THE BELL SYSTEM TECHNICAL JOURNAL

Volume 48 January 1969 Number 1

Copyright © 1969, American Telephone and Telegraph Company

Some Theorems on Properties of DC Equations of Nonlinear Networks

By I. W. SANDBERG and A. N. WILLSON, JR.

(Manuscript received July 3, 1968)

Several results are presented concerning the equation $F(x) + Ax = B$ (with $F(\cdot)$ a "diagonal" nonlinear mapping of real Euclidean n-space E^n into itself, and A a real $n \times n$ matrix) which plays a central role in the dc analysis of transistor networks. In particular, we give necessary and sufficient conditions on A such that the equation possesses a unique solution x for each real n-vector B and each strictly monotone increasing $F(\cdot)$ that maps E^n onto itself.

There are several direct circuit-theoretic implications of the results. For example, we show that if the short-circuit admittance matrix G of the linear portion of the dc model of a transistor network satisfies a certain dominance condition, then the network cannot be bistable. Therefore, a fundamental restriction on the G matrix of an interesting class of switching circuits is that it must violate the dominance condition.

I. INTRODUCTION

For each positive integer n let $\mathfrak{F}^n$ denote that collection of mappings of the real n-dimensional Euclidean space E^n onto itself, defined by: $F \in \mathfrak{F}^n$ if and only if there exist, for $i = 1, \cdots, n$, strictly monotone increasing functions f_i mapping E^1 onto E^1 such that, for each $x \equiv (x_1, \cdots, x_n)' \in E^n$, $F(x) = (f_1(x_1), \cdots, f_n(x_n))'$.

The main purpose of this paper is to report on some results concerning properties of the equation

$$F(x) + Ax = B, \tag{1}$$

where A is an $n \times n$ matrix of real numbers, F maps E^n into E^n, and $B \varepsilon E^n$. In particular, a condition to be satisfied by A is given which is both necessary and sufficient to guarantee that for each $F \varepsilon \mathfrak{F}^n$ and each $B \varepsilon E^n$ there exists a unique solution of equation (1).

We also study the problem of obtaining bounds on the solution of equation (1). These bounds show that (if $F \varepsilon \mathfrak{F}^n$ and our condition on A is satisfied) the solution depends continuously on B. The bounds are often of use in computing the solution by standard iteration methods such as the Newton-Raphson method. By appealing to a theorem of R. S. Palais it is shown that the bounds can also be used to obtain a theorem essentially the same as, but somewhat weaker than, our principal result.

Several results can be found in the literature which specify sufficient conditions for the existence of a unique solution of equation (1). For example, if A is positive semidefinite then a special case of a theorem of Ref. 1 guarantees the existence of a unique solution of equation (1) for all those $F \varepsilon \mathfrak{F}^n$ which have the property that the slope of each f_i is bounded from above and below by positive constants, and for all $B \varepsilon E^n$. This theorem also specifies that a certain iteration scheme will always converge to the solution.

A theorem of G. J. Minty[2], when applied to equation (1), also implies essentially the same result. The boundedness condition on the slopes of the functions f_i is not required by Minty's theorem. On the other hand, Minty's theorem does not provide a procedure for computing the solution of equation (1).

In Ref. 3 it is proved that a sufficient condition for the existence and uniqueness of a solution of equation (1) for all $F \varepsilon \mathfrak{F}^n$ and $B \varepsilon E^n$ is that A satisfy a weak row-sum dominance condition:

$$a_{ii} \geqq \sum_{j \neq i} | a_{ij} |, \qquad i = 1, \cdots, n.*$$

Other information concerning the location and the computation of the solution is also given in Ref. 3.

The class of matrices satisfying the condition of our theorem (which is defined in Section III and denoted by P_0) includes all positive semidefinite matrices as well as all matrices which satisfy any one of several

* Appendix A contains a simpler proof of a similar result and a proof of a new related result. These results specify convergent algorithms for obtaining the solution.

dominance conditions. Many other matrices are included in P_0; and since the condition of our theorem is both a necessary and a sufficient one, we are assured that P_0 is the largest class of matrices A for which equation (1) has a unique solution for all $F \varepsilon \mathfrak{F}^n$ and all $B \varepsilon E^n$.

