Ein direktes Iterationsverfahren zur Hurwitzzerlegung eines Polynoms", Arch. elekt. Ubertr., vol. 9, pp. 285 - 290; 1955

32, W. Bader "Kettenschaltungen mit vorgeschriebener Kettenmatrix", TFT, vol. 32, pp. 119 - 125, 144 - 147; June-July, 1943

33, E.L. Norton, "Constant Resistance Networks with Application to Filter Groups", Bell System Tech. J., vol. 16, pp. 178 - 183; April, 1937

34, G.C. Temes and J.D. MacDonald, "The Approximating Design of a Band Separating Filter-Pair on the Operating-Loss Basis", Northern Electric Co., Ottawa, Can., Tech. Memo. No. 8230-9, 1960

35, G.C. Temes, "Optimal Selectivity of Harmonic Separating Filter Sets", to be published in the June, 1962 issue of IRE Trans. on Circuit Theory

36, F.S. Atiya, "Theorie der maximal-geebneten und quasi-Tchebyscheffschen Filter", Arch. elekt. Ubertragung, vol. 7., pp. 441 - 450; Sept. 1953

37, S.B. Cohn, "Direct-Coupled Resonator Filters", Proc. IRE, vol. 45, pp. 187 - 196; Feb. 1957

38, K. Yamamoto, K. Fujimoto and H. Watanabe, "Programming the Minimum-Inductance Transformation", IRE Trans. on Circuit Theory, vol. CT-9, pp. 184 - 191; Sept. 1961

39, A. Ralston and H.S. Wilf (ed.), "Mathematical Methods for Digital Computers", John Wiley and Sons, Inc., New York, 1960; para. 26, "Network Analysis" by T.R. Bashkow

40, D.T. Bell, "Digital Computers as Tools in Designing Transmission Networks", 1957 IRE Convention Record, pt. 2, pp. 145 - 153

41, K.W. Henderson and W.H. Kautz, "Transient Responses of Conventional Filters", IRE Trans. on Circuit Theory, vol. CT-5, pp. 333 - 347, Dec. 1958

42, G.C. Temes, "A Method for the Estimation and Precorrection of Losses in Terminated LC-Networks", Proc. NEC, vol. 17, pp. 98 - 110; 1961. Also, "First-Order Estimation and Precorrection of Parasitic Effects in Filters", to be published

43, H.J. Orchard, "Predistortion of Singly-Loaded Reactance Networks", IRE Trans. on Circuit Theory, vol. CT-7, pp. 181 - 182; June 1960

44, P.R. Geffe, "A Note on Predistortion", IRE Trans. on Circuit Theory, vol. CT-6, p. 395, Dec. 1959, also Editor's Note, ibid, pp. 395 - 396

45, M. Dishal, "Design of Dissipative Band-Pass Filters . . . .", Proc. IRE, vol. 37, pp. 1050 - 1069; Sept. 1949

46, C.A. Desoer, "Network Design by First-Order Predistortion Technique, IRE Trans. on Circuit Theory, vol. CT-4, pp. 167 - 170; Sept. 1957

47, C.A. Desoer and S.K. Mitra, "Design of Lossy Ladder Filters by Digital Computer", IRE Trans. on Circuit Theory, vol. CT-9, pp. 192 - 201; Sept. 1961

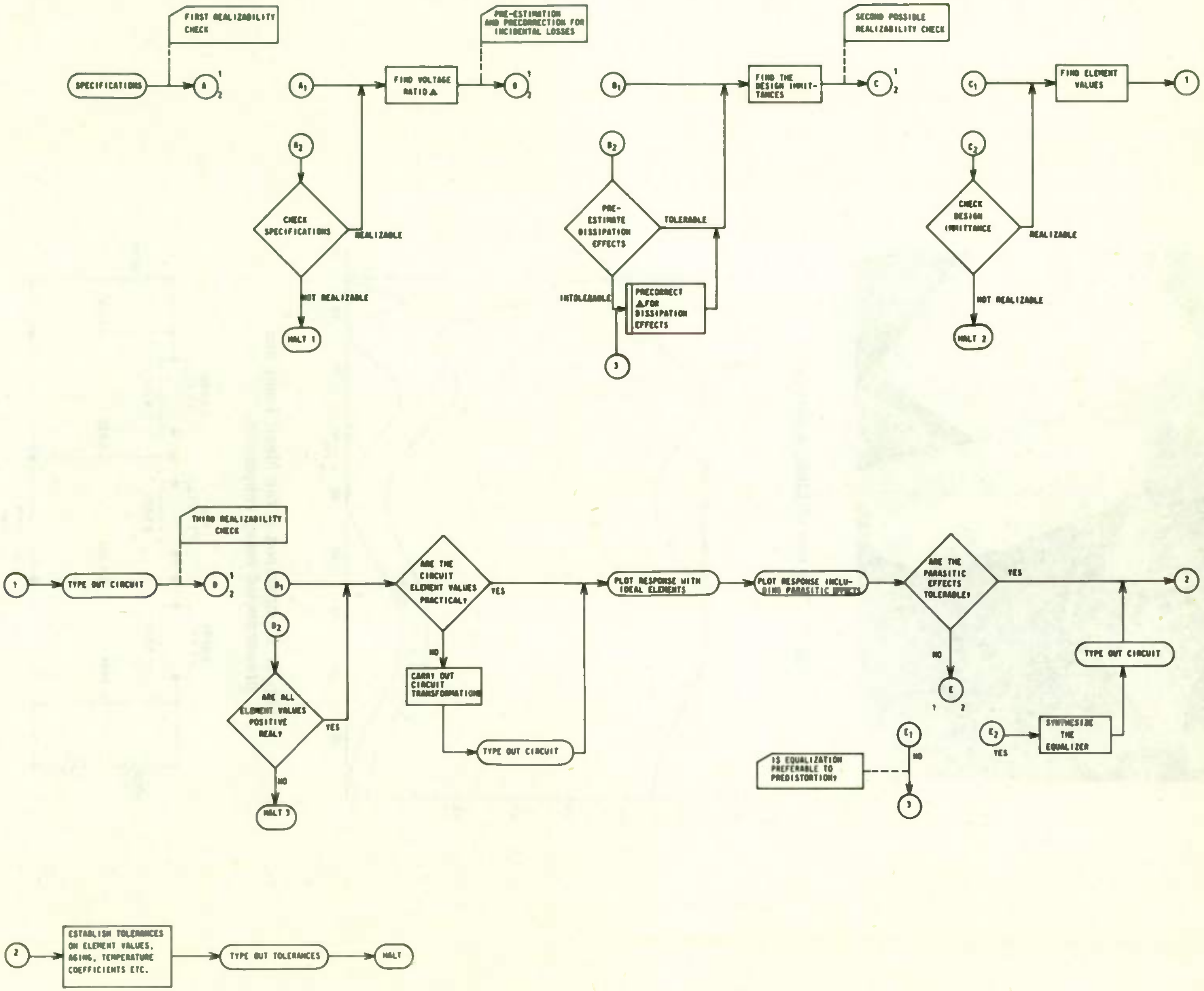

FIRST REALIZABILITY CHECK
SPECIFICATIONS
A
1
2
A1
A2
CHECK SPECIFICATIONS
REALIZABLE
NOT REALIZABLE
HALT 1
FIND VOLTAGE RATIO Δ
PRE-ESTIMATION AND PRECORRECTION FOR INCIDENTAL LOSSES
B
1
2
B1
B2
PRE-ESTIMATE DISSIPATION EFFECTS
TOLERABLE
INTOLERABLE
PRECORRECT Δ FOR DISSIPATION EFFECTS
3
FIND THE DESIGN IMMIT-TANCES
SECOND POSSIBLE REALIZABILITY CHECK
C
1
2
C1
C2
CHECK DESIGN IMMITTANCE
REALIZABLE
NOT REALIZABLE
HALT 2
FIND ELEMENT VALUES
1
1
TYPE OUT CIRCUIT
THIRD REALIZABILITY CHECK
D
1
2
D1
D2
ARE ALL ELEMENT VALUES POSITIVE REAL?
YES
NO
HALT 3
ARE THE CIRCUIT ELEMENT VALUES PRACTICAL?
YES
NO
CARRY OUT CIRCUIT TRANSFORMATIONS
TYPE OUT CIRCUIT
PLOT RESPONSE WITH IDEAL ELEMENTS
PLOT RESPONSE INCLU-DING PARASITIC EFFECTS
ARE THE PARASITIC EFFECTS TOLERABLE?
YES
NO
2
E
1
2
E1
NO
IS EQUALIZATION PREFERABLE TO PREDISTORTION?
3
E2
YES
SYNTHESIZE THE EQUALIZER
TYPE OUT CIRCUIT
2
ESTABLISH TOLERANCES ON ELEMENT VALUES, AGING, TEMPERATURE COEFFICIENTS ETC.
TYPE OUT TOLERANCES
HALT

Fig. 1. The flow chart of a general filter synthesis program.

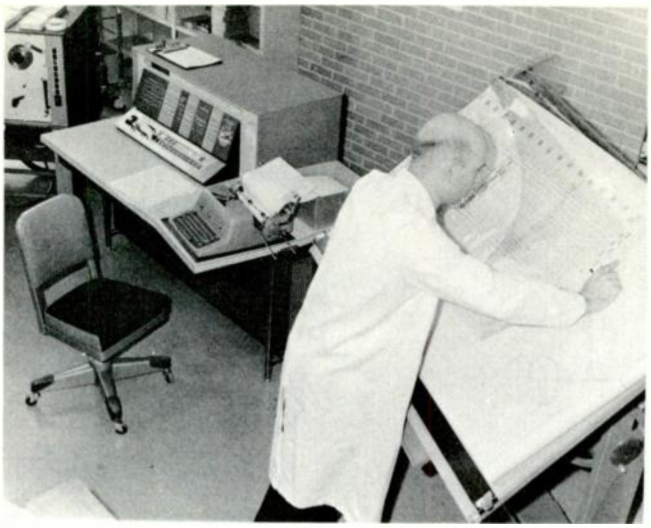


Fig. 2. The use of Linke's curves.

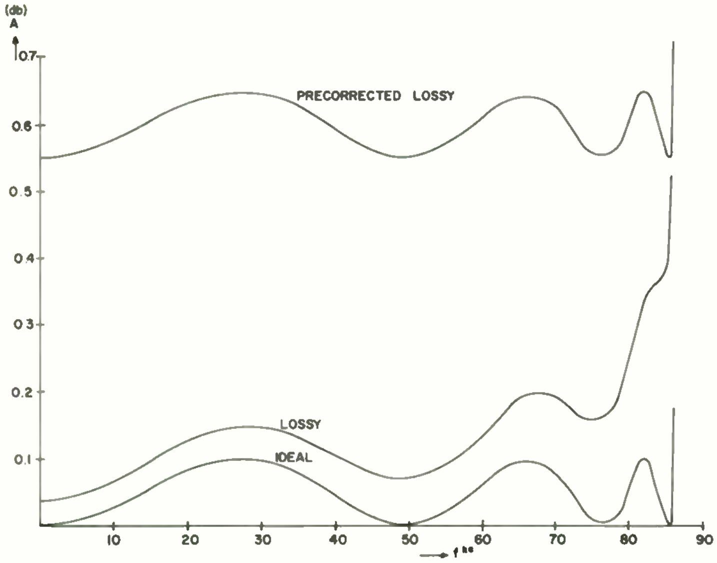


Fig. 3. Comparison of the ideal, lossy and precorrected lossy responses.

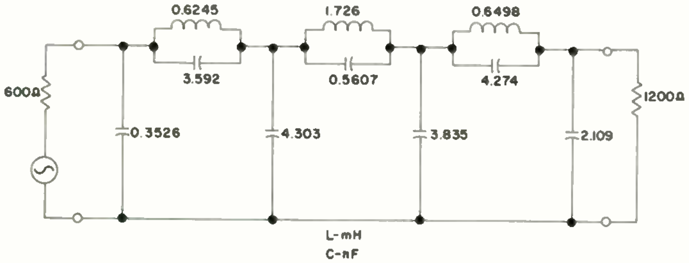


Fig. 4. Precorrected low-pass filter.

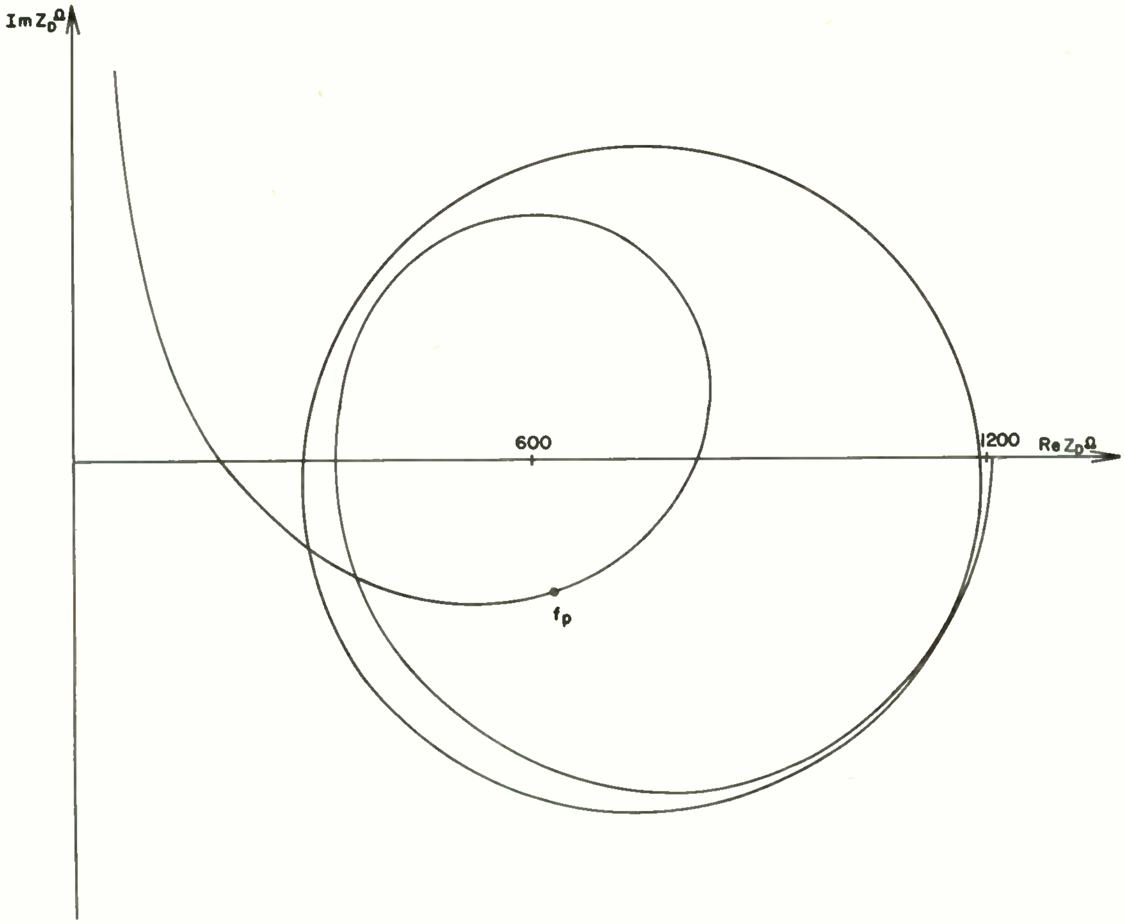


Fig. 5. The driving-point impedance of a precorrected filter.

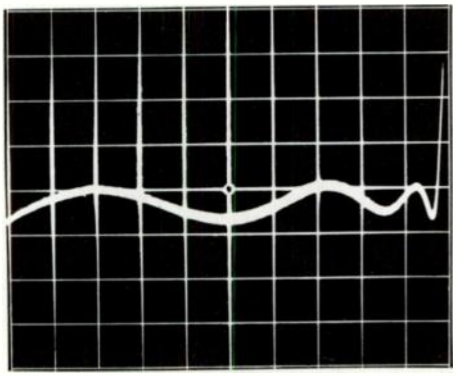


Fig. 6. The measured pass band response. (one division is about 0.1 db)

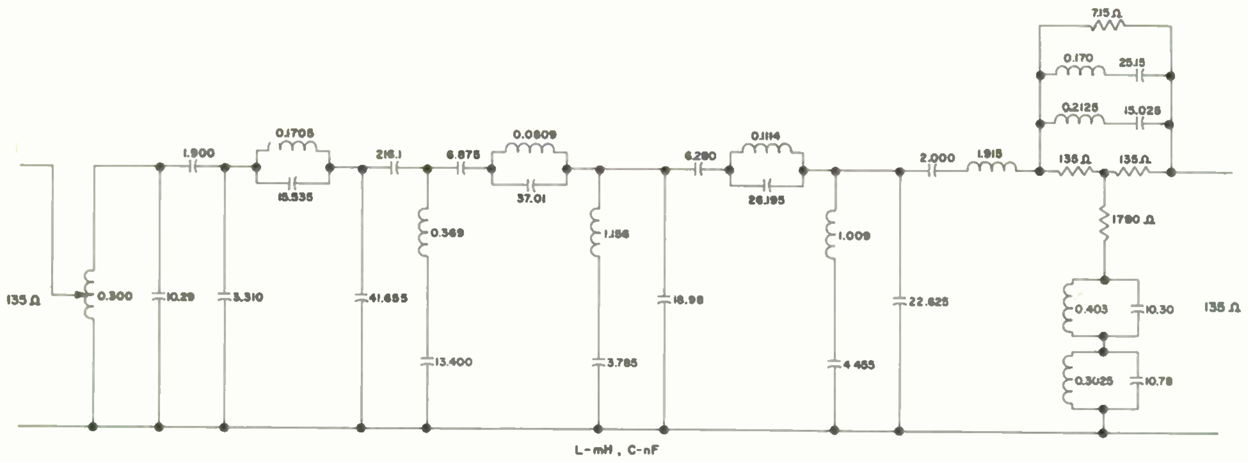


Fig. 7. A minimum-inductance band-pass filter with constant-resistance equalizer.

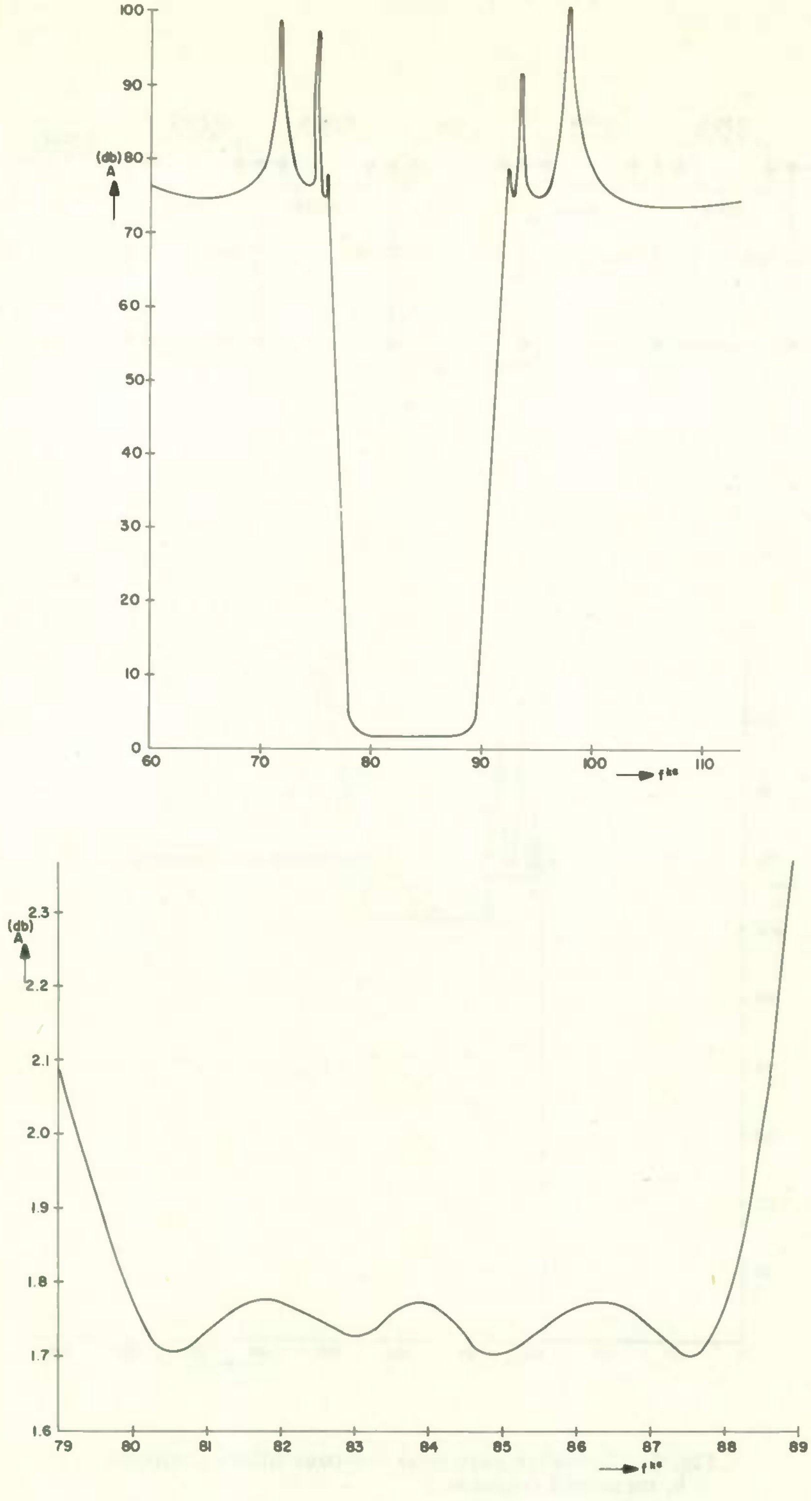


Fig. 8. The response of the band-pass filter and equalizer.