II. NONLINEAR NETWORKS

Equation (1) is often encountered in the study of nonlinear electrical networks. In the case of networks containing only resistors (that is, linear resistors with nonnegative resistance), dependent and independent sources, and two-terminal nonlinear resistors that are described by functions in $\mathfrak{F}^1$ (diodes, for example), this is rather obvious.[3] Even for networks which contain more general nonlinear devices, however, equation (1) can often provide a convenient characterization. For example, D. A. Calahan shows in his recent book that the transistor network of Fig. 1 may be described by the equation

$$\begin{pmatrix} I_{es}(e^{qV_r/kT} - 1) \\ I_{cs}(e^{qV_f/kT} - 1) \end{pmatrix} + \begin{bmatrix} 0.0225 & 0.309 \\ -0.168 & 0.494 \end{bmatrix} \begin{pmatrix} V_r \\ V_f \end{pmatrix} = \begin{pmatrix} 0.00177\,V_{cc} \\ -0.188\,V_{cc} \end{pmatrix}$$

if the Ebers–Moll model is used to represent the transistor. (See pp. 13ff of Ref. 4.) In this equation I_{es}, I_{cs}, q, k, T, and V_{cc} all represent fixed real parameters. It is quite trivial to apply the theory of this paper (in particular, Corollary 3 of Section IV) to Calahan's example and prove that this equation has a solution, the solution is unique, and the solution depends continuously on V_{cc}. We also show how bounds on the solution can be obtained.

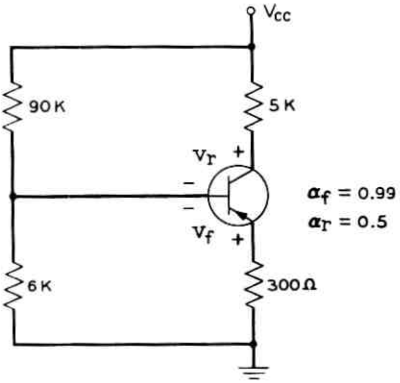


Fig. 1 — Biased transistor-stage.

More generally, it is frequently the case that networks which contain transistors, as well as the previously mentioned linear and nonlinear elements, may be described by the equation

$$TF(x) + Ax = B. \tag{2}$$

In this case, x is a vector whose components are the voltages across the nonlinear resistors and the transistor base-emitter and base-collector voltages. The $n \times n$ matrix A is the y-parameter matrix of the linear n-port network which is obtained by removing all nonlinear resistors and transistors and setting the value of each independent source to zero. The function $TF(x)$ describes the behavior of the nonlinear resistors and the transistors. It happens that the matrix T is nonsingular; therefore equation (2) can be put into the form of equation (1).

Networks which contains inductors and capacitors as well as the memoryless elements already mentioned are of course described by differential equations. Even the study of such networks, however, can often lead to the consideration of equations of the same type as equation (1). One usually finds the solution of such an equation is necessary, for example, when computing the solution of the differential equations by using some implicit numerical integration formula.

The problem of determining the equilibrium states of the above-mentioned dynamic networks is one in which the consideration of equations of type (1) often arises in perhaps a more direct manner. In this regard, if it happens that equation (1) has a unique solution, then the network cannot possibly be bistable.

When the determination of equilibrium states of a transistor network leads first to the consideration of equation (2), then as a rather direct application of our existence and uniqueness theorem it follows that if the matrix A satisfies a weak column-sum dominance condition,

$$a_{ii} \geqq \sum_{j \neq i} | a_{ji} |, \qquad i = 1, \cdots, n,$$

then $T^{-1}A \, \varepsilon \, P_0$ and hence the network has exactly one equilibrium state. This result and related results which are proved in Section IV have the following interesting corollary: One cannot synthesize a bistable network which consists of resistors, inductors, capacitors, diodes, independent voltage and current sources, and one (Ebers–Moll modeled) transistor—or even an arbitrary number of (Ebers–Moll modeled) transistors with a common base connection.

The authors feel that in many respects the main contributions of this paper are in the techniques used to prove the results. For this reason, we have not chosen to summarize all of the results at the outset and relegate proofs to later sections. But rather, the results and the proofs will appear in the order in which they will best illustrate the techniques developed.

III. MATRICES OF CLASSES P AND P_0

The following notation will be used throughout the remainder of the paper: The origin in E^n will be denoted by θ. If D is a diagonal matrix then $D > 0$ ($D \geqq 0$) means that each element of D on the main diagonal is positive (nonnegative).

In Ref. 5 and Ref. 6 M. Fiedler and V. Pták define the classes of matrices denoted by P and P_0. They in fact prove that the following properties of a square matrix A are equivalent:

(*i*) All principal minors of A are positive.

(*ii*) For each vector $x \neq \theta$ there exists an index k such that $x_k y_k > 0$ where $y = Ax$.

(*iii*) For each vector $x \neq \theta$ there exists a diagonal matrix $D_x > 0$ such that the scalar product $\langle Ax, D_x x \rangle > 0$.

(*iv*) For each vector $x \neq \theta$ there exists a diagonal matrix $H_x \geqq 0$ such that $\langle Ax, H_x x \rangle > 0$.

(*v*) Every real eigenvalue of A, as well as of each principal submatrix of A, is positive.

The class of all matrices satisfying one of the above conditions is denoted by P. Fiedler and Pták prove that the following properties of a square matrix A are also equivalent:

(*i*) All principal minors of A are nonnegative.

(*ii*) For each vector $x \neq \theta$ there exists an index k such that $x_k \neq 0$ and $x_k y_k \geqq 0$ where $y = Ax$.

(*iii*) For each vector $x \neq \theta$ there exists a diagonal matrix $D_x \geqq 0$ such that $\langle x, D_x x \rangle > 0$ and $\langle Ax, D_x x \rangle \geqq 0$.

(*iv*) Every real eigenvalue of A, as well as of each principal submatrix of A, is nonnegative.

The class of all matrices satisfying one of the above conditions is denoted by P_0.

The following theorems follow directly from the above definitions.

Theorem 1. If $A \in P_0$ then for every diagonal matrix $\Delta \geqq 0$ ($\Delta > 0$), $\Delta + A \in P_0$ ($\Delta + A \in P$).

Proof: Let $x \neq \theta$. Then, since $A \in P_0$, there exists an index k such that $x_k \neq 0$ and $x_k(Ax)_k \geqq 0$. Thus, $x_k(\Delta x + Ax)_k \geqq 0$ (>0). □

In particular, Theorem 1 implies that if $A \varepsilon P_0$ and $\Delta \geqq 0$ $(\Delta > 0)$ then det $(\Delta + A) \geqq 0$ (>0).

Theorem 2. If $A \varepsilon P$ then $A^{-1} \varepsilon P$.

Proof: Suppose $A \varepsilon P$. Let $x \neq \theta$ be given and let $y = A^{-1}x$. $y \neq \theta$ since A^{-1} is nonsingular. Thus, there exists a diagonal matrix $D > 0$ such that $\langle Ay, Dy\rangle > 0$, which implies $\langle x, DA^{-1}x\rangle > 0$, or $\langle Dx, A^{-1}x\rangle > 0$, or $\langle A^{-1}x, Dx\rangle > 0$. That is, for every $x \neq \theta$ there exists $D > 0$ such that $\langle A^{-1}x, Dx\rangle > 0$. Hence $A^{-1} \varepsilon P$. □

Because of the similarity of the definitions of the classes of matrices P and P_0, one might conjecture that this proposition is also true: *If $A \varepsilon P_0$, and det $A \neq 0$, then $A^{-1} \varepsilon P_0$.* This conjecture is in fact true. Interestingly enough, however, its proof is not obtained as one might at first suspect, by simply modifying the proof of Theorem 2. Moreover, the proof of this conjecture does not even seem to follow directly from any of the above definitions of P_0. Rather, upon making the trivial observation that for every diagonal matrix $D > 0$, det $(A^{-1} + D) =$ det (A^{-1}) $\cdot$det $(D^{-1} + A)\cdot$det (D), the conjecture is easily seen to follow from the fact that det $(D + A) \neq 0$ for every diagonal $D > 0$ if and only if $A \varepsilon P_0$. This fact is a direct corollary to the proof of Theorem 3.

IV. EXISTENCE AND UNIQUENESS THEOREM

The following theorem is the principal result of this paper.

Theorem 3. If A is an $n \times n$ matrix then there exists a unique solution of equation (1) *for each $F \varepsilon \mathfrak{F}^n$ and for each $B \varepsilon E^n$ if and only if $A \varepsilon P_0$.*

Proof: (*if*) Let $A \varepsilon P_0$, $F \varepsilon \mathfrak{F}^n$, and $B \varepsilon E^n$. The solution of equation (1) is then unique (if it exists) since if x and y are both solutions then, using the strict monotonicity property of F, there exists a diagonal matrix $D > 0$ such that $F(x) - F(y) = D(x - y)$. But $[D + A](x - y) = \theta$ and, by Theorem 1, $D + A$ is nonsingular. This means that $x = y$.