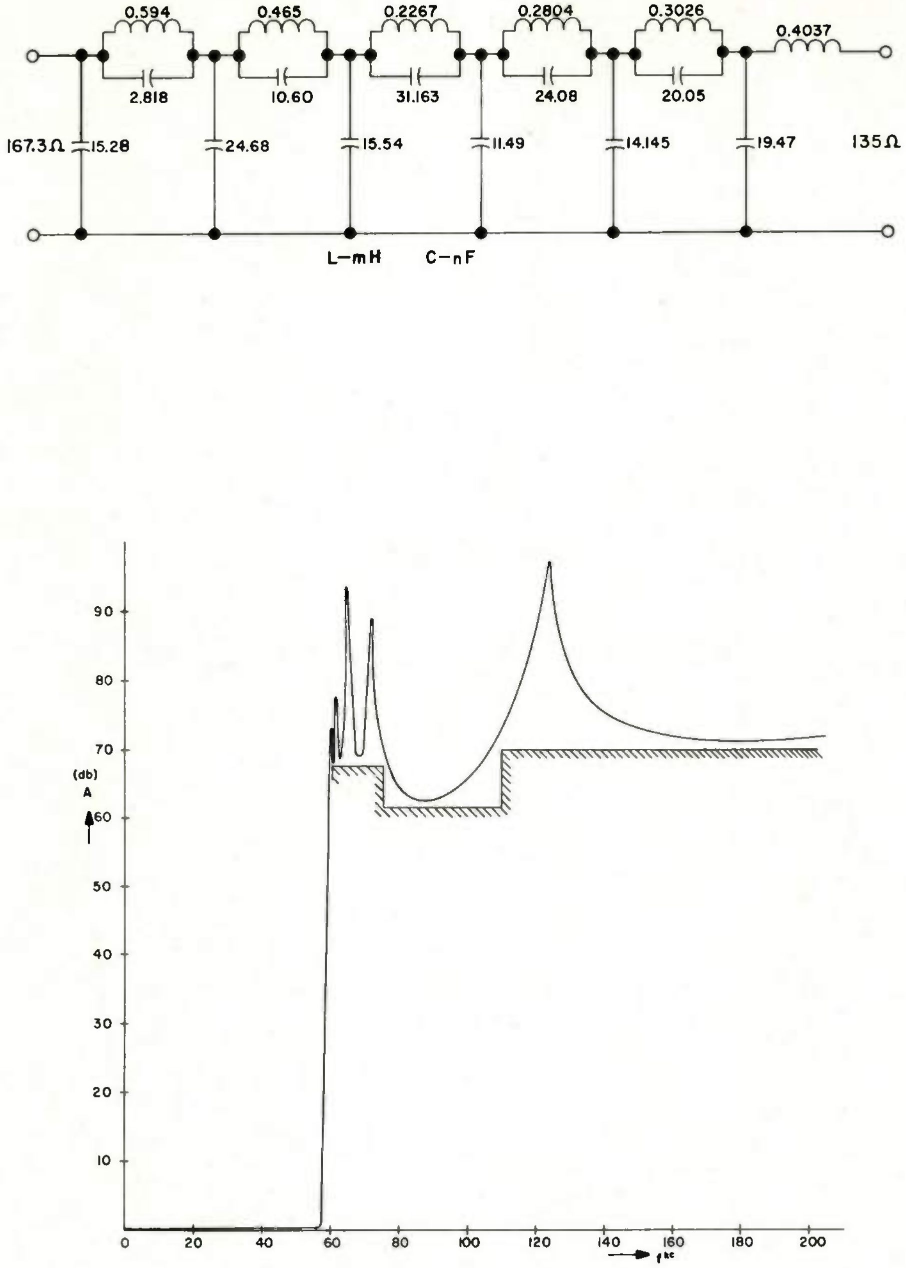


Fig. 9. Chebyshev pass band low-pass filter: a, circuit, b, measured response.

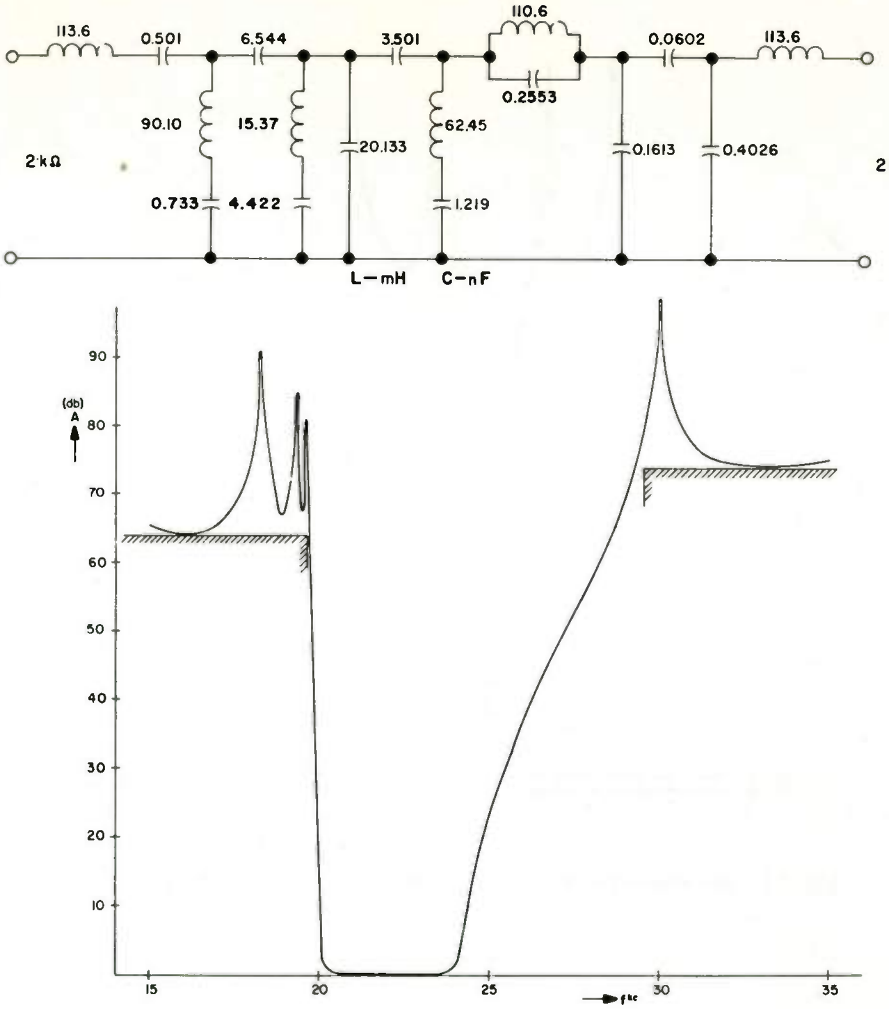


Fig. 10. Chebyshev pass band band-pass filter: a, circuit, b, measured response.

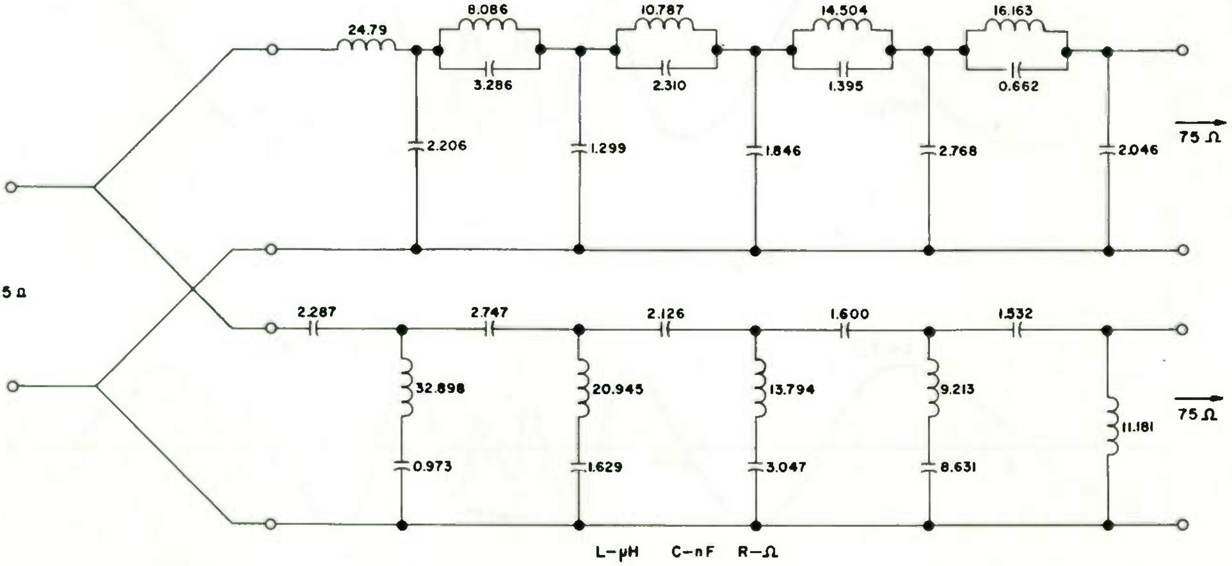


Fig. 11. A band separating filter pair.

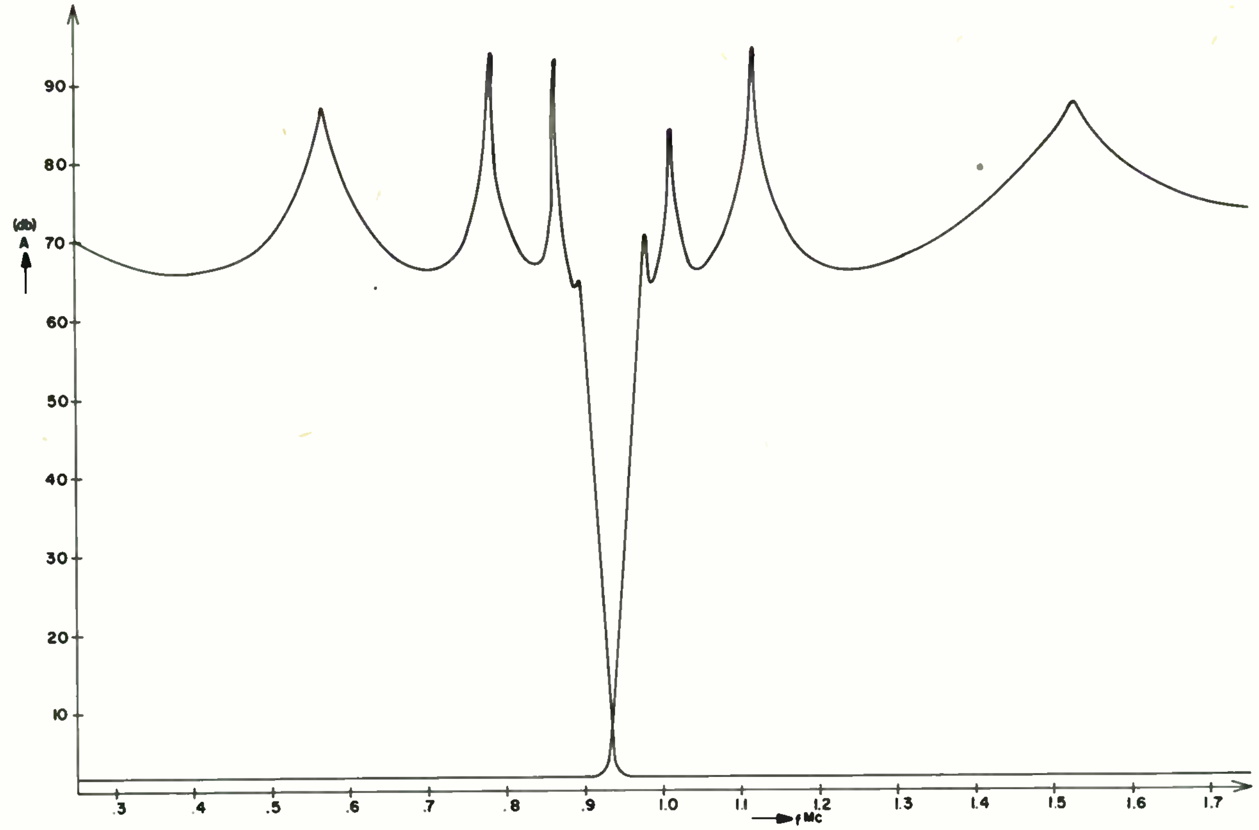


Fig. 12. The measured response of the band separating filter-pair.

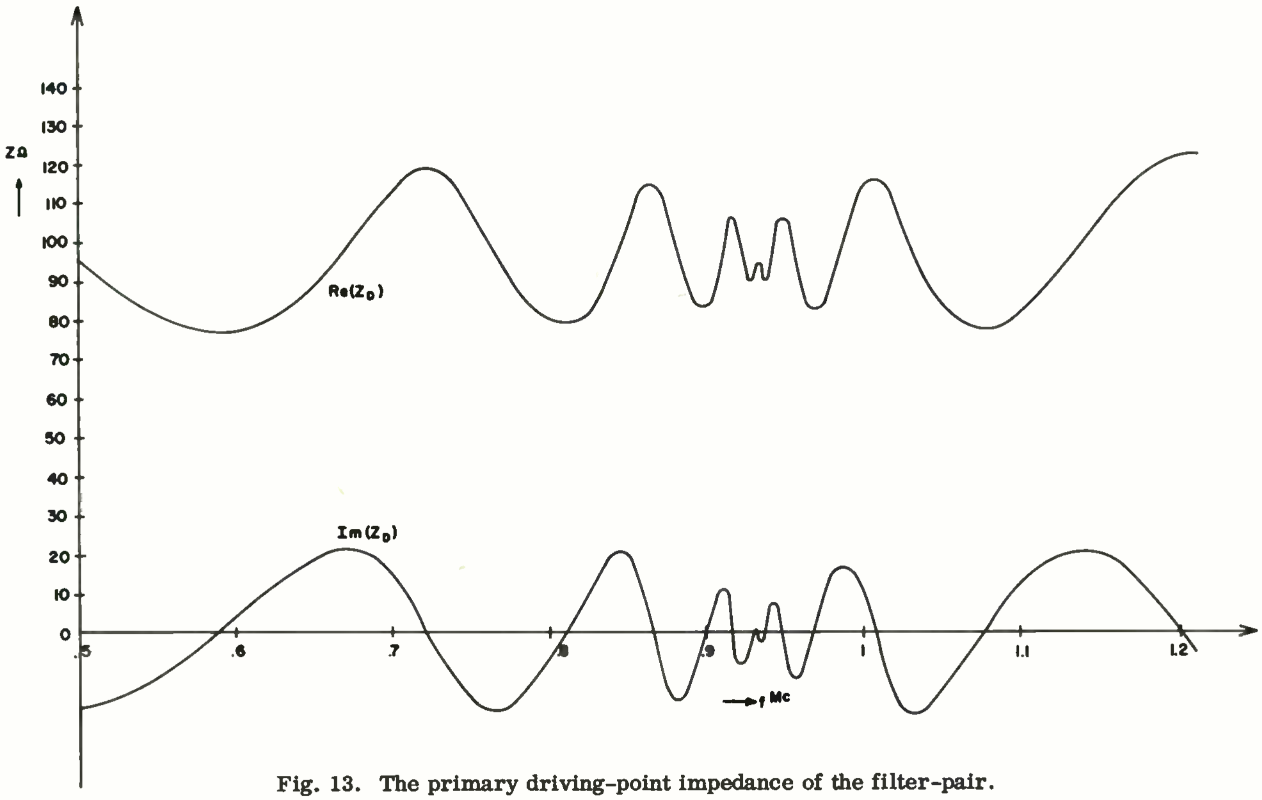


Fig. 13. The primary driving-point impedance of the filter-pair.

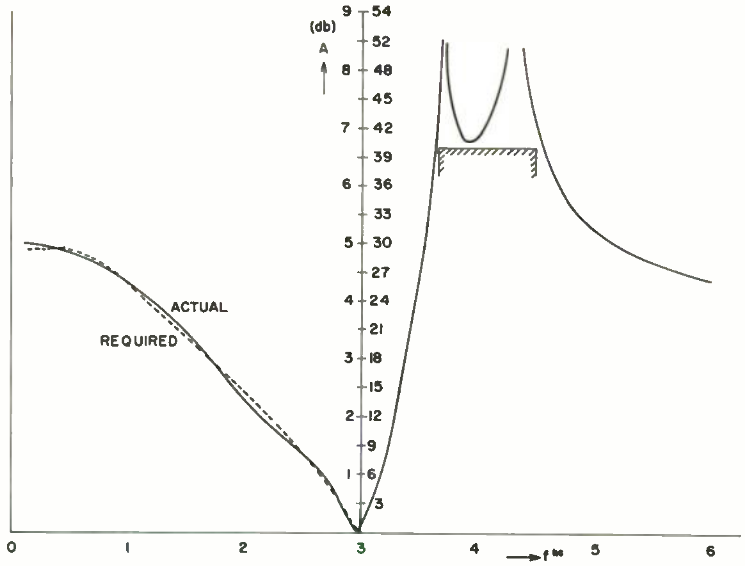


Fig. 14. Specified and actual response of a filter with arbitrary pass band attenuation and prescribed poles.

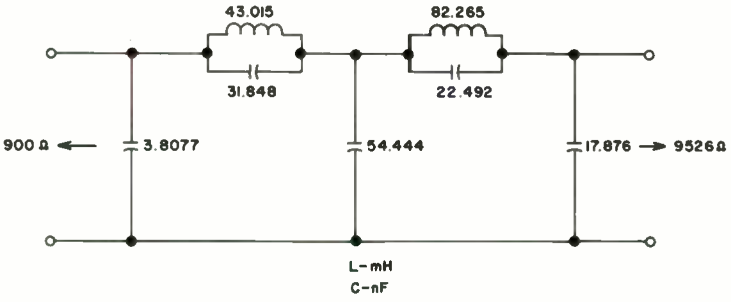


Fig. 15. The circuit of the filter.

# DESIGN OF TRANSISTOR FEEDBACK AMPLIFIERS AND AUTOMATIC CONTROL CIRCUITS WITH THE AID OF A DIGITAL COMPUTER

Omer P. Clark
Bell Telephone Laboratories, Incorporated
North Andover, Massachusetts

## 1. Introduction

A method of designing transistor feedback transmission amplifiers with the aid of a recently developed nodal analysis digital computer program is described. A procedure for using this nodal analysis program during the design of automatic control systems is also included. This computer program will be referred to as NAPANS for "Nodal Analysis of Passive and Active Networks" in the following discussion.

This paper consists of three main parts: The first part explains the functional requirements for transmission amplifiers, the basic terminology used in feedback circuit design, and the general circuit configuration used for transmission amplifiers under development in the 60-kc to 4-mc frequency band. The second part of this paper describes the nodal analysis computer program used to compute open- and closed-loop amplifier frequency response and input and output impedance. The major problem encountered during the development of this program was obtaining transistor and transformer characteristics in a form suitable for a nodal analysis program. The method of doing this with the aid of small computer programs is explained. The third and final part of this paper explains how the circuit engineer prepares the circuit information for the computer programmer. The outstanding feature of the nodal analysis program is shown to be the small amount of effort required by the engineer to prepare the circuit for analysis. The results obtained with the NAPANS program are compared with laboratory measurements for a 2-stage transistor feedback amplifier and a phase-locked oscillator control circuit.

## 2. Transmission Amplifier and Feedback Circuit Terminology

A conventional method of indicating an amplifier with associated coupling circuits to its feedback beta circuit is shown in Fig. 1a.[1] This arrangement is particularly useful in the design of transformer-coupled amplifiers discussed in this paper. The amplifier has a forward gain characteristic indicated by $\mu$. A fraction of the output signal is fed back to the input through a network indicated by $\beta$. The over-all feedback amplifier gain is given by the equation $\mu/1\text{-}\mu\beta$ as shown in Fig. 1b.

Another form of writing the feedback amplifier equation, commonly used for automatic control systems, is indicated in Fig. 1c. Note that a positive sign is normally used in the denominator. The forward gain of the circuit is given as $G(j\omega)$ for frequency response and is indicated by $G(s)$ for Laplace transforms when analyzing the circuit for transient response.[2]

When the design requirements for a feedback amplifier are specified, a general design graph is drawn as shown in Fig. 1b to indicate the open- and closed-loop gain requirements. Past experience has indicated that over 30 db of loop feedback is required to reduce modulation products of the amplifier. A polar plot of a typical $\mu\beta$ gain [also called GH(s) gain] versus phase is given in Fig. 1c. The dashed line indicates network shaping to obtain more feedback gain while preserving phase margin.

A very useful procedure to follow while making manual calculations of a transistor circuit is given in Fig. 2.[3] The first step is to obtain transistor and transformer data at their approximate center values for the equations 1 to 6. Note that these are general equations that must be modified for local shunt or series feedback as given in reference 3. The resulting calculations using these equations will give response curves very close to the desired values of the type shown in Fig. 2. That is, the initial shaping of the $\mu\beta$ curves and phase characteristic curves should be done manually. The method of calculating break-point frequencies and obtaining asymptotic approximation to Bode cutoff is given in another paper.[4]

A typical general purpose 2-stage transistor transmission amplifier for the 60-kc to 4-mc frequency band is shown in Fig. 3. This amplifier provides an output of 300 milliwatts with 20-db closed-loop gain and more than 30 db of loop feedback. The $\mu\beta$ gain curve for a practical amplifier design must have a slope not greater than an average of 10 db per octave for a phase margin of 30 degrees. This means that with 30 db of loop gain the $\mu\beta$ and phase response curves must be controlled to a frequency greater than 32 mc.

A circuit description of the amplifier shown in Fig. 3 is as follows: The bias circuit for Q1 and Q2 is provided by the voltage drop across CR1. This aids in controlling the amplifier bias point for large variations of supply voltage.

T1 and T2 provide high-side bridge feedback[1] in the amplifier. This type of feedback is used to control the input and output impedances of the amplifier for large variations of transistor parameters. C3 and C5 are used for final adjustment on input and output return loss, and are used in conjunction with C7 to shape the $\mu\beta$ gain and phase frequency response curves. The transformer turns ratios of T1 and T2 are selected to give $1/\beta$ for final amplifier gain and to match the transistor impedance to the transmission line. These impedance ratios are also selected to provide an optimum noise figure at the amplifier input and to give optimum amplifier power output.

The 2-stage transistor amplifier in Fig. 3 is particularly useful for high-level output power. The phase reversal for feedback is obtained from the emitter of Q2, which serves as a low impedance in series with R7 to give the proper termination to the feedback winding of T2. A different method of obtaining the reversed phase for negative feedback in a 3-stage amplifier is used as shown in Fig. 8. This 3-stage amplifier is useful as a low-noise, low input level amplifier, and can serve as a good preamplifier for the 2-stage amplifier shown in Fig. 3.

## 3. Description of the Nodal Analysis Digital Computer Program

The nodal analysis program will reduce any general network, through 24 nodes, to from two to ten ports and compute the following outputs:

(a) Short-circuited admittance parameters.

(b) Open-circuited impedance parameters.

(c) Current and voltage scattering parameters.

(d) Transistor Y-parameters from hybrid parameters.

(e) ABCD parameters for two ports.

(f) Y-parameters from ABCD parameters.

(g) $Z_{IN}$ and $Y_{IN}$ at each port.

The required input data for the nodal analysis program is as follows:

(a) Number of ports.

(b) Number of nodes.

(c) Number of frequencies.

(d) Configuration of each branch.

(e) Node connections.

(f) Element values.

(g) Input matrices, if any.

(h) Output parameters.

The restrictions on the program are:

(a) Maximum number of nodes is 24 (including the reference node).

(b) Network must have a nonsingular admittance matrix.

(c) Maximum number of ports is ten.

(d) A special matrix is required for transformers.