We prove the existence of a solution of equation (1) by induction. For $k = 1, \cdots, n$, let

$$F_k(x) = \begin{bmatrix} f_1(x_1) \\ \vdots \\ f_k(x_k) \end{bmatrix}, \qquad A_k = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \cdot & \cdot & \cdot \\ a_{k1} & \cdots & a_{kk} \end{bmatrix}, \qquad B_k = \begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix}.$$

Clearly, $A_k \varepsilon P_0$, $F_k \varepsilon \mathfrak{F}^k$, and $B_k \varepsilon E^k$. Also, it is clear that there exists a unique solution of $F_1(x) + A_1x = B_1$ for each $F_1 \varepsilon \mathfrak{F}^1$ and for each $B_1 \varepsilon E^1$, and that this solution is a continuous function of b_1.

Assume that there exists a unique solution of $F_k(x)+A_kx=B_k$ for each $F_k \varepsilon \mathfrak{F}^k$, $B_k \varepsilon E^k$, and that this solution depends continuously on any scalar parameter η upon which B_k depends continuously. Let the matrices $A_{k,k+1}$ and $A_{k+1,k}$ be defined by

$$A_{k,k+1} = \begin{bmatrix} a_{1,k+1} \\ \vdots \\ a_{k,k+1} \end{bmatrix},$$

$$A_{k+1,k} = [a_{k+1,1} \cdots a_{k+1,k}].$$

Then, for every real number x_{k+1}, the equation

$$F_k(x) + A_kx + A_{k,k+1}x_{k+1} = B_k \tag{3}$$

has a (unique) solution which is a continuous function of x_{k+1} and of η. Let the components of this solution be denoted by $x_i = m_i(x_{k+1}, \eta)$, for $i = 1, \cdots, k$, and define the vector $M_k(x_{k+1}, \eta)$ by $M_k = (m_1, \cdots, m_k)^t$.

We now prove that the function

$$\varphi(x_{k+1}, \eta) \equiv A_{k+1,k}M_k(x_{k+1}, \eta) + a_{k+1,k+1}x_{k+1} - b_{k+1}(\eta)$$

is monotone increasing in x_{k+1}: Let $x_{k+1}^1, x_{k+1}^2 \varepsilon E^1$ with $x_{k+1}^1 < x_{k+1}^2$. Then, if M^1 (M^2) denotes the solution of equation (3) when $x_{k+1} = x_{k+1}^1$ (x_{k+1}^2), we have

$$F_k(M^2) - F_k(M^1) + A_k(M^2 - M^1) + A_{k,k+1}(x_{k+1}^2 - x_{k+1}^1) = \theta.$$

Because of the strict monotonicity of the function F_k, however, there exists a $k \times k$ diagonal matrix $\Delta > 0$ such that

$$F_k(M^2) - F_k(M^1) = \Delta(M^2 - M^1).$$

Hence,

$$M^2 - M^1 = -[\Delta + A_k]^{-1}A_{k,k+1}(x_{k+1}^2 - x_{k+1}^1).$$

Thus,

$$\varphi(x_{k+1}^2) - \varphi(x_{k+1}^1) = \{a_{k+1,k+1} - A_{k+1,k}[\Delta + A_k]^{-1}A_{k,k+1}\}(x_{k+1}^2 - x_{k+1}^1).$$

But then, from the easily verified relation

$$a_{k+1,k+1} - A_{k+1,k}[\Delta + A_k]^{-1}A_{k,k+1} = \frac{\det\left(\begin{bmatrix} \Delta & \begin{matrix} 0 \\ \vdots \end{matrix} \\ 0 \cdots & 0 \end{bmatrix} + A_{k+1}\right)}{\det(\Delta + A_k)}$$

and since

$$\det\,(\Delta + A_k) > 0, \qquad \det\left[\begin{bmatrix} \Delta & \begin{matrix} 0 \\ \vdots \end{matrix} \\ 0 \cdots & 0 \end{bmatrix} + A_{k+1}\right] \geqq 0 \text{ (Theorem 1)},$$

and $x^2_{k+1} - x^1_{k+1} > 0$, it follows that $\varphi(x^2_{k+1}) \geqq \varphi(x^1_{k+1})$.

Now since φ is monotone increasing and, obviously, continuous in x_{k+1}, it follows that the left side of the equation

$$f_{k+1}(x_{k+1}) + \varphi(x_{k+1}) = 0 \tag{4}$$

is a strictly monotone increasing function mapping E^1 onto E^1, and hence equation (4) has a unique solution. If x^0_{k+1} denotes this solution then

$$x^0 \equiv \begin{bmatrix} m_1(x^0_{k+1}) \\ \vdots \\ m_k(x^0_{k+1}) \\ x^0_{k+1} \end{bmatrix}$$

is the (unique) solution of

$$F_{k+1}(x) + A_{k+1}x = B_{k+1}\,.$$

We must now prove that this solution is a continuous function of any scalar parameter η upon which B_{k+1} depends continuously. It suffices to prove that x_{k+1} depends continuously on η (see equation (3)). This may be done as follows:

Let x^0_{k+1} be the solution of equation (4) corresponding to $\eta = \eta^0$. That is, let

$$f_{k+1}(x^0_{k+1}) + \varphi(x^0_{k+1}\,,\,\eta^0) = 0,$$

and let $\epsilon > 0$ be given. Since f_{k+1} is a strictly monotone increasing mapping of E^1 onto E^1, so is f^{-1}_{k+1}, and hence f^{-1}_{k+1} is continuous. Hence, there exists $\delta' > 0$ such that if $|\, f_{k+1}(x^0_{k+1}) - f_{k+1}(x_{k+1}) \,| < \delta'$ then $|\, x^0_{k+1} - x_{k+1} \,| < \epsilon$. Since φ is a continuous function of η, there exists $\delta > 0$ such that $|\, \eta^0 - \eta \,| < \delta$ implies $|\, \varphi(x^0_{k+1}\,,\,\eta^0) - \varphi(x^0_{k+1}\,,\,\eta) \,| < \delta'$. If $|\, \eta^0 - \eta \,| < \delta$, and

$$f_{k+1}(x_{k+1}) + \varphi(x_{k+1}\,,\,\eta) = 0,$$

then,

$$f_{k+1}(x^0_{k+1}) - f_{k+1}(x_{k+1}) + \varphi(x^0_{k+1}, \eta) - \varphi(x_{k+1}, \eta)$$
$$= -[\varphi(x^0_{k+1}, \eta^0) - \varphi(x^0_{k+1}, \eta)].$$

But since both f_{k+1} and φ are monotone increasing in x_{k+1},

$$(x^0_{k+1} - x_{k+1})[f_{k+1}(x^0_{k+1}) - f_{k+1}(x_{k+1})] \geqq 0,$$

and

$$(x^0_{k+1} - x_{k+1})[\varphi(x^0_{k+1}, \eta) - \varphi(x_{k+1}, \eta)] \geqq 0.$$

Therefore,

$$|(x^0_{k+1} - x_{k+1})[f_{k+1}(x^0_{k+1}) - f_{k+1}(x_{k+1})]|$$
$$\leqq |(x^0_{k+1} - x_{k+1})[\varphi(x^0_{k+1}, \eta^0) - \varphi(x^0_{k+1}, \eta)]|.$$

Now, if $x^0_{k+1} = x_{k+1}$ then of course $|x^0_{k+1} - x_{k+1}| < \epsilon$. Otherwise,

$$|f_{k+1}(x^0_{k+1}) - f_{k+1}(x_{k+1})| \leqq |\varphi(x^0_{k+1}, \eta^0) - \varphi(x^0_{k+1}, \eta)|.$$

But then,

$$|f_{k+1}(x^0_{k+1}) - f_{k+1}(x_{k+1})| < \delta',$$

and hence $|x^0_{k+1} - x_{k+1}| < \epsilon$. Thus, x_{k+1} is a continuous function of η.

(*only if*) Suppose $A \notin P_0$. If $\det A < 0$ then for sufficiently small $\zeta > 0$, $\det(\zeta I + A) < 0$. For sufficiently large ζ, however,

$$\det(\zeta I + A) = \zeta^n \cdot \det\left(I + \frac{1}{\zeta} A\right) > 0.$$

Thus, since $\det(\zeta I + A)$ is a continuous function of ζ, there is some value of $\zeta > 0$ such that $\det(\zeta I + A) = 0$. For this value of ζ let $F(x) = \zeta I x$. Clearly, for this choice of $F \in \mathfrak{F}^n$, equation (1) cannot have a unique solution.

If $\det A \geqq 0$, but A has a negative principal minor, we can still find a diagonal matrix $\Delta > 0$ such that $\det(\Delta + A) = 0$; however, in this case Δ will not, in general, be simply the identity matrix multiplied by a positive constant ζ.