(e) For numbering nodes:

(1) Reference or ground 0

(2) Inputs and outputs 1, 2, ....,p

(3) Remaining nodes p+1,.....,n

(f) If scattering parameters are to be computed, a termination must be given for each port.

The above tabulations summarize the capabilities and limitations of the nodal analysis program. Many of the required network calculations can be made using simple mesh computer programs. However, to obtain over-all closed-loop gain and input-output impedances, it is convenient to use a nodal analysis program.

The method used by NAPANS to reduce a network for analysis is shown in Fig. 4. First, the node points are assigned and the computer organizes the network into an admittance matrix as indicated in Fig. 4. This matrix may consist of as many as 23 rows and 23 columns. The computer reduces the matrix a row and a column at a time until two of each remain. These represent the network in terms of an input port and an output port.

The transformer characteristics are measured as indicated in Fig. 5a. A special computer program is used to convert the test data into the form shown in Fig. 5b. The resistance and capacitance elements of the transformer are handled as circuit elements and the inductances of the transformer are arranged into a matrix, Fig. 5c, that handles the signal polarities and turns ratios. NAPANS uses this matrix directly, and signal polarities are obtained by the sequence of numbering the matrix as indicated in Fig. 5b.

Special measurements are required to obtain accurate transistor parameters at high frequencies. The method of measuring the transistor and calculating its h parameters is shown in Fig. 6. Transmission line measuring techniques[5] are used to measure special characteristics of the transistor. These measurements provide $Z_{IN}$ with the output terminated in g (50 ohms), $Y_o$ (output admittance) with the input terminated in r (50 ohms). These measurements also provide $S_{21}$ (the forward voltage gain) and $S_{12}$ (the reverse voltage transfer ratio). A special computer program is used to solve for the transistor h parameters from these four measurements by the equations given in Fig. 6.

After the proper transformer, transistor, and circuit element parameters have been provided for the program and the program has reduced the network to an input and output port, a scattering matrix calculation is made on the network by the computer program. For feedback amplifiers, this network is reduced to the form shown in Fig. 7. The results of the scattering current and voltage calculations provide data to calculate the network open- and closed-loop gain, $\mu\beta$ curves and phase angle, and input and output impedances. The gain is given in decibels and degrees, and the input and output impedances are given in complex numbers.

## 4. Application of NAPANS to Amplifier Design

One of the main features of the nodal analysis computer program is the ease in which the circuit engineer can use it. Node points may be assigned to the complete circuit including the power supply, if the number of nodes is less than 23, or the circuit may be simplified to include only the ac circuits as shown in Fig. 8 for a 3-stage transistor amplifier. This circuit arrangement shows the transformers in their proper form, and requires the present maximum of 23 nodes for analysis. In many cases, node points can be saved by rearranging the circuit by the use of equivalent networks.

After the engineer has drawn the circuit to be analyzed, assigned node points, provided element values and transformer and transistor data, a digital computer programmer writes instructions for punching approximately 100 cards for the computer. This quantity varies according to the number of nodes and the frequencies computed. The NAPANS program is written for the IBM 7090 computer, and requires 32,000 words of core storage. The approximate computing time for 25 frequency points, and 24 nodes is 1 minute on the IBM 7090 computer.

The circuit shown in Fig. 8 is modified for $\mu\beta$ gain and phase measurements by terminating the input and output and applying signal input at terminal 23, terminated by a resistance. The output is obtained at node 22. Fig. 9 shows the computed data compared with the measured data for amplifier open- and closed-loop gain versus frequency response. The small variation in results is due to the problem of obtaining exact transistor and transformer data. The computed and measured $\mu\beta$ gain and phase results are shown in Fig. 10. After the engineer receives the calculated results from the computer, he selects networks to improve the shape of the $\mu\beta$ gain and phase curves and repeats the computer calculations. The networks also affect the closed-loop gain flatness of the amplifier; therefore, considerable experience is required in the design of feedback amplifiers to select networks that will give the desired gain flatness while still preserving the proper $\mu\beta$ gain and phase stability margins.

At the present time, the NAPANS program will tell the engineer precisely and quickly how wise his choice was in specifying a compensating network. Programming work is continuing to have this network selection performed automatically.

## 5. Application of NAPANS to Automatic Control System Design

The NAPANS program is ideally suited for use in the frequency analysis of automatic control systems. First, each functional part of the automatic control circuit is analyzed by the nodal analysis program method to check laboratory measurements with its ac equivalent circuit. Then, the complete ac equivalent circuit for the automatic control circuit is combined into one final circuit and analyzed in the manner outlined for the 3-stage transistor feedback amplifier, Fig. 8. A block diagram of a frequency phase-locked oscillator system[6] is shown in Fig. 11. The transient response characteristics of the automatic control system are obtained from the frequency response data in the conventional manner. The ac equivalent circuit used by NAPANS to analyze the phase-locked oscillator is given in Fig. 12.

## 6. Conclusion

A nodal analysis digital computer program that is capable of analyzing closed-loop active networks, with less than 24 nodes, has been developed. This program is intended to serve as an aid to the engineer while designing feedback circuits. The program calculates precise frequency response data on complete closed-loop feedback circuits. Information regarding transient response and stability is obtained from this frequency response data by conventional methods.

This computer program has been given the title NAPANS for "Nodal Analysis of Passive and Active Networks." It is a general-type program that can be used on any type of active or passive circuit, provided the transformer and transistor data is arranged in the proper form as outlined in this paper.

The results from NAPANS, for frequencies less than 20 megacycles, check within 0.1 db and 10 degrees in phase with laboratory measurements for amplifier open- and closed-loop frequency response and $\mu\beta$ gain and phase. The method of analyzing feedback amplifier circuits is also applied in a similar manner to the design of automatic control circuits.

## Acknowledgement

The author wishes to acknowledge the assistance contributed by Mrs. G. L. Ebbe and Miss B. A. Bielko in the development of the NAPANS program and the guidance of R. S. Graham during the design of the transistor transmission amplifiers.

## References

1. H. Bode, "Network Analysis and Feedback Amplifier Design," D. Van Nostrand, Inc., New York, N. Y. (1945).

2. J. A. Aseltine, "Transform Method in Linear System Analysis," McGraw-Hill (1958).

3. R. Riddle and M. Ristenbatt, "Transistor Physics and Circuits," Prentice-Hall, Inc., (1958).

4. F. H. Blecher, "Design Principles Single-Loop Transistor Feedback Amplifier," IRE Transactions on Circuit Theory, Vol. CT-4 (September, 1957), pp. 145-156.

5. D. Leed and O. Kummer, "A Loss and Phase Set for Measuring Transistor Parameters and Two-Port Networks Between 5 and 250 mc," B. S. T. J. Vol. XL (May, 1961).

6. R. D. Barnard, "Variational Techniques Applied to Capture in Phase-Controlled Oscillators," B. S. T. J. Vol. XLI (January, 1962).

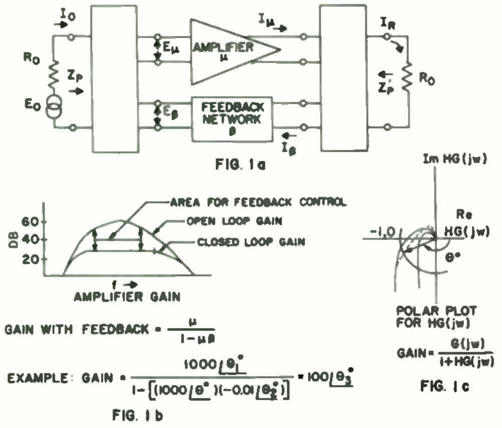


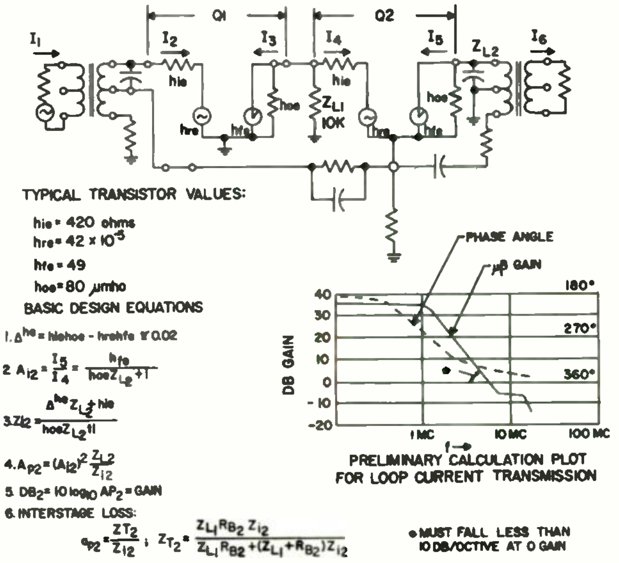
FIG. 1a

FIG. 1b

FIG. 1c

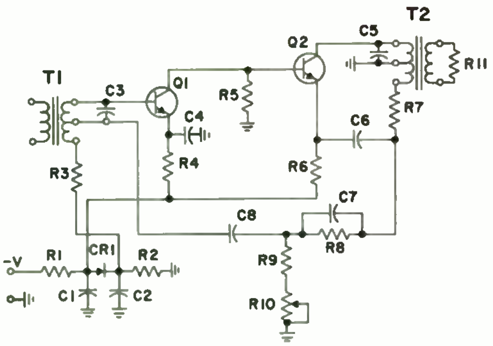
BASIC FEEDBACK AMPLIFIER DEFINITIONS

FIG. 1


Q1
Q2
I1
I2
I3
I4
I5
I6
ZL2
hie
hre
hfe
hoe
ZL1 10K
TYPICAL TRANSISTOR VALUES:
hie = 420 ohms
hre = 42 × 10^-5
hfe = 49
hoe = 80 μmho
BASIC DESIGN EQUATIONS
1. Δhe = hiehoe − hrehfe ≅ 0.02
2. Ai2 = I5/I4 = hfe / (hoeZL2 + 1)
3. Zi2 = (ΔheZL2 + hie) / (hoeZL2 + 1)
4. Ap2 = (Ai2)² ZL2/Zi2
5. DB2 = 10 log10 AP2 = GAIN
6. INTERSTAGE LOSS:
ap2 = ZT2/Zi2 ; ZT2 = ZL1RB2Zi2 / (ZL1RB2 + (ZL1 + RB2)Zi2)
PHASE ANGLE
μβ GAIN
DB GAIN
40
30
20
10
0
−10
−20
180°
270°
360°
1 MC
10 MC
100 MC
f →
PRELIMINARY CALCULATION PLOT FOR LOOP CURRENT TRANSMISSION
● MUST FALL LESS THAN 10 DB/OCTIVE AT 0 GAIN

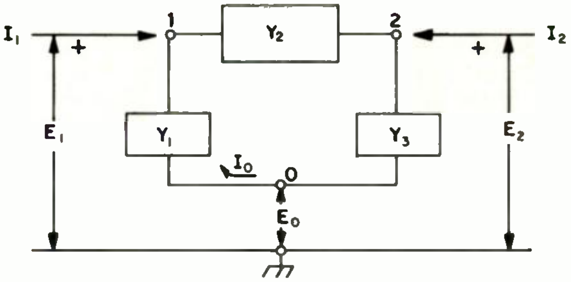
SHAPING THE HIGH FREQUENCY END OF $\mu\beta$

FIG 2



TWO STAGE TRANSISTOR TRANSMISSION AMPLIFIER

FIG. 3



NETWORK TO BE COMPUTED BY NODAL ANALYSIS PROGRAM

1. THE ADMITTANCE EQUATIONS FOR THE BRANCH BETWEEN NODES 1,0 ARE

$I_0 = Y_1 \; E_0 - Y_1 \; E_1$ ; SIMILARLY FOR BRANCHES CONNECTED
$I_1 = Y_1 \; E_0 + Y_1 \; E_1$ BY 1,2 AND 2,0.

2. WHEN ALL BRANCH PARAMETERS ARE SUMMED THEY WILL RESULT IN THE FOLLOWING ADMITTANCE MATRIX DESCRIBING THE NETWORK.

$$\begin{bmatrix} I_0 \\ I_1 \\ I_2 \end{bmatrix} = \begin{bmatrix} Y_1 + Y_3 & -Y_1 & -Y_3 \\ -Y_1 & Y_1 + Y_2 & -Y_2 \\ -Y_3 & -Y_2 & Y_2 + Y_3 \end{bmatrix} \begin{bmatrix} E_0 \\ E_1 \\ E_2 \end{bmatrix}$$

## NETWORK AND MATRIX FOR NODAL ANALYSIS PROGRAM

FIG. 4

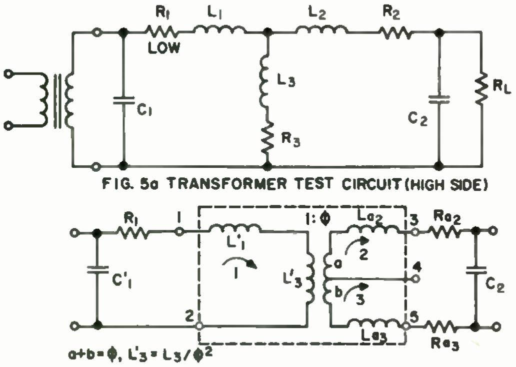


FIG. 5a TRANSFORMER TEST CIRCUIT (HIGH SIDE)

$a+b=\phi, \; L'_3 = L_3/\phi^2$

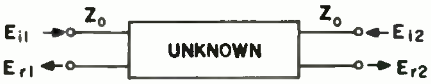
FIG. 5b TRANSFORMER ARRANGED FOR NODAL PROGRAM

$$\begin{bmatrix} L'_1 + L'_3 & L'_3 a & L'_3 b \\ L'_3 a & La_2 + L'_3 a^2 & abL'_3 \\ L'_3 b & abL'_3 & La_3 + b^2 L'_3 \end{bmatrix}$$

FIG. 5c TRANSFORMER MATRIX USED BY NODAL ANALYSIS PROGRAM

## TRANSFORMER PARAMETERS USED BY NODAL ANALYSIS PROGRAM

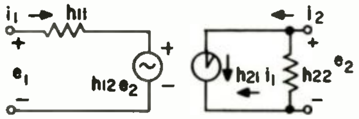
FIG. 5



$E_{r1} = S_{11} \; E_{i1} + S_{12} \; E_{i2}$

$E_{r2} = S_{21} \; E_{i1} + S_{22} \; E_{i2}$

SCATTERING PARAMETER EQUATIONS



TRANSISTOR IN H PARAMETER FORM

$Z_{IN} = h11 - \dfrac{h12 \; h21}{h22 + g}$

$Y_O = h22 - \dfrac{h12 \; h21}{h11 + r}$

$S_{21} = \dfrac{-2h21}{(g + h22)(r + h11) - h12 \; h21}$

$S_{12} = \dfrac{2h12}{(g + h22)(r + h11) - h12 \; h21}$

TYPICAL

$e_1 = h11 \; i_1 + h12 \; e_2$ ; $h21 = -0.98 = -0.18\text{db}, 180°$

$i_2 = h21 \; i_1 + h22 \; e_2$ ; $h12 = 6 \times 10^{-4} = -64\text{db}, 0°$

$h11 = 6.5$ OHMS

$h22 = 1.4 \times 10^{-6}$ MHO

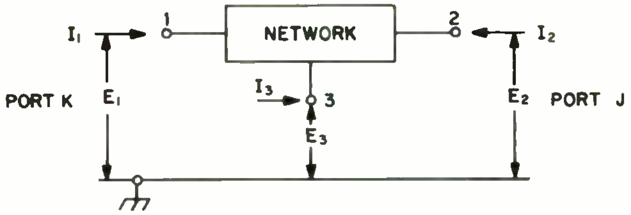
NODAL PROGRAM CONVERTS COMMON BASE VALUES TO COMMON EMITTER VALUES.

WHERE: $Z_{IN}$ = INPUT IMPEDANCE WITH OUTPUT TERMINATED IN r
$Y_O$ = OUTPUT ADMITTANCE WITH INPUT TERMINATED IN g
$S_{21}$ = FORWARD VOLTAGE INSERTION GAIN
$S_{12}$ = REVERSE VOLTAGE INSERTION GAIN

A COMPUTER PROGRAM IS USED TO SOLVE FOR h21, h12, h11, h22

## METHOD OF OBTAINING H.F. TRANSISTOR PARAMETERS

FIG. 6



THE VOLTAGE SCATTERING MATRIX FOR THIS NETWORK IS

$$\begin{bmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{bmatrix} \begin{bmatrix} E_{i1} \\ E_{i2} \\ E_{i3} \end{bmatrix} = \begin{bmatrix} E_{r1} \\ E_{r2} \\ E_{r3} \end{bmatrix} \qquad \text{WHERE } S_{11} = \frac{E_{r1}}{E_{i1}} \quad S_{21} = \frac{E_{r2}}{E_{i1}} \quad \text{ETC.}$$

WHERE $\frac{E_{rj}}{E_{ij}}$ IS THE RATIO OF THE REFLECTED VOLTAGE AT PORT J TO THE INCIDENT VOLTAGE AT PORT K WHEN ALL OTHER INCIDENT VOLTAGES ARE ZERO.

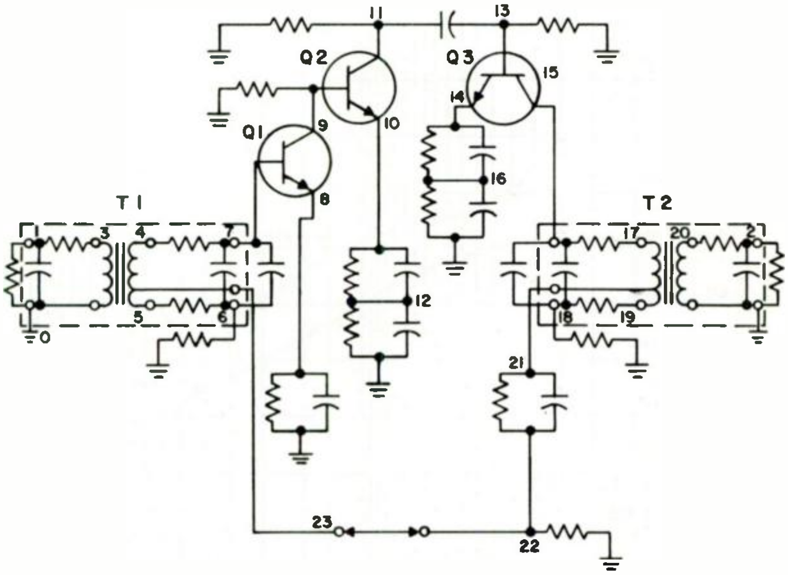
BY DEFINITION: $E_1 = E_{r1} + E_{i1}$ IF $E_{i1} = 1$, $E_{i2} = E_{i3} = 0$, THEN

$E_2 = E_{r2} + E_{i2}$ $E_1 = 1 + S_{11}$

$E_3 = E_{r3} + E_{i3}$ $E_2 = S_{21}$

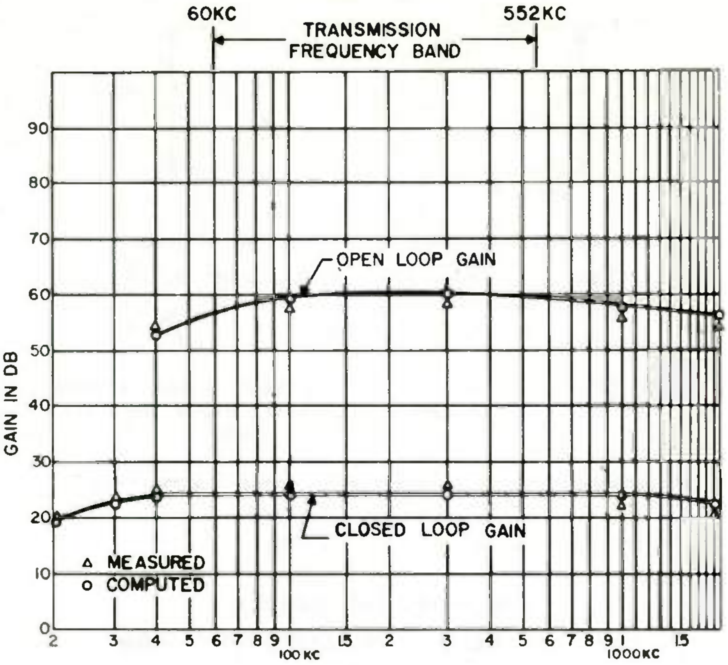
$E_3 = S_{31}$

SCATTERING MATRIX
USED BY NODAL ANALYSIS PROGRAM
FIG. 7



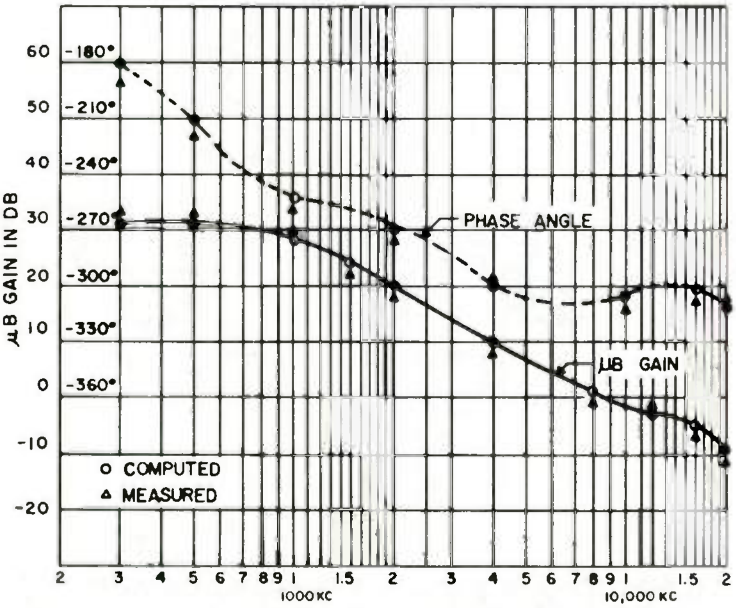
A.C. CIRCUIT OF A 3-STAGE TRANSMISSION AMPLIFIER
ARRANGED FOR NODAL COMPUTER PROGRAM

FIG. 8



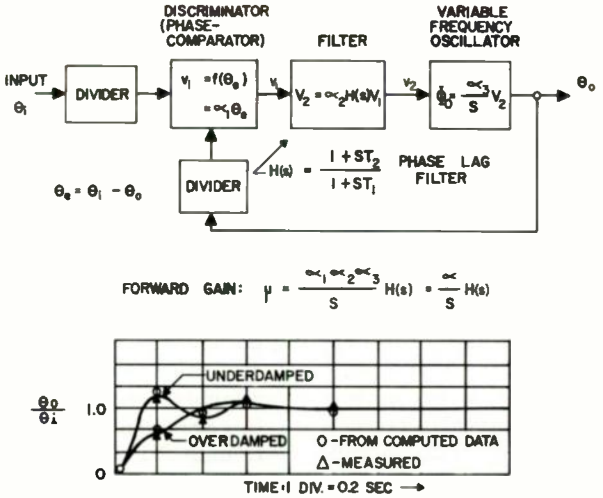
COMPUTED & MEASURED AMPLIFIER GAIN VS FREQ