For some positive integer $k < n$ let A have a $k \times k$ principal minor which is negative and let

$$\Delta^{(1)} = \operatorname{diag}[\delta_1, \cdots, \delta_n].$$

Since the determinant of $\Delta + A$ is not altered if any two rows and then the corresponding pair of columns are interchanged we may, without loss of generality, assume that the matrix A is partitioned as

$$A = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix},$$

where A_1 is a $k \times k$ matrix with $\det A_1 < 0$. Let $\xi > 0$ be chosen so small that $\det (\xi I + A_1) < 0$, and let $\delta_1 = \cdots = \delta_k = \xi$. Now, if $\delta_{k+1} = \cdots = \delta_n = \zeta > 0$, then

$$\det (\Delta^{(1)} + A) = \det \begin{bmatrix} \xi I + A_1 & A_2 \\ A_3 & \zeta I + A_4 \end{bmatrix}$$

$$= \zeta^{n-k} \cdot \det \begin{bmatrix} \xi I + A_1 & A_2 \\ \frac{1}{\zeta} A_3 & I + \frac{1}{\zeta} A_4 \end{bmatrix}.$$

Thus, for $\zeta > 0$ chosen to be sufficiently large, $\det (\Delta^{(1)} + A) < 0$. ($\det (\Delta^{(1)} + A) \to \zeta^{n-k} \cdot \det (\xi I + A_1) < 0$ as $\zeta \to \infty$.) Now, if for $\eta > 0$, $\Delta^{(2)} = \eta I$, then it is clear that for η chosen sufficiently large, $\det [\Delta^{(2)} + A] = \eta^n \cdot \det (I + (1/\eta)A) > 0$. Thus, if

$$\Delta(\epsilon) = \epsilon \Delta^{(1)} + (1 - \epsilon) \Delta^{(2)},$$

it is clear, since $\det [\Delta(0) + A] > 0$ and $\det [\Delta(1) + A] < 0$ and since $\det [\Delta(\epsilon) + A]$ is a continuous function on $0 \leqq \epsilon \leqq 1$, that there is a value of $\epsilon > 0$ $(0 < \epsilon < 1)$ such that $\det [\Delta(\epsilon) + A] = 0$. For this value of ϵ, $\Delta(\epsilon) > 0$ is the required diagonal matrix. □

Notice that our proof shows that if $F \varepsilon \mathfrak{F}^n$ and $A \varepsilon P_0$, then the solution of equation (1) depends continuously on any scalar parameter upon which B depends continuously. The arguments of Section V show, under these assumptions on F and A, that the operator $(F + A)^{-1}$ is in fact a continuous map of E^n into itself.

In the proof of Theorem 3 we see that the uniqueness of the solution follows simply from the hypotheses that each f_i is strictly monotone increasing and that $A \varepsilon P_0$. The additional hypotheses that each f_i is continuous and maps E^1 *onto* E^1 are not necessary (continuity of each f_i is not explicitly hypothesized, but follows from the "monotonicity" and "onto" hypotheses). Hence, we have:

Corollary 1. If, for $i = 1, \cdots, n$, S_i is a subset of E^1, and if $S = S_1 \times \cdots \times S_n$, and if $F(x) = (f_1(x_1), \cdots, f_n(x_n))'$, where each f_i maps E^1 into E^1 and is strictly monotone increasing on S_i, then if $A \varepsilon P_0$ and $B \varepsilon E^n$, there exists at most one solution of equation (1) *in S.*

We now prove another interesting corollary of Theorem 3. We first define some additional notation.

For each positive integer n let $\mathfrak{S}^n$ denote the collection of all subsets of E^n defined by: $S \varepsilon \mathfrak{S}^n$ if and only if $S = S^1 \times \cdots \times S^n$ where, for $i = 1, \cdots, n$, $S^i \subset E^1$ and S^i has the same cardinality as E^1. For each $S \subset \mathfrak{S}^n$ we define the collection $\mathfrak{F}^n(S)$ of functions mapping S onto E^n by: $F \varepsilon \mathfrak{F}^n(S)$ if and only if there exist, for $i = 1, \cdots, n$, strictly monotone increasing functions f_i mapping S^i onto E^1 such that for each $x \varepsilon S^n$, $F(x) = (f_1(x_1), \cdots, f_n(x_n))^t$.

Corollary 2. If A is an $n \times n$ matrix and the collection $\mathfrak{F}^n(S)$ is non-empty then there exists a unique solution of the equation

$$F_1(x) + AF_2(x) = B \tag{5}$$

for each $F_1 \varepsilon \mathfrak{F}^n(S)$, $F_2 \varepsilon \mathfrak{F}^n(S)$, and each $B \varepsilon E^n$ if and only if $A \varepsilon P_0$.

Proof: Since $F_2 \varepsilon \mathfrak{F}^n(S)$, $F_2^{-1} : E^n \to S$ exists and $F_1 \circ F_2^{-1} \varepsilon \mathfrak{F}^n$. Thus, there exists a unique solution of equation (5) if and only if there exists a unique solution of

$$F_1(F_2^{-1}(y)) + Ay = B. \quad \square$$

As special cases of Corollary 2 we have: there exists a unique solution of each of the equations

$$F_1(x) + AF_2(x) = B,$$

and

$$x + AF(x) = B,$$

for each F_1, F_2, $F \varepsilon \mathfrak{F}^n$ and each $B \varepsilon E^n$ if and only if $A \varepsilon P_0$.

In Theorem 3 (and Corollary 2) the hypothesis that each of the functions f_i is an onto mapping is quite necessary in order to guarantee the *existence* of a solution for each $A \varepsilon P_0$. In the following example all of the hypotheses of Theorem 3 except this one are satisfied:

$$\begin{aligned} e^{x_1} + x_1 - x_2 &= 1 \\ e^{x_2} - x_1 + x_2 &= -2. \end{aligned}$$

It is of course impossible for these equations to have a solution since, by adding both sides, we find that the solution would have to satisfy

$$e^{x_1} + e^{x_2} = -1,$$

which is absurd.

Even though the functions f_i are not "onto," it is still possible to specify sufficient conditions for the existence of a unique solution of equation (5) [and equation (1)] by strengthening the hypothesis on the matrix A—namely, by requiring that $A \in P$. This is the essence of Corollary 3. We first require additional notation.

With $\mathcal{S}^n$ defined as above, we define, for each $S \in \mathcal{S}^n$, the collection of functions $\mathfrak{F}_0^n(S)$ mapping S into E^n by: $F \in \mathfrak{F}_0^n(S)$ if and only if there exist, for $i = 1, \cdots, n$, monotone increasing functions f_i mapping S^i onto a connected set in E^1 such that, for each $x \in S$, $F(x) = (f_1(x_1), \cdots, f_n(x_n))^t$. When $S = E^n$ we denote $\mathfrak{F}_0^n(S)$ by $\mathfrak{F}_0^n$.

Corollary 3. If A is an $n \times n$ matrix then there exists a unique solution of equation (5) *for each $F_1 \in \mathfrak{F}_0^n(S)$, $F_2 \in \mathfrak{F}^n(S)$, or $F_1 \in \mathfrak{F}^n(S)$, $F_2 \in \mathfrak{F}_0^n(S)$, and for each $B \in E^n$, if $A \in P$.*

Proof: If $F_2 \in \mathfrak{F}^n(S)$, $F_2^{-1} : E^n \rightarrow S$ exists and $F_1 \circ F_2^{-1} \in \mathfrak{F}_0^n$. Thus, in this case, there exists a unique solution of equation (5) if there exists a unique solution of

$$F_1(F_2^{-1}(y)) + Ay = B. \tag{6}$$

Now, since $A \in P$, it follows from the fact that the determinant of a matrix is a continuous function of each of its elements, that there is a matrix $A^* \in P \subset P_0$ and an $\epsilon > 0$, such that $A = \epsilon I + A^*$. Hence, equation (6) is equivalent to

$$F(y) + A^*y = B, \tag{7}$$

where we have defined

$$F(y) \equiv F_1(F_2^{-1}(y)) + \epsilon Iy.$$

But, since $F_1 \circ F_2^{-1} \in \mathfrak{F}_0^n$ and $\epsilon I \in \mathfrak{F}^n$, it follows that $F \in \mathfrak{F}^n$. Therefore, since $A^* \in P_0$, equation (7) and hence equation (6) and hence equation (5) have unique solutions.