FIG. 9



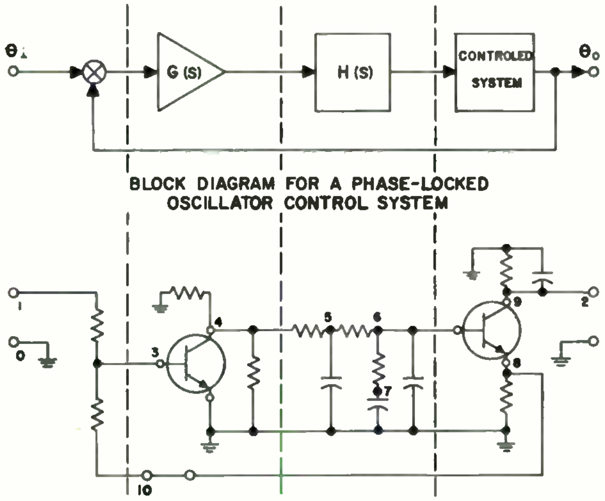
COMPUTED & MEASURED μB GAIN & PHASE

FIG. 10


DISCRIMINATOR (PHASE-COMPARATOR)
FILTER
VARIABLE FREQUENCY OSCILLATOR
INPUT $\theta_i$
DIVIDER
$v_1 = f(\theta_e) = \alpha_1\theta_e$
$v_1$
$V_2 = \alpha_2 H(s) V_1$
$V_2$
$\dot{\theta}_0 = \frac{\alpha_3}{S} V_2$
$\theta_o$
DIVIDER
$\theta_e = \theta_i - \theta_o$
$H(s) = \frac{1+ST_2}{1+ST_1}$
PHASE LAG FILTER
FORWARD GAIN: $\mu = \frac{\alpha_1\alpha_2\alpha_3}{S} H(s) = \frac{\alpha}{S} H(s)$
$\frac{\theta_o}{\theta_i}$
1.0
0
UNDERDAMPED
OVERDAMPED
O-FROM COMPUTED DATA
Δ-MEASURED
TIME: 1 DIV. = 0.2 SEC ⟶

PHASE-LOCKED OSCILLATOR DESIGNED WITH THE AID OF THE DIGITAL COMPUTER

FIG. 11



BLOCK DIAGRAM FOR A PHASE-LOCKED OSCILLATOR CONTROL SYSTEM

EQUIVALENT CIRCUIT OF THE ϕ-LOCKED OSCILLATOR FOR NAPANS

FIG. 12

# D-C AND TRANSIENT ANALYSIS OF NETWORKS USING A DIGITAL COMPUTER

Franklin H. Branin, Jr.
Development Laboratory, Data Systems Division
International Business Machines Corporation
Poughkeepsie, New York

## Summary

An experimental program is described for computing the d-c and transient response of transistor switching circuits of arbitrary configuration and size (up to 20 transistors) using the IBM 704 computer. One important feature of the program which is discussed is its ability to compile all the necessary equations automatically from input data describing the circuit parameters and configuration. Another is the solution of the transient problem by numerical integration of the differential equations for the linear part of the circuit separately from those describing the transistors, the output from each set of equations being used periodically as input for the other set. Considerable increase in speed of integration is obtained in this manner.

The method of d-c analysis is based on a topological-matrix formulation of the linear part of the problem, and its solution by Kron's method, followed by an iterative procedure for satisfying certain nonlinear side conditions imposed by the transistors. Although the transient analysis also uses a matrix formulation of the required differential equations, it is not based on a topological approach. However, a generalized topological-matrix formulation of the transient problem is given in an appendix.

The nature of a serious theoretical limit on the rate of integration of the network equations. is discussed since it constitutes the principal computational barrier to a rapid solution of the transient problem. An outline of the more important programming procedures involved in the topological-matrix formulation is also given.

Certain shortcomings of the program, and pitfalls to be avoided are pointed out. In particular, the importance of being able to modify or replace the transistor equivalent circuit (network model) is emphasized.

Finally, the computed responses of a four-transistor switching circuit are displayed and shown to agree well with the observed responses.

## Introduction

This paper is based on the experience gained in writing an experiental program for analyzing transistor switching circuits using the IBM 704 computer. This program, called TAP for "transistor analysis program",[1, 2] was developed to provide circuit-design engineers with the ability to carry out "computational experiments" to aid in understanding, as well as designing, switching circuits.

Although this objective was reached, the program has become obsolete because it was restricted to the analysis of circuits containing a certain type of diffused base transistor which is of limited interest. Consequently, the program is not being maintained nor is it being made available for outside distribution. Nevertheless, the lessons learned from writing TAP are felt to be worth sharing for the benefit of those who may be interested in writing a similar program. In addition to describing the principal features of TAP, this paper will point out the major difficulties and pitfalls encountered; it will also suggest some improvements in technique which may prove helpful. Finally, some actual and potential applications of a digital computer program of this type will be described.

TAP is capable of performing the entire d-c and transient analysis of a multi-transistor switching circuit of arbitrary configuration and size up to 20 transistors. Its most valuable feature, from the user's point of view, is its ability to compile automatically all the necessary circuit equations using a simple input card format which specifies only the parameters and connections of the circuit components. By compiling these equations for him, TAP relieves the user (ostensibly an engineer rather than a programmer) of the tedious and error-prone chore of writing all the circuit equations himself and then setting up a program to solve them. At the same time, TAP makes it convenient to modify parameters or connections in the circuit simply by changing the appropriate input cards.

It is interesting to note that the ability to compile the circuit equations from simple input data has also been included in the DYANA program recently developed by the General Motors Research Laboratories for mechanical and electrical network analysis.[3] Since this feature is particularly helpful to the user, it is strongly recommended that any general purpose network analysis programs developed in future also include this compiling facility.

The d-c analysis portion of TAP, which provides the initial values of voltage and current needed in the transient problem, takes account of the nonlinearities due to the transistors. Fortunately, it has been found possible to view this d-c problem as one of solving a linear network problem subject to an appropriate set of nonlinear side conditions. The linear network is treated by Kron's method of "interconnecting solutions"[4, 5, 6] while the nonlinear side conditions are satisfied by using an iterative method of successive approximations.[2]

After the initial conditions of voltage and current are obtained, the transient response of the network, excited by a ramp-function input pulse of arbitrary amplitude and duration, is calculated by numerical integration of the differential equations describing the system. This part of the computation is the most time-consuming and represents the biggest bottleneck in the entire analysis. In the attempt to speed up the numerical integration process in TAP, the linear and nonlinear sets of differential equations are integrated separately, over short intervals of time, and only periodically rejoined so as to provide a meaningful solution. The reasons for and advantages of doing this will be described below.

One extremely important aspect of the analsis of networks containing nonlinear devices is that of developing suitable network models for such devices. A network model must, of course, represent the device in question to an acceptable degree of approximation. It should also be made as simple as possible to minimize the computational burden. Both of these desiderata are fairly well satisfied by the transistor model used in TAP. However, an addition consideration should be pointed out whose importance was not recognized until after TAP had been written: namely, the ability to change the network representation of the transistor, either in part or in toto, without extensively altering the program.

Since TAP was intended for analyzing only circuits containing a certain type of transistor for which an adequate model existed,[7] the ability to alter this model was not considered important. Although provision was made to vary each parameter of the model, its basic configuration and the character of its nonlinearities were fixed and were made an integral part of the program. Accordingly, once TAP had been tested and proved practical for its intended purpose, attempts to extend it to the analysis of circuits containing other nonlinear devices were frustrated by the amount of reprogramming required to alter or replace the transistor model. This shortcoming, unfortunately, forced the program into premature obsolescence.

Although a network analysis program necessitates the invention of network models for nonlinear devices, by its very nature it also provides the means for validating these models. This is another reason for setting up the program in such a way as to facilitate changing these models easily. No recommendations can be made as to how this may be accomplished, but the importance of doing so needs to be recognized.

In order to conserve space in this paper, only the essence of the mathematical and programming techniques used in TAP will be given. Reference to published material will be made for any mathematical details omitted; but programming details that should be apparent to those skilled in the art will be omitted entirely. Although TAP was programmed for the IBM 704 computer, the techniques used are amenable to any binary computer and, with suitable modification, to decimal machines as well.

The three main sections of the paper describe the compilation process, d-c analysis, and transient analysis. A generalized formulation of the linear transient network problem is described in Appendix I. This formulation is recommended in place of that actually used in TAP. Other appendices are included which describe certain procedures used in determining and using the topological matrices.

## Compilation of Input Data

The scheme used in TAP for compiling the differential equations for the transient analysis of the linear part of the network is based on a matrix formulation[8] in which the coefficient matrices are determined by inspection of the network connections rather than by means of topological matrices. This formulation, using a combination of node voltages and mesh currents as variables, establishes a simultaneous system of first order differential equations, similar in form to that described by Bashkow[9]. The d-c analysis, which was incorporated into the program at a later date, makes use of a topological matrix formulation.[2, 5, 6]

A disadvantage arises from formulating the transient problem in terms of differential equations only since every node voltage or mesh current must be computed in terms of a corresponding differential equation, even though an algebraic equation might have been more appropriate. This presents no problem as far as mesh currents are concerned for these are introduced only when inductors are actually present. But since each node voltage requires a differential equation to describe it, there must be a capacitive path from each node to ground in order to define this differential equation. Hence, additional ("stray") capacitances must be inserted wherever the required capacitive path is absent in the original circuit.

Admittedly, these stray capacitances do exist in actual circuits and this is why they were included in the original formulation. However, their effect may be negligible in many cases and yet by their very presence, these capacitances may slow down the numerical integration process appreciably. Therefore, it is important to formulate the transient problem with sufficient generality to admit algebraic equations, when required, as well as differential equations.

A formulation, using a topological-matrix approach, is given in Appendix I. This formulation, which extends the recent work of Bashkow,[9] suffices for the d-c analysis as well and gives directly all the initial values required by the transient problem.

The input scheme developed for TAP permits arbitrary connections and parameters to be specified both for the linear part of the network and for the transistors. Only the transistor model is fixed in its configuration and nonlinear characteristics.[1] The input information required, therefore is the following: (1) type of circuit component, such as transistor, resistor, capacitor, or inductor; (2) serial number of each component (actually required only for transistors); (3) parameter (or parameters) pertinent to each component; (4) component connections as designated by node numbers; (5) voltage and/or current sources; (6) input pulse characteristics.

This information, punched into cards, is read into the computer and compiled either in matrices or in tables, some of which are later converted to matrix form. Since the cards specifying the linear part of the circuit are read in first, with R, L, and C cards intermixed, the compilation of the tables required for the d-c analysis goes on simultaneously with that of the coefficient matrices for the transient problem.

The information from the cards specifying the transistor parameters and connections is read in, and tabulated. Since the resistors in the transistor model are linear, these data are included in the tables for the d-c analysis. The transistors are simulated by current sources driving a linear circuit in both the d-c and transient analysis.

## D-C Analysis

The tables required by the matrix formulation used in the d-c analysis[5,6] are the following: (1) branch resistances (RDATA); (2) branch connections (RCON) showing initial and final node numbers for each branch; (3) voltage sources (EDATA) and (4) current sources (IDATA) in each branch. (In TAP, the IDATA table is omitted since the only current sources encountered are due to transistors. In a more general program, however, the IDATA table should be included.)

These tables are not converted into a matrix form of the network equations, but rather into the appropriate "solution matrices" in terms of which the solution may be computed directly. Actually, the voltages are computed from the nodal solution matrix (inverse of the nodal admittance matrix) while the currents are computed from the mesh solution matrix (inverse of the mesh impedance matrix). To be sure, either solution matrix would suffice for computing all the voltages and currents but it was felt that round-off errors would be reduced by the method adopted.

To handle networks of arbitrary configuration, this program requires an unambiguous procedure for identifying and specifying just the right number of meshes. Such a procedure has been developed with the aid of the well-known topological concepts of tree, link, and basic mesh.

A tree is defined as any network structure devoid of closed paths. The network tree is a tree which includes all nodes of the network and a link is any branch of the tree-complement. When a single link is connected to the tree, it forms a unique closed path called a basic mesh; each basic mesh contains but one link. Hence, if all the network branches are classified as either tree-branches or links, the basic meshes may be unambiguously identified and they are just sufficient in number.[5,6]

In computing the nodal solution matrix, a modification of Kron's method called the link-at-a-time (or LAT) algorithm is used. This algorithm, which is explained in reference 5, is a constructive method for obtaining the nodal solution matrix for the complete network by adding a link at a time to the network tree and modifying the corresponding nodal solution matrix at each successive step. Since the nodal solution matrix for the tree is obtained without matrix inversion and since the general formula for modification of this matrix at each step is the same, the algorithm is easy to use.

To minimize round-off errors at each successive step of this algorithm, it is desirable to chose the network tree of minimum total resistance.[5] Accordingly, the first step in the d-c analysis is to sort the RDATA table in order of increasing resistance. The RCON, EDATA and IDATA tables are then rearranged to conform to this new sequence of the network branches.

Next, starting with the branch of smallest resistance, the RCON table is examined to determine if each suceeding branch does or does not form a closed path with the then-defined tree. If a closed path is formed, the branch is classed as a link; if not, it is classed as a tree branch and added to the tree. When all the branches have been examined in this manner, the tree of minimum resistance is defined. Appendix II is a description of the tree-link sorting procedure.

The RDATA, RCON, EDATA and IDATA tables are rearranged to conform to this new classification with tree-branches in one group and links in another. The next step is to convert the RCON table into the appropriate topological matrices.

## Topological Matrices

Three topological matrices are required in the d-c analysis and all of them consist of the elements +1, -1, and/or 0 only. Using a binary computer such as the IBM 704, it is convenient to represent each such matrix element as a pair of binary digits (bits) in order to conserve storage -- and also to speed up certain parts of the computation. In TAP, the storage format used is 16-bit-pairs per word for each 16 elements of a column (or, sometimes, row) of a topological matrix. An alternative format, which has certain advantages, is to store the magnitude bits in one word and corresponding sign bits in an adjacent word.

By definition,[5, 6] the elements of the branch-node matrix are:

$a_{ij}$ = (+1, -1, 0) if the i-th branch is (positively, negatively, not) incident on the j-th node.

Now the i-th entry (word) of the RCON table corresponds to the i-th row of the branch-node matrix since both specify the initial (+) and final (-) node of the i-th branch. The RCON table is stored with the initial node number in the address field and the final node number in the decrement field of a single word. The task of converting this table to the bit-pair matrix format (stored columnwise) is therefore a straight-forward programming task that need not be elaborated.

Initially, the branch-node matrix includes the column corresponding to the datum or reference (ground) node, but later this column is deleted since it contains redundant information.[5, 6] The matrix remaining after this deletion is then designated as the A matrix. The tree portion of this matrix, $A_T$, is a square matrix whose inverse may be obtained topologically rather than computationally. The inverse of $A_T$ may be shown[5, 6] to be identical with the transpose of a matrix, $B_T$, called the node-to-datum-path matrix of the tree. The elements of this matrix are defined as follows:

$b_{ij}$ = (+1, -1, 0) if the i-th branch is (positively, negatively, not) included in the j-th node-to-datum path.

This $B_T$ matrix is used in computing both the nodal solution matrix of the tree and the branch-mesh matrix described below. The procedure for determining the $B_T$ matrix, by means of an exhaustive search of the network tree from the datum node outward, is explained in Appendix III.

The third topological matrix required in the d-c analysis is the branch-mesh matrix, designated C, whose elements are defined thus:

$c_{ij}$ = (+1, -1, 0) if the i-th branch is (positively, negatively, not) included in the j-th basic mesh.

By adopting the convention that both the orientation and ordering of the basic meshes agree with those of the corresponding links, the link portion of the branch-mesh matrix, $C_L$, turns out to be a unit matrix and does not even need to be written. The tree portion $C_T$ then, contains all the essential information about the branch mesh matrix.

Fortunately, $C_T$ is readily computed from $B_T$ and the transpose of $A_L$, the link portion of the branch-node matrix, by means of the equation[5, 6]

$$C_T = -B_T A_L^t \qquad (1)$$

This computation amounts to taking the difference of two columns of $B_T$ to get each column of $C_T$. Accordingly, it is easier to use the link portion of the RCON table than the $A_L$ matrix for this purpose since the address and decrement of each RCON word tell directly which columns of $B_T$ are to be added or subtracted. Thus, all the necessary topological matrices may be obtained directly or indirectly from the RCON table.

## Nodal Solution Matrix

The primitive impedance matrix, Z, and its inverse, the primitive admittance matrix, Y, characterize the branches of a network independently of their interconnections.[5, 6] In the d-c problem, the RDATA table corresponds to a real diagonal Z matrix and its reciprocal to a real diagonal Y matrix. The nodal solution matrix, which is the inverse of the nodal admittance matrix $A^tYA$, is obtained by means of the LAT algorithm starting with the inverse of the nodal admittance matrix for the tree, $A^t_T\ Y_T\ A_T$. Using the relation $A^{-1}_T = B^t_T$ mentioned above, it is easily shown that

$$(A^t_T\ Y_T\ A_T)^{-1} = B^t_T\ Z_T\ B_T \tag{2}$$

Hence, the nodal solution matrix for the tree may be calculated without any matrix inversion. Appendix IV describes how the triple matrix product $B^t_T\ Z_T\ B_T$ is computed, taking advantage of the diagonal nature of $Z_T$ and using the compact storage format of $B_T$.

The LAT algorithm makes use of a recursion formula which modifies a given nodal solution matrix to take account of the addition of a single link. Assuming that the impedance of the j-th link is $z^j_L$ and that this link is connected between the p-th and q-th nodes of the network, the modified nodal solution matrix, $Z^j$ is given by the equation[5]

$$Z^j = Z^{j-1} - \frac{(Z^{j-1}_{.p} - Z^{j-1}_{.q})(Z^{j-1}_{p.} - Z^{j-1}_{q.})}{Z^{j-1}_{pp} + Z^{j-1}_{qq} - Z^{j-1}_{pq} - Z^{j-1}_{qp} + z^j_L} \tag{3}$$

where $Z^{j-1}_{.p}$ and $Z^{j-1}_{p.}$ are the p-th column and p-th row of the previous nodal solution matrix, $Z^{j-1}$. Clearly $Z^0 = B^t_T Z_T B_T$.

Since all the $Z^j$ are symmetric, only the diagonal and subdiagonal elements are computed and stored, thereby conserving both computation time and memory space. Furthermore, since the tree is chosen for minimum total resistance, the link resistances will be as large as possible, thereby minimizing the denominator of the correction term in Eq. (3) and reducing round-off errors.

## Mesh Solution Matrix

Computation of the mesh-solution matrix is based on a recursion formula similar to Eq. (3). The corresponding algorithm adds one tree-branch at a time to the set of all links and modifies the mesh-solution matrix accordingly at each successive step. The starting point of this algorithm, of course, is the mesh solution matrix of the links which, fortunately, is diagonal. A discussion of this algorithm is given in reference 5.

Computation of the mesh and nodal solution matrices could, of course, have been done by the usual methods of matrix inversion and with less programming effort. However, since it was desired to evaluate the potentialities of Kron's method of interconnecting solutions, advantage was taken of the experimental nature of TAP to program and test the link-at-a-time and tree-branch-at-a-time algorithms in an actual situation. (The computational efficiency and applications of these algorithms are discussed in reference 5.)

## Solution of the Nonlinear Problem

As previously stated, the nonlinearities introduced by the presence of transistors in the network are handled by imposing a set of nonlinear side conditions on the solution of the linear network problem. This is accomplished by representing the effect of the transistors on the linear network by means of current sources at the appropriate nodes and adjusting these current sources so that certain of the voltage responses satisfy the nonlinear side conditions.

The d-c portion of the equivalent circuit used in TAP to represent a nonsaturating, diffused base PNP transistor is shown in Fig. 1. The base and collector resistances, $R_{bb}'$ and $R_{cc}'$, as well as the leakage resistance $R_c$, are assumed to be constant and so are included as part of the linear network. The only nonlinearities, then, are due to the current sources $\alpha I_h$ and $(1-\alpha)I_h$ where $\alpha$, the current gain factor, is assumed to be a nonlinear function of $I_h$. (See reference 1 for details.)

The current source $I_h$ (hole current) is exponentially related to the emitter-base voltage $V_{eb}$ by the diode equation

$$I_h = I_{es}\ (e^{\Lambda V_{eb}'} - 1) \tag{4}$$

where $I_{es}$ is the reverse saturation current and $\Lambda = q/kT = 1/.026$ volts at room temperature.

As far as the linear network is concerned, $I_h$ can be chosen arbitrarily since this choice merely specifies the current sources $\alpha I_h$ and $(1-\alpha)\ I_h$. The resulting voltage responses of the linear network may then be calculated by the matrix equation[5, 6]

$$e' = (A^t\ YA)^{-1}\ A^t\ (I-YE) \tag{5}$$

where e' is the node-to-datum voltage vector and I and E are the current source and voltage source vectors (corresponding to the IDATA and EDATA tables.)

Since the only current sources considered in TAP are those due to transistor action, one may write the matrix relation

$$A^t I = \bar{\alpha} I_h \qquad (6)$$

to define the equivalent nodal current sources depicted in Fig. 2. In Eq. (6), $I_h$ represents the vector of $I_h$ values for all transistors while $\bar{\alpha}$ represents a matrix whose only nonzero elements are $-1$, $\alpha$, and $(1-\alpha)$ placed so as to assign current sources $-I_h$, $\alpha I_h$ and $(1-\alpha) I_h$ to the appropriate nodes of each transistor model.

The voltages $V_{eb}'$ for each transistor may be computed from the node voltage relation $e_e' - e_b''$. In matrix form, this may be written as

$$V_{eb}' = He' \qquad (7)$$

where H is a matrix whose only nonzero elements are +1 and -1 placed so as to perform the appropriate linear combination of the elements of e'.

Combining Eqs. (5), (6) and (7), we may write

$$V_{eb}' = \left[H(A^tYA)^{-1}\bar{\alpha}\right] I_h - H(A^tYA)^{-1} A^tYE \qquad (8)$$

which expresses the vector $V_{eb}'$ directly as a function of $I_h$. The trick now is to choose $I_h$ so that both Eq. (8) and Eq. (5) are satisfied.

The nature of this problem is shown graphically in Fig. 3 where the diode curve represents Eq. (4) and where the load line represents the almost linear relation between $V_{eb}'$ and $I_h$ in Eq. (8) -- ignoring the slight dependency of $\bar{\alpha}$ on $I_h$. The two situations of major interest shown in Fig. 3 correspond to load lines for "on" and "off" transistors.

The iterative method of solving this problem may be explained as follows: For each transistor, an initial estimate $I_h^{(o)}$ is made. The diode curve is then approximated by a tangent line at the point $(I_h^{(o)}, V_{eb}^{(o)\prime})$ and its intersection with the load line determined. This results in a new estimate $I_h^{(o)}$ as shown in Fig. 4. The process is then repeated until convergence is obtained.

The tangent-line approximation to the diode curve is given by the point-slope formula:

$$V_{eb}' = \left[\frac{1}{\Lambda(I_h^{(o)} + I_{es})}\right] I_h + V_{eb}^{(o)\prime} - \left[\frac{1}{\Lambda(I_h^{(o)}+I_{es})}\right] I_h^{(o)} \qquad (9)$$

This relation may be considered as a matrix equation describing the tangent lines for all transistors if the quantity $\left[1/\Lambda(I_h^{(o)} + I_{es})\right]$ is regarded as a diagonal matrix.

Now for transistors operating in the "off" region, close to the asymptotic limit of $-I_{es}$, the computation of $I_h + I_{es}$ may involve serious round-off error. As a result, the larger $1/\Lambda\,(I_h + I_{es})$ becomes, the more inaccurate it is. This error can be avoided, however, by making the change of variable,

$$J = I_h + I_{es} \qquad (10)$$

and not calculating $I_h$ explicitly.

Eliminating $V_{eb}'$ from Eqs. (8) and (9) and introducing the variable J, we obtain the following equation for the iteration procedure:

$$\left[H(A^tYA)^{-1}\bar{\alpha}^{(n-1)} - \frac{1}{\Lambda J^{(n-1)}}\right] J^{(n)} = V_{eb}^{(n)} - \left[\frac{1}{\Lambda J^{(n-1)}}\right] + \left[H(A^tYA)^{-1}\bar{\alpha}^{(n-1)}\right] I_{es} + H(A^tYA)^{-1}A^tYE \qquad (11)$$

where $\bar{\alpha}^{(n-1)} = \bar{\alpha}(I_h^{(n-1)})$ is assumed to be constant. In this expression, whatever accuracy is inherent in $J^{(n-1)}$ is retained in $1/\Lambda J^{(n-1)}$.

One difficulty which arises in applying Eq. (11) is that an "out-of-bounds" intersection of the tangent-line and load line for an "off" transistor may occur, as depicted in Fig. 5. This problem is handled by testing each element of the $J^{(n)}$ vector for proper boundedness and replacing the out-of-bounds elements by the arbitrary value of $I_{es} \times 10^{-6}$.

A second difficulty is that Eq. (11) yields inaccurate J values for the "off" transistors. These values may be improved by the method of successive substitutions depicted in Fig. 6. After each iteration, the elements of the J vector. corresponding to each "off" transistor, are tested for boundedness, replaced by $I_{es} \times 10^{-6}$ if necessary, and then substituted into the equation

$$V_{eb}' = \left[H(A^tYA)^{-1}\bar{\alpha}\right] J - \left[H(A^tYA)^{-1}\bar{\alpha}\right] I_{es} - H(A^tYA)^{-1} A^tYE \qquad (12)$$

The J values are then recomputed using the expression

$$J = I_{es} (e^{\Lambda V_{eb}'}) \tag{13}$$

This process of successive substitutions converges if the slope of the load line is greater than that of the diode curve at their intersection point and if the starting point of the process is close enough to this intersection; otherwise, the process diverges. In order to guard against incipient divergence, the values of $\Lambda V_{eb}'$ are monitored and if a value in excess of 4.0 is detected, the process of successive substitutions is terminated. Otherwise, three iterations are made.

The entire procedure of successive approximations is repeated until the fractional change in the length of the J vector is $10^{-7}$, with a maximum of thirty iterations being allowed. Although no attempt has been made to establish theoretically the conditions under which convergence will be assured, it is believed that only those circuits which involve large positive feedback would be likely to cause trouble. In such cases, it may be necessary to resort to the transient analysis procedure, using whatever initial conditions are obtained from the d-c analysis, and allow the integration to proceed with no input pulse until a steady state is reached.

In the normal case, after convergence has been attained, all the voltages and currents of the network are computed, completing the d-c analysis. Not all of these data are required as initial conditions for the transient analysis, however, and so only the desired voltages and currents are selected.

## Transient Analysis

After the initial conditions have been obtained from the d-c analysis, they are verified by running the numerical integration for a short time without any input pulse. When a satisfactory steady state has been reached, the input pulse is initiated. The transient computation may then be continued as long as desired. At suitable intervals during this computation, the entire set of voltages and currents in the network is printed out. It would be preferrable to display certain portions of this information graphically, either on a printer or cathode ray tube, but this feature was not included in TAP.

The solution of the transient problem, as mentioned above, is based on integrating the linear differential equations of the network separately from the nonlinear differential equations describing the transistors. To explain the practical importance of this modus operandi, a brief outline of the basic theory of the numerical integration is needed first.

The numerical integration of the single differential equation

$$\frac{dx}{dt} = f(x, t) \tag{14}$$

may be effected by means of the generalized expression

$$x_n = \sum_{j=1}^{N} a_j x_{n-j} + h \sum_{j=0}^{M-1} b_j (dx/dt)_{n-j} \tag{15}$$

where $x_n = x(t_o + nh)$ and $h = \Delta t$ is the integration interval. If $b_o = 0$, the integration formula is called a "predictor"; if $b_o \neq 0$, it is called a "corrector." Usually, a combined predictor-corrector scheme is employed with the coefficient $a_j$ and $b_j$ selected to give the accuracy desired.

In the special case of a constant coefficient, linear differential equation, where $f(x, t) = gx$, Eq. (15) becomes

$$(1 - hgb_o) x_n = (a_1 + hgb_1) x_{n-1} \tag{16}$$

$$+ (a_2 + hgb_2) x_{n-2} + \ldots + (a_r + hgb_r) x_{n-r}$$

where at least one of the coefficients $a_r$ or $b_r$ is nonzero. This finite difference equation, of order r, may be shown[10] to have the general solution

$$x_n = \sum_{j=1}^{r} c_j p_j^n \tag{17}$$

where the coefficients $c_j$ are determined by the initial values of $x_o$, $x_1$, $\ldots x_{r-1}$ and where $p_1$, $p_2, \ldots p_r$ are their roots of the characteristic equation

$$(1-hgb_o)p^r - (a_1 - hgb_1)\, p^{r-1} - (a_2 - hgb_2)p^{r-2}$$

$$\ldots\ldots -(a_r - hgb_r) = 0 \tag{18}$$

obtained by substituting the particular solution $x_n = p^n x_o$ into Eq. (16).

One of these roots, say $p_1$, will generate the principal solution to the difference equation. The other r-1 roots will generate parasitic solutions which arise because the order of the finite difference equation is (r-1) greater than that of the differential equation being approximated. Accordingly, if any one of the parasitic roots is greater in magnitude than unity, then the corresponding term $c_j p_j^n$ in Eq. (17) increases without bound as n increases, thereby vitiating the desired solution.[10] This situation, called numerical where e' is the node-to-datum voltage vector and I and E are the current source and voltage source vectors (corresponding to the IDATA and EDATA tables.)

Since the only current sources considered in TAP are those due to transistor action, one may write the matrix relation

$$A^t I = \bar{\alpha} I_h \qquad (6)$$

to define the equivalent nodal current sources depicted in Fig. 2. In Eq. (6), $I_h$ represents the vector of $I_h$ values for all transistors while $\bar{\alpha}$ represents a matrix whose only nonzero elements are $-1$, $\alpha$, and $(1-\alpha)$ placed so as to assign current sources $-I_h$, $\alpha I_h$ and $(1-\alpha) I_h$ to the appropriate nodes of each transistor model.

The voltages $V_{eb}'$ for each transistor may be computed from the node voltage relation $e_e' - e_b''$. In matrix form, this may be written as

$$V_{eb}' = He' \qquad (7)$$

where H is a matrix whose only nonzero elements are +1 and -1 placed so as to perform the appropriate linear combination of the elements of e'.

Combining Eqs. (5), (6) and (7), we may write

$$V_{eb}' = \left[H(A^tYA)^{-1}\bar{\alpha}\right] I_h - H(A^tYA)^{-1} A^tYE \qquad (8)$$

which expresses the vector $V_{eb}'$ directly as a function of $I_h$. The trick now is to choose $I_h$ so that both Eq. (8) and Eq. (5) are satisfied.

The nature of this problem is shown graphically in Fig. 3 where the diode curve represents Eq. (4) and where the load line represents the almost linear relation between $V_{eb}'$ and $I_h$ in Eq. (8) -- ignoring the slight dependency of $\bar{\alpha}$ on $I_h$. The two situations of major interest shown in Fig. 3 correspond to load lines for "on" and "off" transistors.

The iterative method of solving this problem may be explained as follows: For each transistor, an initial estimate $I_h^{(o)}$ is made. The diode curve is then approximated by a tangent line at the point $(I_h^{(o)}, V_{eb}^{(o)\prime})$ and its intersection with the load line determined. This results in a new estimate $I_h^{(o)}$ as shown in Fig. 4. The process is then repeated until convergence is obtained.

The tangent-line approximation to the diode curve is given by the point-slope formula:

$$V_{eb}' = \left[\frac{1}{\Lambda(I_h^{(o)} + I_{es})}\right] I_h + V_{eb}^{(o)\prime} - \left[\frac{1}{\Lambda(I_h^{(o)}+I_{es})}\right] I_h^{(o)} \qquad (9)$$

This relation may be considered as a matrix equation describing the tangent lines for all transistors if the quantity $\left[1/\Lambda(I_h^{(o)} + I_{es})\right]$ is regarded as a diagonal matrix.

Now for transistors operating in the "off" region, close to the asymptotic limit of $-I_{es}$, the computation of $I_h + I_{es}$ may involve serious round-off error. As a result, the larger $1/\Lambda\ (I_h + I_{es})$ becomes, the more inaccurate it is. This error can be avoided, however, by making the change of variable,

$$J = I_h + I_{es} \qquad (10)$$

and not calculating $I_h$ explicitly.

Eliminating $V_{eb}'$ from Eqs. (8) and (9) and introducing the variable J, we obtain the following equation for the iteration procedure:

$$\left[H(A^tYA)^{-1}\bar{\alpha}^{(n-1)} - \frac{1}{\Lambda J^{(n-1)}}\right] J^{(n)} = V_{eb}^{(n)\prime} - \left[\frac{1}{\Lambda J^{(n-1)}}\right] + \left[H(A^tYA)^{-1}\bar{\alpha}^{(n-1)}\right] I_{es} + H(A^tYA)^{-1}A^tYE \qquad (11)$$

where $\bar{\alpha}^{(n-1)} = \bar{\alpha}(I_h{}^{(n-1)}$ is assumed to be constant. In this expression, whatever accuracy is inherent in $J^{(n-1)}$ is retained in $1/\Lambda J^{(n-1)}$ .

One difficulty which arises in applying Eq. (11) is that an "out-of-bounds" intersection of the tangent-line and load line for an "off" transistor may occur, as depicted in Fig. 5. This problem is handled by testing each element of the $J^{(n)}$ vector for proper boundedness and replacing the out-of-bounds elements by the arbitrary value of $I_{es} \times 10^{-6}$.

A second difficulty is that Eq. (11) yields inaccurate J values for the "off" transistors. These values may be improved by the method of successive substitutions depicted in Fig. 6. After each iteration, the elements of the J vector, corresponding to each "off" transistor, are tested for boundedness, replaced by $I_{es} \times 10^{-6}$ if necessary, and then substituted into the equation

$$V_{eb}' = \left[H(A^tYA)^{-1}\bar{\alpha}\right] J - \left[H(A^tYA)^{-1}\bar{\alpha}\right] I_{es} - H(A^tYA)^{-1} A^tYE \qquad (12)$$

The J values are then recomputed using the expression

$$J = I_{es} (e^{\Lambda V_{eb}'}) \qquad (13)$$

This process of successive substitutions converges if the slope of the load line is greater than that of the diode curve at their intersection point and if the starting point of the process is close enough to this intersection; otherwise, the process diverges. In order to guard against incipient divergence, the values of $\Lambda V_{eb}'$ are monitored and if a value in excess of 4.0 is detected, the process of successive substitutions is terminated. Otherwise, three iterations are made.

The entire procedure of successive approximations is repeated until the fractional change in the length of the J vector is $10^{-7}$, with a maximum of thirty iterations being allowed. Although no attempt has been made to establish theoretically the conditions under which convergence will be assured, it is believed that only those circuits which involve large positive feedback would be likely to cause trouble. In such cases, it may be necessary to resort to the transient analysis procedure, using whatever initial conditions are obtained from the d-c analysis, and allow the integration to proceed with no input pulse until a steady state is reached.

In the normal case, after convergence has been attained, all the voltages and currents of the network are computed, completing the d-c analysis. Not all of these data are required as initial conditions for the transient analysis, however, and so only the desired voltages and currents are selected.

## Transient Analysis

After the initial conditions have been obtained from the d-c analysis, they are verified by running the numerical integration for a short time without any input pulse. When a satisfactory steady state has been reached, the input pulse is initiated. The transient computation may then be continued as long as desired. At suitable intervals during this computation, the entire set of voltages and currents in the network is printed out. It would be preferrable to display certain portions of this information graphically, either on a printer or cathode ray tube, but this feature was not included in TAP.

The solution of the transient problem, as mentioned above, is based on integrating the linear differential equations of the network separately from the nonlinear differential equations describing the transistors. To explain the practical importance of this modus operandi, a brief outline of the basic theory of the numerical integration is needed first.

The numerical integration of the single differential equation

$$\frac{dx}{dt} = f(x, t) \qquad (14)$$

may be effected by means of the generalized expression

$$x_n = \sum_{j=1}^{N} a_j x_{n-j} + h \sum_{j=0}^{M-1} b_j (dx/dt)_{n-j} \qquad (15)$$

where $x_n = x(t_o + nh)$ and $h = \Delta t$ is the integration interval. If $b_o = 0$, the integration formula is called a "predictor"; if $b_o \neq 0$, it is called a "corrector." Usually, a combined predictor-corrector scheme is employed with the coefficient $a_j$ and $b_j$ selected to give the accuracy desired.

In the special case of a constant coefficient, linear differential equation, where $f(x, t) = gx$, Eq. (15) becomes

$$(1 - hgb_o) x_n = (a_1 + hgb_1) x_{n-1} \qquad (16)$$

$$+ (a_2 + hgb_2) x_{n-2} + \ldots + (a_r + hgb_r) x_{n-r}$$

where at least one of the coefficients $a_r$ or $b_r$ is nonzero. This finite difference equation, of order r, may be shown[10] to have the general solution

$$x_n = \sum_{j=1}^{r} c_j p_j^n \qquad (17)$$

where the coefficients $c_j$ are determined by the initial values of $x_o$, $x_1$, ... $x_{r-1}$ and where $p_1$, $p_2$, ... $p_r$ are their roots of the characteristic equation

$$(1-hgb_o)p^r - (a_1 - hgb_1) p^{r-1} - (a_2 - hgb_2) p^{r-2}$$

$$\ldots . -(a_r - hgb_r) = 0 \qquad (18)$$

obtained by substituting the particular solution $x_n = p^n x_o$ into Eq. (16).

One of these roots, say $p_1$, will generate the principal solution to the difference equation. The other r-1 roots will generate parasitic solutions which arise because the order of the finite difference equation is (r-1) greater than that of the differential equation being approximated. Accordingly, if any one of the parasitic roots is greater in magnitude than unity, then the corresponding term $c_j p_j^n$ in Eq. (17) increases without bound as n increases, thereby vitiating the desired solution.[10] This situation, called numerical instability, arises if the integration interval is made too large.

Similar considerations apply to the numerical integration of a system of differential equations. For example, the system of equations describing a linear, constant parameter RLC network (see Appendix I) are of the form

$$P \dot{X} + Q X = F(t) \qquad (19)$$

where P and Q are matrices, $\dot{X} = dX/dt$, X is a vector of voltages and currents, and F(t) is a vector of voltage and/or current sources, some of which may be time-dependent. By inverting the matrix P, we may write,

$$\dot{X} = -P^{-1} Q X + P^{-1}F(t) = S X + G(t) \qquad (20)$$

The vector counterpart of Eq. (15) for numerical integration of Eq. (20), in predictor-corrector form, is:

$$\text{predictor:} \quad X_n = \sum_{j=1}^{N} a_j \overline{X}_{n-j} + h \sum_{j=1}^{M} b_j \dot{X}_{n-j} \qquad (21)$$

$$\text{corrector:} \quad X_n = \sum_{j=1}^{T} c_j X_{n-j} + h \sum_{j=0}^{W-1} d_j \dot{X}_{n-j} \qquad (22)$$

Here, $\overline{X}_n$, the corrected solution vector, is obtained by using the predicted derivative vector, $\dot{X}_n = SX_n + G(t_n)$.

To avoid numerical instability in the use of Eqs. (21) and (22), the integration interval h must be made less in magnitude than the reciprocal of the largest eigenvalue $\lambda$max of the matrix S in Eq. (20). This eigenvalue corresponds to the largest natural frequency, and its reciprocal to the smallest natural time constant $\tau_{min}$ of the network. Ironically, this eigenvalue, through its exponential function, $e^{-\lambda_{max} t}$, contributes least to the (analytical) solution of Eq. (20) and yet it forces the numerical integration to proceed at a rate determined by the condition,

$$h < 0.25\, \tau_{min} \qquad (23)$$

The permissible maximum value of h is in this range but depends somewhat on the actual choice of coefficients $a_j$, $b_j$, $c_j$ and $d_j$ in Eqs. (21) and (22).[11,12]

Eqs. (21) and (22) may also be used to approximate the solution of the differential equations describing nonlinear or time-varying networks, since the use of numerical integration implies that the network is linear and time-invariant at every particular time step, $t_n$, even though its parameters change from one time step to the next. Accordingly, the matrices P and Q will, in general, need to be recomputed and Eq. (19) either solved for $X_n$ or converted to the form of Eq. (20 by inverting P at each time step. This will change the matrix S and its eigenvalues. Hence, the integration interval, h, may need to be adjusted from time to time in order to prevent instability, when $\tau_{min}$ decreases, or to permit faster integration when $\tau_{min}$ increases.

In the circuits handled by TAP, the smallest time constant is due primarily to the parameters in the transistor model. This time constant, at best, is about 1/10 that of the linear part of the network considered by itself. Therefore, by integrating the equations for the linear network separately from those pertaining to the transistors, a significant increase in speed of integration is obtained. This increase is due not only to the larger integration interval permitted but also to the fact that the P matrix for the linear system need be inverted but once.

The integration scheme used for the linear network in TAP is based on a modified Euler predictor - corrector formula:

$$\text{predictor:} \quad X_n = \overline{X}_{n-1} + h_L \dot{X}_{n-1} \qquad (24)$$

$$\text{corrector:} \quad \overline{X}_n = \overline{X}_{n-1} + \frac{h_L}{2} (\dot{X}_n + \dot{X}_{n-1}) \qquad (25)$$

with $\dot{X}_n = SX_n + G(t_n)$ and with $h_L$ held constant. The vector $G(t_n)$, which is updated at each time step, contains current source terms that describe the effect of each transistor. The values of these current sources are obtained from the solution of the nonlinear differential equations describing each transistor's behavior.

These equations, two for each transistor, are based on the equivalent circuit shown in Fig. 7 where $R_{bb}'$, $R_c'$, $R_{cc}'$ and the current sources $\alpha I_h$ and $I_h(1-\alpha)I_h$ are identical to their counterparts in Fig. 1. The collector-base capacitance $C_{tc}$ is assumed constant but the emitter-base capacitance $C_e$ is assumed to be the sum of the two nonlinear capacitance

$$C_{se} = \frac{\Lambda (I_h + I_{es})}{2\pi f_{\alpha co}} \qquad (26)$$

and

$$C_{te} = \frac{K}{(V_o - V_{eb}')^n} \qquad (27)$$

where $f_{\alpha co}$ is the common-base cutoff frequency, K is a proportionality constant computed by the program, $V_o$ is the contact potential, and n is a constant dependent on the grading of the junction. All the basic parameters of this model, except K, are specified on the input cards for each transistor and tabulated during the compilation process. It should be pointed out that $C_e$ is defined as a differential or small-signal, capacitance dQ/dV and not as a static capacitance Q/V. This definition is preferable from the standpoint of measuring $C_e$ but care must be exercised in using small-signal capacitances, as discussed in Appendix I.

The nonlinear differential equations for each transistor are integrated separately using a modified Adam's predictor-corrector formula:

$$\text{predictor:}\quad X_n = \overline{X}_{n-1} + (h_N/24)(55\dot{X}_{n-1} - 59\dot{X}_{n-2} + 37\dot{X}_{n-3} - 9\dot{X}_{n-4}) \qquad (28)$$

$$\text{corrector:}\quad \overline{X}_n = \overline{X}_{n-1} + (h_N/24)(9\dot{X}_n + 19\dot{X}_{n-1} - 5\dot{X}_{n-2} + \dot{X}_{n-3}) \qquad (29)$$

The integration interval $h_N$ is continuously monitored by comparing the difference between predicted and corrected values of the solution. When this difference increases beyond a certain limit, the integration interval is halved; when this difference decreases sufficiently, the interval is doubled. In this way, since each transistor is treated independently of the others, the integration proceeds at close to the maximum safe rate for each transistor instead of at the rate of the slowest.

The transistor equations require as input data the response voltages at each of the nodes of the linear network to which a transistor terminal is connected. These voltages are computed at each integration interval $h_L$ of the linear system of equations and they are supplied to the nonlinear equations as driving forces. These driving forces are assumed constant over the next integration interval $h_L$ while the nonlinear equations are integrated using the variable interval $h_N$. The integration process for each transistor is carried along, with $h_N$ being adjusted enroute, until a period exactly equal to $h_L$ has been covered. The terminal currents (emitter, base, and collector) computed at the end of this period are then supplied to the $G(t_n)$ vector for the linear system for its next integration step. In this way, the integration of the linear and nonlinear equations is carried out alternately with both systems being joined at each interval $h_L$.

## Results and Applications

As an indication of the adequacy of TAP in predicting the transient behavior of a transistor switching circuit, the results of the analysis of the circuit shown in Fig. 8 are displayed in Figs. 9-16 together with the observed responses.[1] These results were obtained in about 10 minutes of computing time on the IBM 704. The close agreement between the computed and observed results is really a testimony to the faithfulness of the transistor model since the analytical and computational techniques are in themselves quite dependable.

The potential value of a network analysis program as an experimental tool is indicated by Figs. 17-22 which show the different responses resulting from the variation of a single parameter in the network of Fig. 8. Admittedly, some of these variations can be explored with the actual hardware. But the variation of transistor parameters, such as $f_{\alpha co}$ and $V_o$, cannot be achieved on demand in any practical sense. Evidently then, a network analysis program such as TAP offers the design engineer a direct means of studying the behavior of circuits and/or devices in intimate detail either for the purpose of increasing his understanding or for helping him to optimize circuit performance.

Beyond the obvious electrical applications of a network analysis program, there are many possible applications to nonelectrical problems. The DYANA program previously mentioned is already taking advantage of this fact by solving mechanical as well as electrical problems. But this is barely scratching the surface of a vast and fertile field. Actually, a significant portion of theoretical physics is amenable to a network approach.[5, 13] Indeed, network models exist for so many different physical systems as to force the conclusion that a general purpose network analysis program is capable of converting a digital computer into a versatile and powerful analog machine.

## Conclusions and Remarks

An experimental program has been written which is capable of automatically formulating and solving both the d-c and transient analysis problems relating to transistor switching circuits of arbitrary configuration. The program yields computed results which are in reasonable agreement with observations, a fact which proves the adequacy of the transistor model as well as that of the program itself.

The principal features of this program are: (1) its ability to formulate the network problem automatically on the basis of simple input data specifying the network parameters and configuration; (2) its use of topological-matrix methods for handling part of the formulation and analysis; (3) its faster solution of the transient problem by separately integrating the linear and nonlinear sets of differential equations.

The main failing of the program is the difficulty of altering or replacing the transistor equivalent circuit. Another disadvantage is the printing, rather than plotting, of the computed responses. Both of these disadvantages have been eliminated from a more recent program for circuit analysis of nonlinear systems (PE CANS) developed for the IBM 7090 computer by Beaudette and Honkanen.[14] This program compiles the equations for a network including arbitrary nonlinear elements. Hence, it can handle equivalent circuits for a variety of nonlinear devices.

A. F. Malmberg, at Los Alamos Scientific Laboratory, has also written a network analysis program for the MANIAC II computer.[15] This program is based on the topological-matrix formulation described in Appendix I and uses a network model capable of describing saturating transistors.

It has been amply demonstrated, therefore, that it is quite feasible to program a digital computer to both formulate and solve the algebraic and/or differential equations of an arbitrary network -- including at least certain types of nonlinear device. It now remains to refine the techniques described here and to develop new ones so that the full potentialities of a general purpose network analysis program can be realized. Clearly, the practical utility of such a program will depend almost as much on its input/output facilities as on its speed. Accordingly, due attention must be given to such user-oriented features as input format (including original network specifications and modifications thereto) and output display.

The central difficulty, however, is still that of solving the transient problem. Much can be gained by refining the techniques of programming predictor-corrector formulas. But what is really needed is a genuine analytical breakthrough which will lead to an orders-of-magnitude increase in speed. Such a breakthrough, it would appear, cannot possibly come unless some way over, around, or through the minimum time-constant barrier can be found. This is a frontier which offers the greatest challenge and most promising rewards.

## Acknowledgements

The TAP program represents the combined efforts of several people whose contributions deserve to be recognized. The originator and director of the project was R. J. Domenico. The transient analysis program, developed primarily by Mrs. Nancy G. Brooks with the assistance of H. von Horn and E. J. Skiko, was based on the earlier work of G. L. Lasher and J. C. Morgan. The numerical integration program, used in the nonlinear transient analysis, was written by M. K. Haynes. The d-c analysis was developed by the present author and implemented by him with the assistance of H. S. Long and J. T. Urbain.

## Appendix I

## Formulation of the Transient Network Problem

The following topological-matrix formulation of the linear transient network problem leads to a simultaneous system of algebraic and first order differential equations similar to that previously described by Bashkow.[9] The present formulation, however, avoids the introduction of the extraneous capacitors and inductors which Bashkow's derivation requires. Moreover, it is in a form which is suitable for programming on a digital computer by an extension of the techniques described elsewhere in this paper. The terminology and notation to be used are essentially the same as in previous work of the author.[5,6]

Instead of employing either the mesh method or the node method of analysis, the present formulation of the transient problem makes use of a combination of these two methods. Actually, the tree method,[5,6] rather than the node method, is combined with the mesh method. This combination, which is also implicit in Bashkow's formulation, is made necessary by the requirement to establish first order differential equations rather than integro-differential equations to characterize the reactive elements of the network. A formal description of this combined method of analysis will be given first. The necessary extension to the transient problem then follows easily.

It is assumed that the network branches are first divided into two categories: admittances, designated by the subscript y, and impedances, designated by the subscript z. It is also assumed that there is no coupling between any admittance branch and any impedance branch although branches within the same category may be coupled. arbitrarily with one another. Ohm's law, instead of being written either in the admittance form $J = YV$ or in the impedance form $V = ZJ$, is now written in the mixed form

$$\begin{bmatrix} J_y \\ V_z \end{bmatrix} = \begin{bmatrix} Y_y & 0 \\ 0 & Z_z \end{bmatrix} \begin{bmatrix} V_y \\ J_z \end{bmatrix} \tag{30}$$

where $Y_y$ and $Z_z$ are the primitive admittance and primitive impedance matrices, and where J and V are the coil current and coil voltage vectors.[5,6] Using the relations J = I + i and V = E + e, where I and E are the current and voltage source vectors while i and e are the branch current and branch voltage (response) vectors, Eq. (30) may also be written as follows:

$$\begin{bmatrix} I_y + i_y \\ E_z + e_z \end{bmatrix} = \begin{bmatrix} Y_y & 0 \\ 0 & Z_z \end{bmatrix} \begin{bmatrix} E_y + e_y \\ I_z + i_z \end{bmatrix} \tag{31}$$

Disregarding the question of ordering the branches, it is now assumed that the admittance branches are classified as either tree-branches or links, using the procedure outlined in Appendix II. Then, with the resulting admittance tree as a starting point, the impedance branches are similarly classified. The network tree obtained in this fashion will, of course, contain both admittance and impedance branches. However, since all the admittances will have been subjected first to the tree-link sorting procedure, all the basic meshes defined by admittance links will necessarily include only admittance tree-branches.

On the other hand, the basic meshes defined by the impedance links may include both admittance and impedance tree-branches. As a consequence, the $C_T$ matrix contains one null submatrix. For if the matrices $B_T$ and $A_T$ are partitioned into submatrices thus.

$$B_T = \begin{bmatrix} B_{Ty} \\ B_{Ty} \end{bmatrix} \tag{32}$$

and

$$A_L = \begin{bmatrix} A_{Ly} \\ A_{Ly} \end{bmatrix} \tag{33}$$

it follows from Eq. (1) that

$$C_T = \begin{bmatrix} C_{Tyy} & C_{Tyz} \\ 0 & C_{Tzz} \end{bmatrix} = \begin{bmatrix} B_{Ty} A^t_{Ly} & -B_{Ty} A^t_{Lz} \\ 0 & -B_{Tz} A^t_{Lz} \end{bmatrix} \tag{34}$$

since $C_{Tzy} = -B_{Tz} A^t_{Ly} = 0$, as explained above.

In accord with the tree method of analysis, the branch voltages e for the entire network are expressed as a linear combination of the tree branch voltages $e_T$ using the relation $e = De_T$, where D is the basic cut-set matrix for the entire network.[5,6] At the same time, in keeping with the mesh method, the branch currents i are expressed as a linear combination of the link currents $i_L$ (which, by convention, are identical with the mesh currents,) using the relation $i = Ci_L$.

These relations, together with the four-way classification of branches described above, lead to the expressions

$$\begin{bmatrix} e_{Ty} \\ e_{Tz} \\ e_{Ly} \\ e_{Lz} \end{bmatrix} = \begin{bmatrix} U_{Ty} & 0 \\ 0 & U_{Tz} \\ -C^t_{Tyy} & 0 \\ -C^t_{Tyz} & -C^t_{Tzz} \end{bmatrix} \begin{bmatrix} e_{Ty} \\ e_{Tz} \end{bmatrix} \tag{35}$$

and

$$\begin{bmatrix} i_{Ty} \\ i_{Tz} \\ i_{Ly} \\ i_{Lz} \end{bmatrix} = \begin{bmatrix} C_{Tyy} & C_{Tyz} \\ 0 & C_{Tzz} \\ U_{Ly} & 0 \\ 0 & U_{Lz} \end{bmatrix} \begin{bmatrix} i_{Ly} \\ i_{Lz} \end{bmatrix} \tag{36}$$

where use has been made of the fact that for the basic cut-set matrix, $D_T = U_T$ (a unit matrix) and $D_L = -C^t_T$.[5,6]

Next, Eq. (31) is rearranged to give

$$\begin{bmatrix} (I_y - Y_y E_y) \\ (E_z - Z_z I_z) \end{bmatrix} = \begin{bmatrix} Y_y & 0 \\ 0 & Z_z \end{bmatrix} \begin{bmatrix} e_y \\ i_z \end{bmatrix} - \begin{bmatrix} i_y \\ e_z \end{bmatrix} \tag{37}$$

where $e_y$ includes both subvectors $e_{Ty}$ and $e_{Ly}$, $i_z$ includes both $i_{Tz}$ and $i_{Lz}$, etc. It now becomes necessary to introduce the admittance cut-set matrix,

$$D_y = \begin{bmatrix} U_{Ty} \\ -C^t_{Tyy} \end{bmatrix} \tag{38}$$

and the impedance mesh matrix,

$$C_z = \begin{bmatrix} C_{Tzz} \\ U_{Lz} \end{bmatrix} \tag{39}$$

to extract from Eq. (35) the expression $e_y = D_y e_{Ty}$, or

$$e_y = \begin{bmatrix} e_{Ty} \\ e_{Ly} \end{bmatrix} = \begin{bmatrix} U_{Ty} \\ -C^t_{Tyy} \end{bmatrix} e_{Ty} \tag{40}$$

The principal features of this program are: (1) its ability to formulate the network problem automatically on the basis of simple input data specifying the network parameters and configuration; (2) its use of topological-matrix methods for handling part of the formulation and analysis; (3) its faster solution of the transient problem by separately integrating the linear and nonlinear sets of differential equations.

The main failing of the program is the difficulty of altering or replacing the transistor equivalent circuit. Another disadvantage is the printing, rather than plotting, of the computed responses. Both of these disadvantages have been eliminated from a more recent program for circuit analysis of nonlinear systems (PE CANS) developed for the IBM 7090 computer by Beaudette and Honkanen.[14] This program compiles the equations for a network including arbitrary nonlinear elements. Hence, it can handle equivalent circuits for a variety of nonlinear devices.

A. F. Malmberg, at Los Alamos Scientific Laboratory, has also written a network analysis program for the MANIAC II computer.[15] This program is based on the topological-matrix formulation described in Appendix I and uses a network model capable of describing saturating transistors.

It has been amply demonstrated, therefore, that it is quite feasible to program a digital computer to both formulate and solve the algebraic and/or differential equations of an arbitrary network -- including at least certain types of nonlinear device. It now remains to refine the techniques described here and to develop new ones so that the full potentialities of a general purpose network analysis program can be realized. Clearly, the practical utility of such a program will depend almost as much on its input/output facilities as on its speed. Accordingly, due attention must be given to such user-oriented features as input format (including original network specifications and modifications thereto) and output display.

The central difficulty, however, is still that of solving the transient problem. Much can be gained by refining the techniques of programming predictor-corrector formulas. But what is really needed is a genuine analytical breakthrough which will lead to an orders-of-magnitude increase in speed. Such a breakthrough, it would appear, cannot possibly come unless some way over, around, or through the minimum time-constant barrier can be found. This is a frontier which offers the greatest challenge and most promising rewards.

## Acknowledgements

The TAP program represents the combined efforts of several people whose contributions deserve to be recognized. The originator and director of the project was R. J. Domenico. The transient analysis program, developed primarily by Mrs. Nancy G. Brooks with the assistance of H. von Horn and E. J. Skiko, was based on the earlier work of G. L. Lasher and J. C. Morgan. The numerical integration program, used in the nonlinear transient analysis, was written by M. K. Haynes. The d-c analysis was developed by the present author and implemented by him with the assistance of H. S. Long and J. T. Urbain.

## Appendix I

### Formulation of the Transient Network Problem

The following topological-matrix formulation of the linear transient network problem leads to a simultaneous system of algebraic and first order differential equations similar to that previously described by Bashkow.[9] The present formulation, however, avoids the introduction of the extraneous capacitors and inductors which Bashkow's derivation requires. Moreover, it is in a form which is suitable for programming on a digital computer by an extension of the techniques described elsewhere in this paper. The terminology and notation to be used are essentially the same as in previous work of the author.[5,6]

Instead of employing either the mesh method or the node method of analysis, the present formulation of the transient problem makes use of a combination of these two methods. Actually, the tree method,[5,6] rather than the node method, is combined with the mesh method. This combination, which is also implicit in Bashkow's formulation, is made necessary by the requirement to establish first order differential equations rather than integro-differential equations to characterize the reactive elements of the network. A formal description of this combined method of analysis will be given first. The necessary extension to the transient problem then follows easily.

It is assumed that the network branches are first divided into two categories: admittances, designated by the subscript y, and impedances, designated by the subscript z. It is also assumed that there is no coupling between any admittance branch and any impedance branch although branches within the same category may be coupled. arbitrarily with one another. Ohm's law, instead of being written either in the admittance form $J = YV$ or in the impedance form $V = ZJ$, is now written in the mixed form

$$\begin{bmatrix} J_y \\ V_z \end{bmatrix} = \begin{bmatrix} Y_y & 0 \\ 0 & Z_z \end{bmatrix} \begin{bmatrix} V_y \\ J_z \end{bmatrix} \tag{30}$$

where $Y_y$ and $Z_z$ are the primitive admittance and primitive impedance matrices, and where J and V are the coil current and coil voltage vectors.[5, 6] Using the relations $J = I + i$ and $V = E + e$, where I and E are the current and voltage source vectors while i and e are the branch current and branch voltage (response) vectors, Eq. (30) may also be written as follows:

$$\begin{bmatrix} I_y + i_y \\ E_z + e_z \end{bmatrix} = \begin{bmatrix} Y_y & 0 \\ 0 & Z_z \end{bmatrix} \begin{bmatrix} E_y + e_y \\ I_z + i_z \end{bmatrix} \tag{31}$$

Disregarding the question of ordering the branches, it is now assumed that the admittance branches are classified as either tree-branches or links, using the procedure outlined in Appendix II. Then, with the resulting admittance tree as a starting point, the impedance branches are similarly classified. The network tree obtained in this fashion will, of course, contain both admittance and impedance branches. However, since all the admittances will have been subjected first to the tree-link sorting procedure, all the basic meshes defined by admittance links will necessarily include only admittance tree-branches.

On the other hand, the basic meshes defined by the impedance links may include both admittance and impedance tree-branches. As a consequence, the $C_T$ matrix contains one null submatrix. For if the matrices $B_T$ and $A_T$ are partitioned into submatrices thus.

$$B_T = \begin{bmatrix} B_{Ty} \\ B_{Ty} \end{bmatrix} \tag{32}$$

and

$$A_L = \begin{bmatrix} A_{Ly} \\ A_{Ly} \end{bmatrix} \tag{33}$$

it follows from Eq. (1) that

$$C_T = \begin{bmatrix} C_{Tyy} & C_{Tyz} \\ 0 & C_{Tzz} \end{bmatrix} = \begin{bmatrix} B_{Ty} A^t_{Ly} & -B_{Ty} A^t_{Lz} \\ 0 & -B_{Tz} A^t_{Lz} \end{bmatrix} \tag{34}$$

since $C_{Tzy} = -B_{Tz} A^t_{Ly} = 0$, as explained above.

In accord with the tree method of analysis, the branch voltages e for the entire network are expressed as a linear combination of the tree branch voltages $e_T$ using the relation $e = De_T$, where D is the basic cut-set matrix for the entire network.[5, 6] At the same time, in keeping with the mesh method, the branch currents i are expressed as a linear combination of the link currents $i_L$ (which, by convention, are identical with the mesh currents,) using the relation $i = Ci_L$.

These relations, together with the four-way classification of branches described above, lead to the expressions

$$\begin{bmatrix} e_{Ty} \\ e_{Tz} \\ e_{Ly} \\ e_{Lz} \end{bmatrix} = \begin{bmatrix} U_{Ty} & 0 \\ 0 & U_{Tz} \\ -C^t_{Tyy} & 0 \\ -C^t_{Tyz} & -C^t_{Tzz} \end{bmatrix} \begin{bmatrix} e_{Ty} \\ e_{Tz} \end{bmatrix} \tag{35}$$

and

$$\begin{bmatrix} i_{Ty} \\ i_{Tz} \\ i_{Ly} \\ i_{Lz} \end{bmatrix} = \begin{bmatrix} C_{Tyy} & C_{Tyz} \\ 0 & C_{Tzz} \\ U_{Ly} & 0 \\ 0 & U_{Lz} \end{bmatrix} \begin{bmatrix} i_{Ly} \\ i_{Lz} \end{bmatrix} \tag{36}$$

where use has been made of the fact that for the basic cut-set matrix, $D_T = U_T$ (a unit matrix) and $D_L = -C^t_T$.[5, 6]

Next, Eq. (31) is rearranged to give

$$\begin{bmatrix} (I_y - Y_y E_y) \\ (E_z - Z_z I_z) \end{bmatrix} = \begin{bmatrix} Y_y & 0 \\ 0 & Z_z \end{bmatrix} \begin{bmatrix} e_y \\ i_z \end{bmatrix} - \begin{bmatrix} i_y \\ e_z \end{bmatrix} \tag{37}$$

where $e_y$ includes both subvectors $e_{Ty}$ and $e_{Ly}$, $i_z$ includes both $i_{Tz}$ and $i_{Lz}$, etc. It now becomes necessary to introduce the admittance cut-set matrix,

$$D_y = \begin{bmatrix} U_{Ty} \\ -C^t_{Tyy} \end{bmatrix} \tag{38}$$

and the impedance mesh matrix,

$$C_z = \begin{bmatrix} C_{Tzz} \\ U_{Lz} \end{bmatrix} \tag{39}$$

to extract from Eq. (35) the expression $e_y = D_y e_{Ty}$, or

$$e_y = \begin{bmatrix} e_{Ty} \\ e_{Ly} \end{bmatrix} = \begin{bmatrix} U_{Ty} \\ -C^t_{Tyy} \end{bmatrix} e_{Ty} \tag{40}$$

and from Eq. (36) the expression $i_z = C_z i_{Lz}$, or

$$i_z = \begin{bmatrix} i_{Tz} \\ i_{Lz} \end{bmatrix} = \begin{bmatrix} C_{Tzz} \\ U_{Lz} \end{bmatrix} i_{Lz} \tag{41}$$

Then, considering only the first column of D in Eq. (35) and using the fundamental relation $D^t\, i = 0$, it is easily shown that

$$D_y^t\, i_y = \begin{bmatrix} U_{Ty} & -C_{Tyy} \end{bmatrix} \begin{bmatrix} i_{Ty} \\ i_{Ly} \end{bmatrix} = -C_{Tyz}\, i_{Lz} \tag{42}$$

Similarly, from the second column of C in Eq. (36) and the equation $C^t\, e = 0$, it follows that

$$C_z^t e_z = \begin{bmatrix} C_{Tzz}^t & U_{Lz} \end{bmatrix} \begin{bmatrix} e_{Tz} \\ e_{Lz} \end{bmatrix} = -C_{Tyz}^t e_{Ty} \tag{43}$$

Finally, premultiplication of the first row of Eq. (37) by $D_y^t$ and of the second row by $C_z^t$, followed by substitution of $D_y e_{Ty}$ in place of $e_y$ and of $C_z i_{Lz}$ in place of $i_z$, yields the result,

$$\begin{bmatrix} D_y^t(I_y - Y_y E_y) \\ C_z^t(E_z - Z_z I_z) \end{bmatrix} = \begin{bmatrix} D_y^t Y_y D_y & 0 \\ 0 & C_z^t Z_z C_z \end{bmatrix} \begin{bmatrix} e_{Ty} \\ i_{Lz} \end{bmatrix} - \begin{bmatrix} D_y^t i_y \\ C_z^t e_z \end{bmatrix} \tag{44}$$

which, together with Eqs. (42) and (43) may be condensed to the desired expression,

$$\begin{bmatrix} D_y^t(I_y - Y_y E_y) \\ C_z^t(E_z - Z_z I_z) \end{bmatrix} = \begin{bmatrix} D_y^t Y_y D_y & -C_{Tyz} \\ C_{Tyz}^t & C_z^t Z_z C_z \end{bmatrix} \begin{bmatrix} e_{Ty} \\ i_{Lz} \end{bmatrix} \tag{45}$$

Thus, Eq. (45) amounts to a tree analysis of the admittance branches alone followed by a mesh analysis of the impedance branches alone, the resulting two sets of equations being coupled together by the submatrix $C_{Tyz}$. But this submatrix denotes those admittance tree branches which belong to basic meshes defined by impedance links. Hence it follows that the corresponding admittance-tree-branch voltages and impedance-link currents will exhibit a reciprocal interaction.

Now in order to guarantee that only first order differential (as well as algebraic) equations will result from the application of Eq. (45) to the transient problem, all capacitors must be classified as admittances and all inductors as impedances. Resistors, however, may be put into either category. The matrices $Y_y$ and $Z_z$, then, contain both algebraic and differential operators and may be written thus:

$$Y_y = G_y + \frac{d}{dt} K_y \tag{46}$$

and

$$Z_z = R_z + \frac{d}{dt} L_z \tag{47}$$

where the symbols G, K, R and L denote conductance, capacitance, resistance and inductance matrices. (K is used for capacitance because C has already been used to designate a topological matrix.) Hence, for the most general case of time-varying capacitances and inductances, Eq. (30) becomes

$$\begin{bmatrix} J_y \\ V_z \end{bmatrix} = \begin{bmatrix} K_y & 0 \\ 0 & L_z \end{bmatrix} \begin{bmatrix} \dot{V}_y \\ \dot{J}_z \end{bmatrix} + \begin{bmatrix} G_y + \dot{K}_y & 0 \\ 0 & R_z + \dot{L}_z \end{bmatrix} \begin{bmatrix} V_y \\ J_z \end{bmatrix}. \tag{48}$$

All admittances having zero capacitance and all impedances having zero inductance will, of course, give rise to zero entries in the $K_y$ and $L_z$ matrices; they will also generate algebraic rather than differential equations. Therefore, to gather these zeros into null matrices and group the algebraic equations together, it is convenient to classify the network branches as follows:

(1) admittances with nonzero capacitance,
(2) admittances with conductance only,
(3) impedances with nonzero inductance,
(4) impedances with resistance only.

These four classes will be denoted by the subscripts k, g, l, and r respectively and it will be assumed that these classes are always in the order shown above. Accordingly, Eq. (48) becomes

$$\begin{bmatrix} J_{yk} \\ J_{yg} \\ V_{zl} \\ V_{zr} \end{bmatrix} = \begin{bmatrix} K_{yk} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & L_{zl} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{V}_{yk} \\ \dot{V}_{yg} \\ \dot{J}_{kl} \\ \dot{J}_{zr} \end{bmatrix} + \begin{bmatrix} G_{yk} + K_{yk} & 0 & 0 & 0 \\ 0 & G_{yg} & 0 & 0 \\ 0 & 0 & R_{zl} + \dot{L}_{zl} & 0 \\ 0 & 0 & 0 & R_{zr} \end{bmatrix} \begin{bmatrix} V_{yk} \\ V_{yg} \\ J_{zl} \\ J_{zr} \end{bmatrix} \tag{49}$$

where the variables with the subscripts yg and zr are involved in purely algebraic equations. It should be noted that all capacitative branches may have nonzero conductance and that all inductive branches may have nonzero resistance. Hence, these particular conductive and/or resistive elements need not be treated as separate branches. This does not preclude their being treated as separate branches, however, if there is some reason for doing so.

If the network branches, ordered by class as shown above (but arbitrarily ordered within each class), are subjected to a tree-link sort, then a sequence of expressions similar to Eqs. (32) to (45), but with twice as many branch categories will result. In particular, it follows that

$$B_T = \begin{bmatrix} B_{Tk} \\ B_{Tg} \\ B_{Tl} \\ B_{Tr} \end{bmatrix} \qquad (50)$$

$$A_L = \begin{bmatrix} A_{Lk} \\ A_{Lg} \\ A_{Lr} \\ A_{Lr} \end{bmatrix} \qquad (51)$$

and

$$C_T = -B_T A_L^t = \begin{bmatrix} C_{Tkk} & C_{Tkg} & C_{Tkl} & C_{Tkr} \\ 0 & C_{Tgg} & C_{Tgl} & C_{Tgr} \\ 0 & 0 & C_{Tll} & C_{Tlr} \\ 0 & 0 & 0 & C_{Trr} \end{bmatrix} \qquad (52)$$

where the submatrices of $C_T$ are defined in the obvious way. Eqs. (38) and (39) now become

$$D_y = \begin{bmatrix} U_{Tk} & 0 \\ 0 & U_{Tg} \\ -C^t_{Tkk} & 0 \\ -C^t_{Tkg} & -C^t_{Tgg} \end{bmatrix} \qquad (53)$$

and

$$C_z = \begin{bmatrix} C_{Tll} & C_{Tlr} \\ 0 & C_{Trr} \\ U_{Ll} & 0 \\ 0 & U_{Lr} \end{bmatrix} \qquad (54)$$

while the submatrix $C_{Tyz}$ of Eqs. (42), (43) and (45) becomes

$$C_{Tyz} = \begin{bmatrix} C_{Tkl} & C_{Tkr} \\ C_{Tgl} & C_{Tgr} \end{bmatrix} \qquad (55)$$

Next, the vectors e and i are written in partitioned form (as row vectors) as:

$$e = (e_{Tk}\ e_{Tg}\ e_{Tl}\ e_{Tr}\ e_{Lk}\ e_{Lg}\ e_{Ll}\ e_{Lr}) \qquad (56)$$

$$i = (i_{Tk}\ i_{Tg}\ i_{Tl}\ i_{Tr}\ i_{Lk}\ i_{Lg}\ i_{Ll}\ i_{Lr}) \qquad (57)$$

while the $Y_y$ and $Z_z$ matrices are partitioned thus:

$$Y_y = \begin{bmatrix} G_{Tk} + \frac{d}{dt}K_{Tk} & 0 & 0 & 0 \\ 0 & G_{Tg} & 0 & 0 \\ 0 & 0 & G_{Lk} + \frac{d}{dt}K_{Lk} & 0 \\ 0 & 0 & 0 & G_{Lg} \end{bmatrix} \qquad (58)$$

and

$$Z_z = \begin{bmatrix} R_{Tl} + \frac{d}{dt}L_{Tl} & 0 & \frac{d}{dt}L_{TLl} & 0 \\ 0 & R_{Tr} & 0 & 0 \\ \frac{d}{dt}L_{LTl} & 0 & R_{Ll} + \frac{d}{dt}L_{Ll} & 0 \\ 0 & 0 & 0 & R_{Lr} \end{bmatrix} \qquad (59)$$

where the matrices $L_{TLl}$ and $L_{LTl}$ allow for inductive coupling (if any) between tree-branches and links.

Finally, with all these relations substituted into Eq. (45), it can be shown that

$$\begin{bmatrix} I'_{Tk} \\ I'_{Tg} \\ E'_{Ll} \\ E'_{Lr} \end{bmatrix} = \begin{bmatrix} K'_{Tkk} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & L'_{Lll} & L'_{Llr} \\ 0 & 0 & L'_{Lrl} & L'_{Lrr} \end{bmatrix} \begin{bmatrix} \dot{e}_{Tk} \\ \dot{e}_{Tg} \\ \dot{i}_{Ll} \\ \dot{i}_{Lr} \end{bmatrix} + \begin{bmatrix} G'_{kk}+\dot{K}'_{kk} & G'_{kg} & -C_{Tkl} & -C_{Tkr} \\ G'_{gk} & G'_{gg} & -C_{Tgl} & -C_{Tgr} \\ C^t_{Tkl} & C^t_{Tgl} & R'_{ll}+\dot{L}'_{ll} & R'_{lr}+\dot{L}'_{lr} \\ C^t_{Tkr} & C^t_{Tgr} & R'_{rl}+\dot{L}'_{rl} & R'_{rr}+\dot{L}'_{rr} \end{bmatrix} \begin{bmatrix} e_{Tk} \\ e_{Tg} \\ i_{Ll} \\ i_{Lr} \end{bmatrix} \quad (60)$$

where the following relations apply:

$$\begin{bmatrix} I'_{Tk} \\ I'_{Tg} \end{bmatrix} = \begin{bmatrix} U_{Tk} & 0 & -C_{Tkk} & -C_{Tkg} \\ 0 & U_{Tg} & 0 & -C_{Tgg} \end{bmatrix} \left\{ \begin{bmatrix} I_{Tk} \\ I_{Tg} \\ I_{Lk} \\ I_{Lg} \end{bmatrix} - \begin{bmatrix} (G_{Tk}+\dot{K}_{Tk})E_{Tk} \\ G_{Tg}\,E_{Tg} \\ (G_{Lk}+\dot{K}_{Lk})E_{Lk} \\ G_{Lg}\,E_{Lg} \end{bmatrix} - \begin{bmatrix} K_{Tk}\dot{E}_{Tk} \\ 0 \\ K_{Lk}\dot{E}_{Lk} \\ 0 \end{bmatrix} \right\} \quad (61)$$

$$\begin{bmatrix} E'_{Ll} \\ E'_{Lr} \end{bmatrix} = \begin{bmatrix} C^t_{Tll} & 0 & U_{Ll} & 0 \\ C^t_{Tlr} & C^t_{Trr} & 0 & U_{Lr} \end{bmatrix} \left\{ \begin{bmatrix} E_{Tl} \\ E_{Tr} \\ E_{Ll} \\ E_{Lr} \end{bmatrix} - \begin{bmatrix} (R_{Tl}+\dot{L}_{Tl})I_{Tl} \\ R_{Tr}I_{Tr} \\ (R_{Ll}+\dot{L}_{Ll})I_{Ll} \\ R_{Lr}I_{Lr} \end{bmatrix} - \begin{bmatrix} L_{Tl}\dot{I}_{Tl}+L_{TLl}\dot{I}_{Ll} \\ 0 \\ L_{Ll}\dot{I}_{Ll}+L_{LTl}\dot{I}_{Tl} \\ 0 \end{bmatrix} \right\} \quad (62)$$

$$K'_{Tkk} = K_{Tk} + C_{Tkk}\,K_{Lk}\,C^t_{Tkk} \quad (63)$$

$$L'_{Lll} = L_{Ll} + C^t_{Tll}\,L_{Tl}\,C_{Tll} + C^t_{Tll}\,L_{TLl} + L_{LTl}\,C_{Tll} \quad (64)$$

$$L'_{Llr} = C^t_{Tll}\,L_{Tl}\,C_{Tlr} + L_{LTl}\,C_{Tlr} \quad (65)$$

$$L'_{Lrl} = C^t_{Tlr}\,L_{Tl}\,C_{Tll} + C^t_{Tlr}\,L_{TLl} \quad (66)$$

$$L'_{Lrr} = C^t_{Tlr}\,L_{Tl}\,C_{Tlr} \quad (67)$$

$$G'_{kk} = G_{Tk}+C_{Tkk}\,G_{Lk}\,C^t_{Tkk}+C_{Tkg}\,G_{Lg}\,C^t_{Tkg} \quad (68)$$

$$G'_{kg} = C_{Tkg}\,G_{Lg}\,C^t_{Tgg} \quad (69)$$

$$G'_{gk} = C_{Tgg}\,G_{Lg}\,C^t_{Tkg} \quad (70)$$

$$G'_{gg} = G_{Tg} + C_{Tgg}\,G_{Lg}\,C^t_{Tgg} \quad (71)$$

$$R'_{ll} = R_{Ll} + C^t_{Tll}\,R_{Tl}\,C_{Tll} \quad (72)$$

$$R'_{lr} = C_{Tll}\,R_{Tl}\,C_{Tlr} \quad (73)$$

$$R'_{rl} = C^t_{Tlr}\,R_{Tl}\,C_{Tll} \quad (74)$$

$$R'_{ll} = C'_{Tlr}\,R_{Tl}\,C_{Tlr} \quad (75)$$

From Eqs. (60), (65), (66) and (67), it is apparent that unless $C_{Tlr} = 0$, a set of differential equations will arise from those purely resistive network branches which are classed as impedances. In other words, if any of the link currents $i_{Lr}$ pass through inductive impedances belonging to the network tree (and this is what the matrix $C_{Tlr}$ specifies) these link currents must be determined as the solution of differential rather than algebraic equations. Therefore, unless some specific reason exists for not doing so, all purely resistive branches of a network should be treated as conductances. This will not only reduce the number of differential equations to be solved but will also result in considerable simplification of Eqs. (49) et seq.

If this is done it can be shown that the remaining algebraic equations in Eq. (60) have the solution

$$e_{Tg} = (G'_{gg})^{-1}\left[I'_{Tg} - G'_{gk}e_{Tk} + C_{Tgl}\,i_{Ll} + C_{Tgr}\,i_{Lr}\right] \quad (76)$$

Hence, if this expression is substituted in Eq. (60), a set of differential equations identical in form to Eq. (19) is obtained. The total number of differential equations derived in this way corresponds to the degree of complexity of the network.

However, if some of the purely resistive branches.are treated as impedances, then more than this number of differential equations may be obtained, as explained above. But this larger number cannot also be equal to the degree of complexity of the network. Accordingly, the true degree of complexity of any capacitive and/or inductive network must be equal to the sum of the number of capacitive tree-branches plus the number of inductive links identified by the foregoing tree-link sorting procedure.

One precaution in using Eq. (60) should be pointed out. This equation is based on the definitions of static capacitance as charge/voltage and static inductance as flux/current. These definitions, in turn, are responsible for the terms $\dot{K}_y V_y + K_y \dot{V}_y$ and $\dot{L}_z J_z + L_z \dot{J}_z$ in Eq. (48). But when differential (small-signal) capacitances and/or inductances are involved, as in the case of $C_e$ defined by Eqs. (26) and (27), the quantities $\dot{K}_y V_y$ and $\dot{L}_z J_z$ must be deleted from Eq. (48). This follows from the definitions of differential capacitance $\bar{K}$ and differential inductance $\bar{L}$ according to the relations

$$J = \dot{Q} = \frac{dQ}{dV} \dot{V} = \bar{K} \dot{V} \qquad (77)$$

and

$$V = \dot{\phi} = \frac{d\phi}{dJ} \dot{J} = \bar{L}\dot{J} \qquad (78)$$

Accordingly, all terms involving $\dot{K}$ or $\dot{L}$, with whatever subscripts, must be eliminated from Eqs. (60), (61) and (62) when differential capacitances and/or inductances are involved.

## Appendix II

### Tree-Link Sorting Procedure

The object of this sorting procedure is to classify each branch of the network as either a tree-branch or a link and then to rearrange the tables RCON, RDATA, EDATA, and IDATA accordingly. It is assumed that these tables are already in some desired sequence, such as in order of increasing resistance, and that a set of tree branches is to be selected from as near the beginning of this sequence as possible. (This will result in choosing the tree of minimum total resistance if the original ordering is that of increasing resistance.)

Starting with the first branch of the sequence, the network tree is constructed stepwise by adding only those branches which do not form a closed path with the partial tree already constructed. This partial tree, as well as the complete network tree, may consist of several disjoint subtrees. Hence, a branch will be classed as a link if both of its nodes are already connected by the branches of one of these subtrees – or if its initial and final nodes are identical, a trivial case which must nevertheless be handled.

Each branch is examined by comparing its initial node number (+N) and its final node number (-N), obtained from the RCON table, against a master list (MLIST) of nodes contained in the partial tree and also against the individual node list (TLIST) for each subtree. Each such node list consists of a string of bits indicating the presence (1) or absence (0) of the node corresponding to a given bit position in the string.

The following criteria form the logical basis of the tree-link sorting procedure:

(1) If +N = -N, the branch is a link.

(2) If either +N or -N or both are absent from the MLIST, the branch is a tree branch.

(3) If both +N and -N are present in MLIST,

(a) The branch is a link if both +N and -N are also in the same TLIST.

(b) The branch is a tree-branch (joining two previously disjoint subtrees) if +N and -N are in different TLISTS.

As each branch is examined, the appropriate node lists are updated. When the classification as tree-branch or link has been made for the I-th branch, its index (I) is stored in the next available location of TREE or LINK, as appropriate. The two resulting sequences of index numbers are then used for rearranging the tables RCON, RDATA, EDATA, and IDATA after the tree-link sort has been completed.

In the following description of the tree-link sorting procedure, the symbols I, J, K, L and M are indices, hence +N(I) and -N(I) represent the initial and final node numbers of the I-th branch and are obtained from the address and decrement of the I-th word of the RCON table. IMAX is the total number of branches in the network.

1. Set I = J = K = 1, clear MLIST and all TLISTS.
2. If +N(I) = -N(I), go to 16.
3. If +N(I) is absent from MLIST, go to 11.
4. If -N(I) is absent from MLIST, go to 15.
5. Find L for which +N(I) is present in TLIST(L) and save L.
6. If -N(I) is present in TLIST(L), go to 16.

7. Find M for which -N(I) is present in TLIST(M) and save M.
8. If M<L, go to 10.
9. Add TLIST(M) to TLIST(L), clear TLIST(M), and go to 19.
10. Add TLIST(L) to TLIST(M), clear TLIST(L), and go to 19.
11. Add +N(I) to MLIST.
12. If -N(I) is absent from MLIST, go to 14.
13. Find L for which -N(I) is present in TLIST(L), add +N(I) to TLIST(L), and go to 19.
14. Add -N(I) to MLIST, find lowest L for which TLIST(L) is clear, add +N(I) and -N(I) to TLIST(L), and go to 19.
15. Add -N(I) to MLIST, find L for which +N(I) is present in TLIST(L), add -N(I) to TLIST(L), and go to 19.
16. Set LINK(J) = I.
17. If I = IMAX, go to 22.
18. Set I = I + 1, J = J + 1, and go to 2.
19. Set TREE(K) = I.
20. If I = IMAX, go to 22.
21. Set I = I + 1, K = K + 1, and go to 2.
22. Rearrange RCON, RDATA, EDATA and IDATA according to the index numbers in TREE and LINK.

## Appendix III

### Determination of the $B_T$ Matrix

The node-to-datum-path matrix, $B_T$, is determined by an exhaustive search of the network tree. Starting at the datum node, and proceeding always along the branch of lowest serial number connected to each node encountered en route, a path is traced out until it terminates at some particular node of the tree. As the path is traced, a path record (PR) is kept in +1, -1, 0 format showing the branches traversed and their orientations relative to this path in the sense of a node-to-datum traversal.

When the path terminates at some node J, the path record is stored in the J-th column of the $B_T$ matrix. The branch leading to node J is then retraced, its entry in PR deleted, and this branch removed from the tree. Next, the outward path is continued, if possible, again taking the branch of lowest serial number at each successive node until the path terminates once more at, say, node K. The PR is stored in the K-th column of $B_T$, the branch leading to node K retraced, its entry in PR deleted, and the branch removed from the tree. By repeating this procedure until all branches of the tree have been exhausted, the entire $B_T$ matrix may be determined.

In actual practice, both the RCON table and the branch-node matrix are used alternately in tracing out these datum-to-node paths. Since the branch-node matrix is stored columnwise, it provides the simplest means of determining the branch of lowest serial number connected to a particular node. The RCON table, however, is more convenient for finding the number of the node at the far end of a given branch, wherever this is required.

During the search procedure described below, the branch-node matrix, which must include the datum column, is destroyed. Since each entry of this matrix, designated A(I,J), consists of a bit-pair, there are four possible values of each bit-pair of which only three are required for the quantities +1, -1, 0. The fourth value, designated -0, is required to designate the "access" branch leading to each successive node of the path being traced. It is this access branch which must be identified whenever it is necessary to retrace and delete a branch from the tree. Deletion of a branch, after retracing it, is accomplished simply by substituting 0 in place of the +1 or -1 value of the appropriate A(I,J).

In the following description of the procedure for determining the $B_T$ matrix, the symbols I and J represent the row (branch) and column (node) indices. (J = 0 designates the datum node.) As in Appendix II, the symbols +N(I) and -N(I) represent the initial and final node numbers of the I-th branch and are obtained from the I-th word of the RCON table. IMAX is the total number of branches in the tree.

1. Set I = 1, J = 0.
2. If A(I,J) = $\pm 1$, go to 5.
3. If I = IMAX, go to 7.
4. Set I = I + 1 and go to 2.
5. If A(I,J) = +1, set J = -N(I), PR(I) = -1, A(I,J) = -0, I = 1, and go to 2.
6. Set J = + N(I), PR(I) = +1, A(I,J) = -0, I = 1, and go to 2.
7. Transfer PR to column J of $B_T$ matrix and set I = 1.
8. If A(I,J) = -0, go to 11.
9. If I = IMAX, go to 15.
10. Set I = I + 1 and go to 8.
11. If PR(I) = +1, go to 13.
12. Set J = +N(I) and go to 14.
13. Set J = -N(I).
14. Set PR(I) = 0, A(I,J) = 0, I = I + 1, and go to 2.
15. If J = 0, stop. Otherwise, an error has occurred.

## Appendix IV

## Computation of $B_T^t Z_T B_T$

By taking advantage of the diagonal nature of the matrix $Z_T$ and the compact storage format of the matrix $B_T$, a very efficient program can be developed for computing the triple matrix product $B_T^t Z_T B_T$. Since $Z_T$ is diagonal, it follows that the ij-th term of this product is given by the expression

$$(B_T^t Z_T B_T)_{ij} = \sum_{k=1}^{p} (b_{ki} b_{kj}) z_{kk} \qquad (79)$$

where p is the number of tree-branches, $b_{ki}$ and/or $b_{kj}$ are elements of $B_T$, and $z_{kk}$ are the diagonal elements of $Z_T$. Since this product is symmetric, only the diagonal (i=j) and subdiagonal (i>j) terms need be computed.

Let the i-th and j-th columns of $B_T$ be designated by $B_{\cdot i}$ and $B_{\cdot j}$ and let the product $(b_{ki} b_{kj})$ for all values of k be represented by the expression

$$B(i,j) = B_{\cdot i} \otimes B_{\cdot j} \qquad (80)$$

where the special operator $\otimes$ signifies multiplication of the corresponding elements of $B_{\cdot i}$ and $B_{\cdot j}$ and where the elements of the p-vector B(i, j) are +1, -1, or 0. Next define a vector Z(T) comprised of the diagonal elements of $Z_T$. Then, the result defined by Eq. (79) is identical with the scalar product of the two vector B(i, j) and Z(T).

The advantage of using this peculiar method for evaluating Eq. (79) is the fact that the compact storage format of $B_{\cdot i}$ and $B_{\cdot j}$ allows the vector B(i, j) to be computed many elements at a time. Moreover, this computation may be effected by means of logical operations (rather than arithmetic operations) on the bit-pair equivalents of the elements +1, -1 and 0 in $B_{\cdot i}$ and $B_{\cdot j}$, the result being the bit-pair representation of B(i, j). Thus, the calculation of B(i, j) may be carried out very rapidly.

The subsequent computation of the scalar product of B(i, j) and Z(T) involves searching B(i, j) for its nonzero elements and then adding or subtracting the corresponding elements of Z(T). The programming details of this task, however, need not be discussed.

The code actually developed for calculating B(i, j) on the IBM 704 computer will not be described. Instead, an equivalent and much simpler scheme will be outlined to illustrate the principles of the computation. It is assumed that p = 6 and that the machine word length is 6 bits since only 6 different combinations of the elements +1, -1, 0 are encountered. To show this, let $B_{\cdot i}$ and $B_{\cdot j}$, written as row vectors, be

$$(B_{\cdot i}) = (1 \;\; 1 \;\; 1 \;\; 0 \;\; 0 \; -1) \qquad (81)$$

$$(B_{\cdot j}) = (1 \;\; 0 \; -1 \;\; 0 \; -1 \; -1) \qquad (82)$$

Hence B(i, j), also written as a row vector, is

$$B(i,j) = (1 \;\; 0 \; -1 \;\; 0 \;\; 0 \;\; 1) \qquad (83)$$

These three vectors may be represented in machine code by using the bits of one word to indicate the magnitudes (M) and bits of another word to indicate the signs (S) of successive elements:

$$(B_{\cdot i}) = \begin{array}{ll} 111001 & M(I) \\ 000001 & S(I) \end{array} \qquad (84)$$

$$(B_{\cdot j}) = \begin{array}{ll} 101011 & M(J) \\ 001011 & S(J) \end{array} \qquad (85)$$

$$B(i,j) = \begin{array}{ll} 101001 & M(I, J) \\ 001000 & S(I, J) \end{array} \qquad (86)$$

The logical "AND" operation, then, suffices to convert M(I) and M(J) into M(I, J):

$$\begin{array}{lll} & 111001 & M(I) \\ \text{"AND"} & \underline{101011} & M(J) \\ & 101001 & M(I, J) \end{array}$$

or,

$$M(I, J) = M(I) \text{ "AND" } M(J). \qquad (87)$$

Two steps are needed, however, for the calculation of S(I, J). First, S(I) and S(J) are combined using the "exclusive or" operation:

$$\begin{array}{lll} & 000001 & S(I) \\ \text{"EXOR"} & \underline{001011} & S(J) \\ & 001010 & \text{RESULT} \end{array}$$

Then, this result is combined with M(I, J) using the "AND" operation:

$$\begin{array}{lll} & 001010 & \text{RESULT} \\ \text{"AND"} & \underline{001001} & M(I, J) \\ & 001000 & S(I, J) \end{array}$$

Hence, it follows that

$$S(I, J) = \left[ S(I) \text{ "EXOR" } S(J) \right] \text{ "AND" } M(I, J) \qquad (88)$$

Since this procedure treats the magnitude bits separately from the sign bits, it is much more efficient than the procedure actually used in TAP.

## References

1. Nancy G. Brooks and H. S. Long, "A Program for Computing the Transient Response of Transistor Switching Circuits — PE TAP," Technical Report 00. 700, IBM Development Laboratory, Poughkeepsie, New York (1959)
2. F. H. Branin, Jr., "D-C Analysis Portion of PE TAP — A Program for Analyzing Transistor Switching Circuits," Technical Report TR 00. 701, IBM Development Laboratory, Poughkeepsie, New York (1960)
3. J. T. Olsztyn and T. J. Theodoroff, "GMR DYANA Programming Manuals I and II," General Motors Research Laboratories, Warren, Michigan (1959).
4. G. Kron, "A Set of Principles to Interconnect the Solutions of Physical Systems," J. Appl. Phys., 24, 965-980 (1953)
5. F. H. Branin, Jr., "The Relation Between Kron's Method and the Classical Methods of Network Analysis," IRE WESCON convention Record, Part 2, 3-28 (1959). Also available as Technical Report 00.686, IBM Development Laboratory, Poughkeepsie, New York, (1959)
6. F. H. Branin, Jr., "Machine Analysis of Networks," Technical Note 00.490, IBM Development Laboratory, Poughkeepsie, New York (1961)
7. R. J. Domenico, "Simulation of Transistor Switching Circuits on the IBM 704," Trans. IRE, Prof. group on Electronic Computers, EC-3, 242-247 (1957)
8. G. L. Lasher and J. C. Morgan, "A General Method of Predicting the Transient Response of a Nonlinear Circuit," IBM Research Report, RC-7 (1957)
9. T. R. Bashkow, "The A Matrix, New Network Description," Trans. IRE, Prof. Group on Circuit Theory, CT-4, 117-119 (1957).
10. F. B. Hildebrand, "Introduction to Numerical Analysis," McGraw-Hill, New York (1956), pp. 202-208
11. H. J. Gray, Jr., "Numerical Methods in Digital Real-Time Simulation," Quarterly of App. Math XII (2), 133-140 (1954)
12. H. M. Gurk, "The Use of Stability Charts in the Synthesis of Numerical Quadrature Formulae," Quarterly of Appl. Math, XIII (1), 73-78 (1955)
13. F. H. Branin, Jr., "An Abstract Mathematical Basis for Network Analogies and Its Significance in Physics and Engineering," AIEE preprint S-128, April 1961. Also available as Technical Report 00.781, IBM Development Laboratory, Poughkeepsie, New York (1961)
14. J. H. Beaudette and P. A. Honkanen, " PE CANS, Circuit Analysis of Nonlinear Systems," (IBM 7090 Program writeup), IBM Development Laboratory, Poughkeepsie, New York (1961)
15. A. F. Malmberg, private communication.

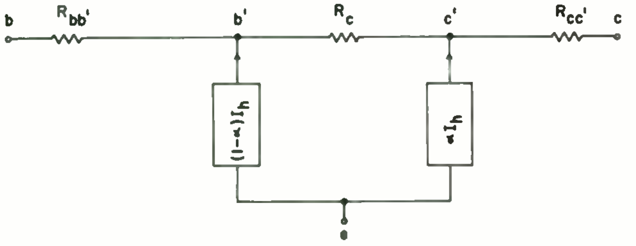


Fig. 1. Nonsaturating PNP Transistor Equivalent Circuit (D-C).

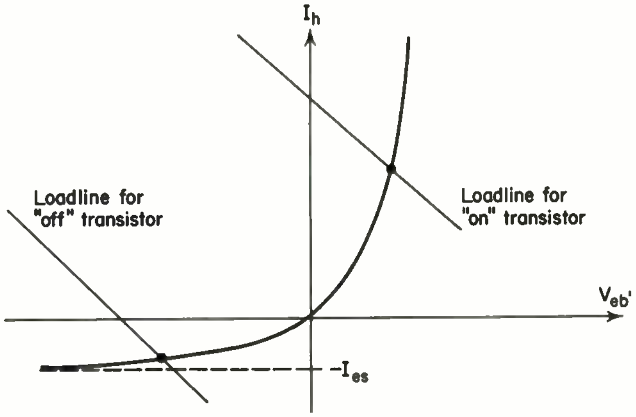


Fig. 3. The Nonlinear D-C Problem.

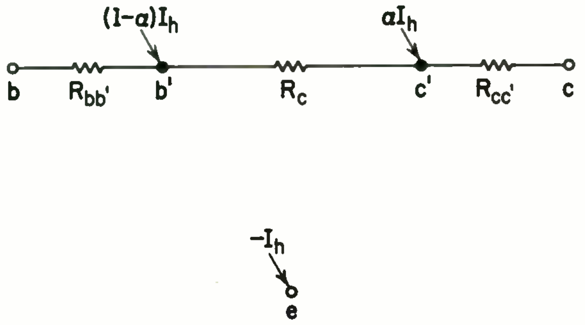


Fig. 2. Nonsaturating PNP Transistor Equivalent Circuit with Nodal Equivalent Current Sources.

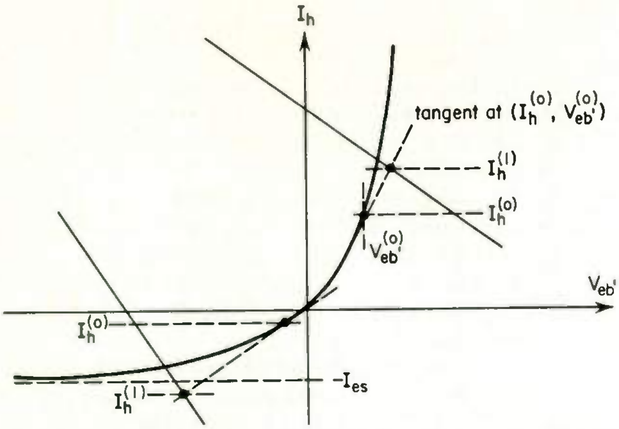


Fig. 4. The Method of Successive Approximations.

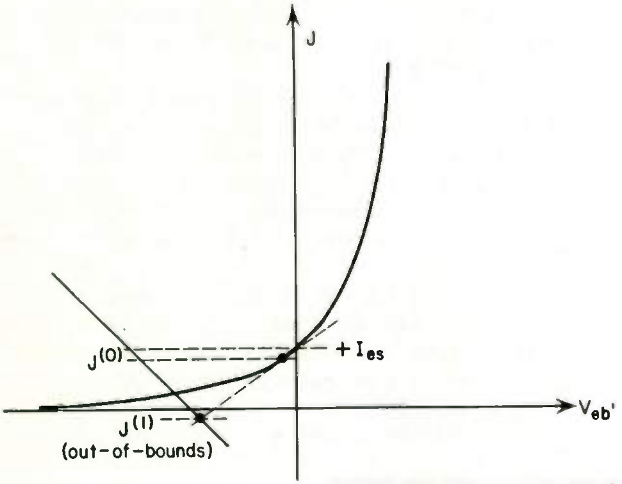


Fig. 5. The Nonlinear D-C Problem in Terms of J.

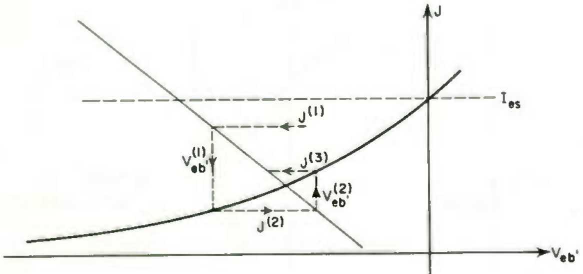


Fig. 6. The Method of Successive Substitutions.

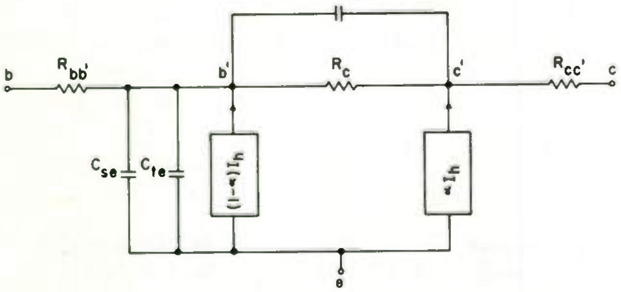


Fig. 7. Nonsaturating PNP Transistor Equivalent Circuit (Transient).

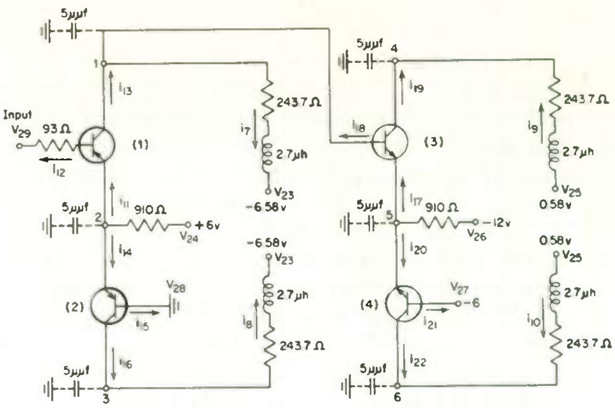


Fig. 8. Four-Transistor Switching Circuit.

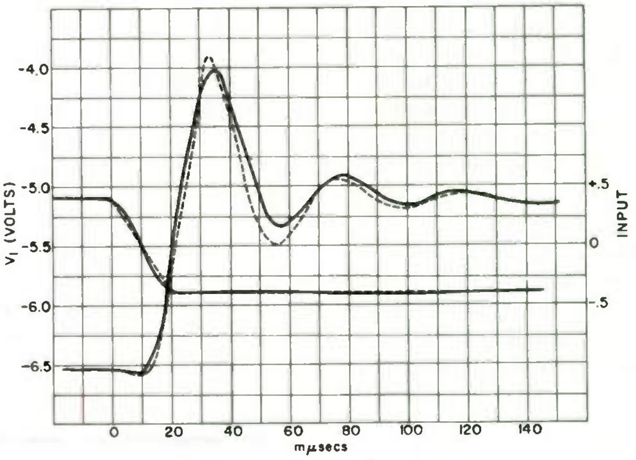


Fig. 9. Collector Voltage for Transistor 1, Negative Input Pulse.

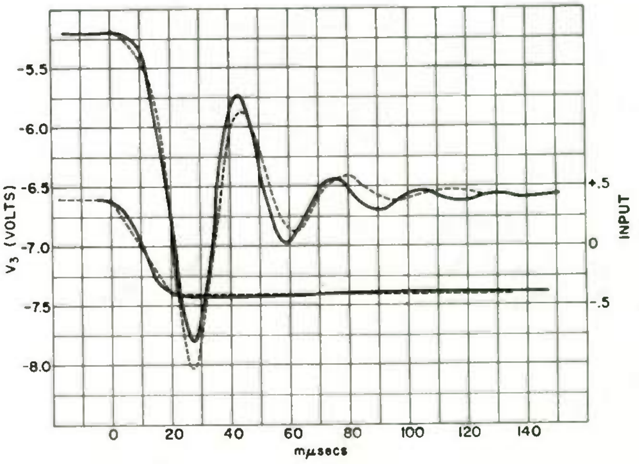


Fig. 10. Collector Voltage for Transistor 2, Negative Input Pulse.

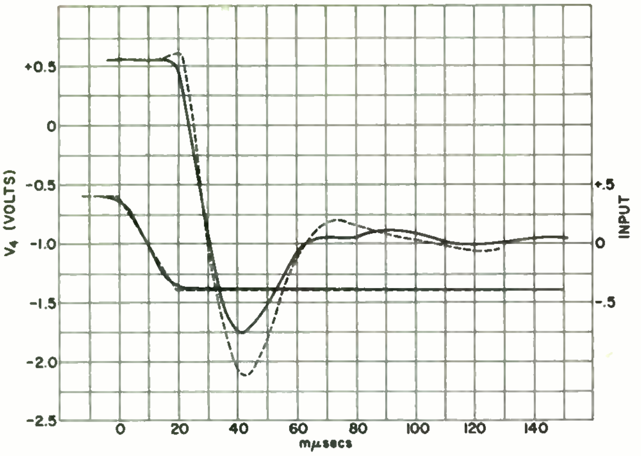


Fig. 11. Collector Voltage for Transistor 3, Negative Input Pulse.

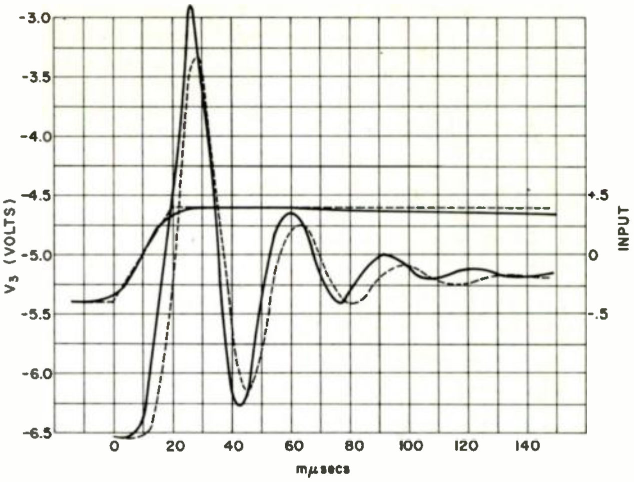


Fig. 14. Collector Voltage for Transistor 2, Positive Input Pulse.

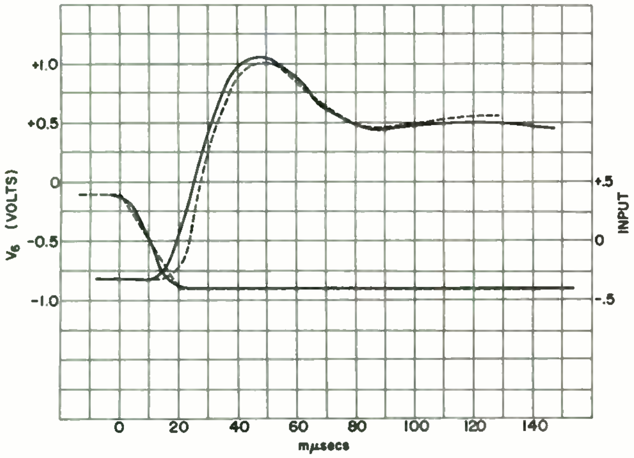


Fig. 12. Collector Voltage for Transistor 4, Negative Input Pulse.

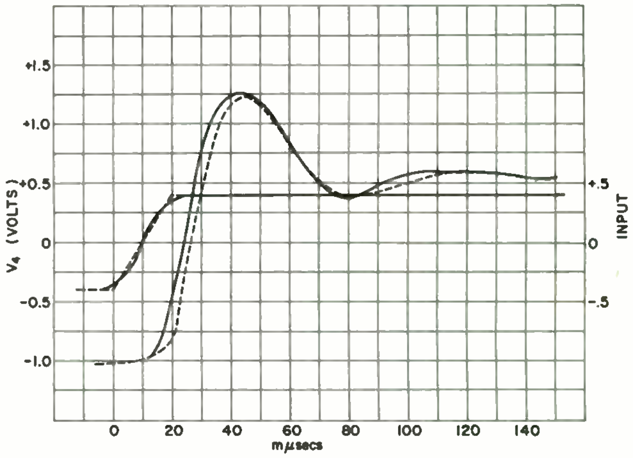


Fig. 15. Collector Voltage for Transistor 3, Positive Input Pulse.

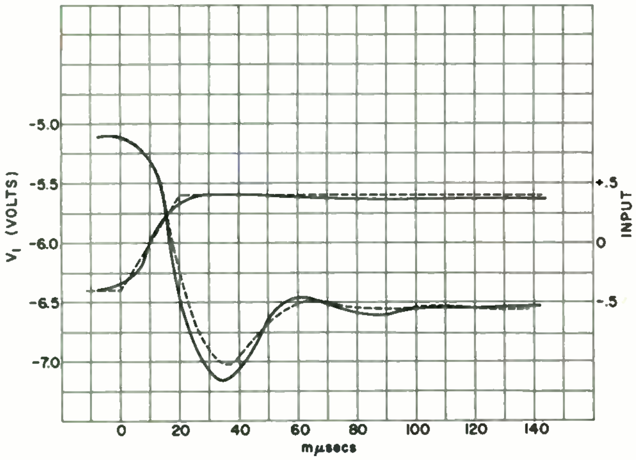


Fig. 13. Collector Voltage for Transistor 1, Positive Input Pulse.

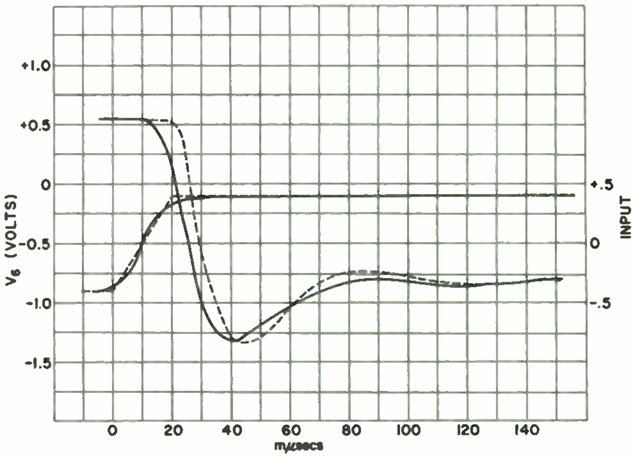


Fig. 16. Collector Voltage for Transistor 4, Positive Input Pulse.

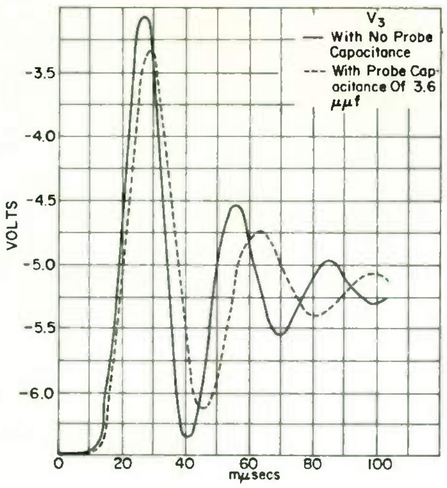


Fig. 17. Collector Voltage for Transistor 2, Positive Input Pulse (Showing Effect of Added Capacitance).

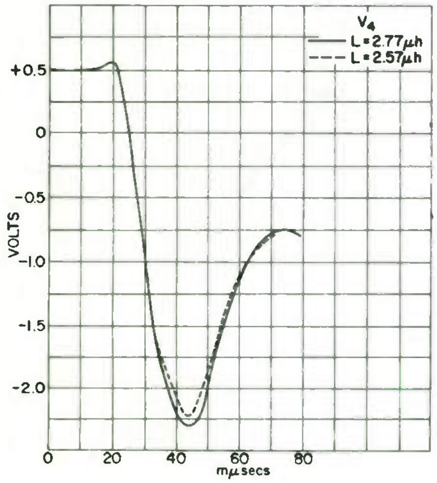


Fig. 18. Collector Voltage for Transistor 3, Negative Input Pulse (Showing Effect of Series Inductance).

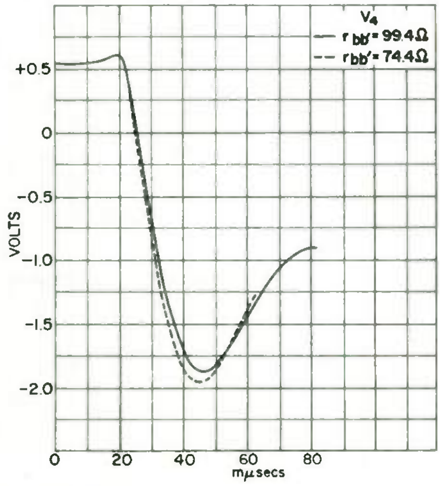


Fig. 19. Collector Voltage for Transistor 3, Negative Input Pulse (Showing Effect of Base Resistance).

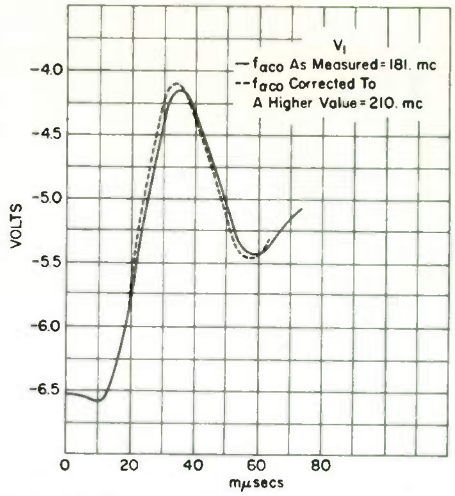


Fig. 20. Collector Voltage for Transistor 3, Negative Input Pulse (Showing Effect of Cutoff Frequency)

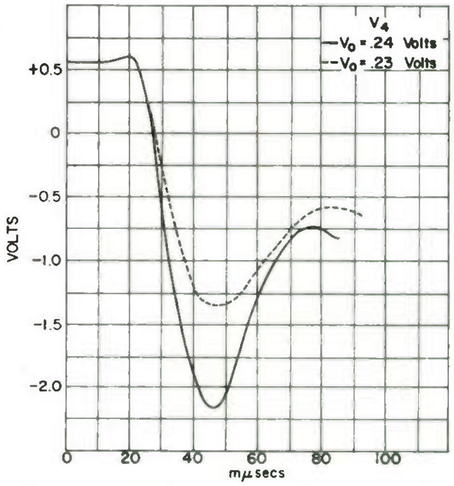


Fig. 21. Collector Voltage for Transistor 3, Negative Input Pulse (Showing Effect of Contact Potential).

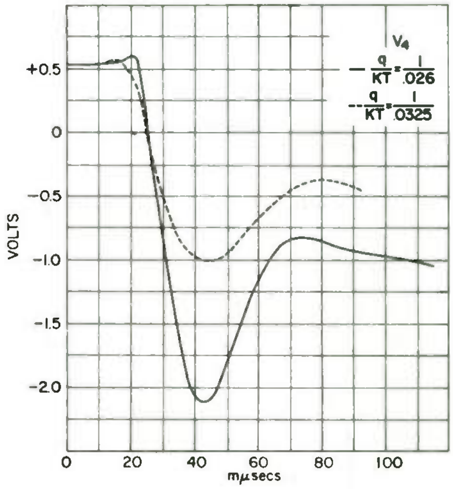


Fig. 22. Collector Voltage for Transistor 3, Negative Input Pulse (Showing Effect of Temperature).