The case when $F_1 \in \mathfrak{F}^n(S)$ and $F_2 \in \mathfrak{F}_0^n(S)$ can be reduced to the case just considered by making the simple observations that, in this case, equation (5) has a unique solution if

$$A^{-1}F_1(x) + F_2(x) = A^{-1}B$$

has a unique solution, and $A \in P$ implies $A^{-1} \in P$ (Theorem 2). □

In Corollary 3 a *sufficient* condition is given for the existence of a unique solution to say equation (1) when the functions f_i which specify F are not necessarily mappings onto E^1. That the condition ($A \in P$) is not *necessary* is easily demonstrated by the counterexample: Let $F \in \mathfrak{F}_0^2$ and

$B \varepsilon E^2$; then the equations

$$f_1(x_1) - x_2 = b_1 , \text{ and } f_2(x_2) + x_1 = b_2$$

have a unique solution in spite of the fact that the matrix

$$A \equiv \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \not\varepsilon P.$$

This is true because the function $f_2(f_1(x_1) - b_1)$ is obviously a continuous monotone increasing function of x_1, and hence the left side of the equation

$$f_2(f_1(x_1) - b_1) + x_1 = b_2 \tag{8}$$

is a strictly monotone increasing mapping of E^1 onto E^1. Thus equation (8) has a unique solution.

V. BOUNDED SOLUTIONS AND RELATED PROBLEMS

For many systems whose behavior is described by an equation having the form of equation (1), the vector B may be regarded as the system's input and the vector x may be regarded as the system's response, or output. Thus, if a sequence $B^1, B^2, B^3, \cdots$ of input vectors for the system is given, the corresponding sequence $x^1, x^2, x^3, \cdots$ of output vectors is specified by equation (1). An important property that such systems might have is that of producing a bounded sequence of output vectors for each bounded sequence of input vectors; that is, the property that whenever an input sequence $B^1, B^2, B^3, \cdots$ is contained in some bounded region of E^n, then the corresponding output sequence $x^1, x^2, x^3, \cdots$ (exists and) also is contained in some bounded region of E^n. By considering matrices A which are not members of P_0, it is easy to demonstrate that all equations having the form of equation (1) do not have this property. For example, if $f(x) \equiv x + e^x$ ($f \varepsilon \mathfrak{F}^1$), then the sequence of solutions of the equation $f(x) + (-1)x = b$ is unbounded, even though the sequence $b = 1, \frac{1}{2}, \frac{1}{3}, \cdots$ of inputs is bounded. The fact that one *must* resort to matrices A which are not in P_0, and the fact that by choosing *any* $A \not\varepsilon P_0$, an example of the above kind can be constructed by an appropriate choice of $F \varepsilon \mathfrak{F}^n$, follows from our next theorem.

Theorem 4. If A is an $n \times n$ matrix then $A \varepsilon P_0$ if and only if for each $F \varepsilon \mathfrak{F}^n$ and each unbounded sequence of points $x^1, x^2, x^3, \cdots$ in E^n, the corresponding sequence $B^1, B^2, B^3, \cdots$ ($B^k = F(x^k) + Ax^k$, $k = 1, 2, 3, \cdots$) is unbounded.

Proof: (*if*) If $A \notin P_0$ then, as shown in the "only if" part of the proof of Theorem 3, there exists a diagonal matrix $D > 0$ such that $D + A$ is singular. Hence, there exists some point $p \in E^n$, $p \neq \theta$, such that $Dp + Ap = \theta$. Let p_j, the j-th component of p, be nonzero. Let the diagonal elements of the matrix D be denoted by $d_1, \cdots, d_n$ and let the mapping $F \in \mathfrak{F}^n$ be defined by

$$f_i(x_i) = \begin{cases} d_i x_i , & \text{for} \quad i \neq j, \\ d_i x_i + e^{x_i}, & \text{for} \quad i = j. \end{cases}$$

If $p_j < 0$ let $\epsilon = 1$, if $p_j > 0$ let $\epsilon = -1$. Consider the unbounded sequence $x^1, x^2, x^3, \cdots$ defined by $x^k = k \cdot \epsilon \cdot p$, for $k = 1, 2, 3, \cdots$. The members of the corresponding sequence $B^1, B^2, B^3, \cdots$ are $B^k = (0, \cdots, 0, e^{k\epsilon p_j}, 0, \cdots, 0)^t$, $k = 1, 2, 3, \cdots$, where the j-th element of each B^k is nonzero. Since for $k = 1, 2, 3, \cdots, k \epsilon p_j < 0$, the sequence $B^1, B^2, B^3, \cdots$ is bounded.

(*only if*) Our proof of the "only if" part of Theorem 4 consists of proving Theorem 5 which is referred to later for another purpose. □

Theorem 5. Let $F \equiv (f_1(\cdot), \cdots, f_n(\cdot))^t \in \mathfrak{F}^n$, $A \in P_0$, *and, for* $i = 1, \cdots, n$, $\alpha_i \leqq \beta_i$ *be given. There exist, for* $i = 1, \cdots, n$, *real numbers* $\gamma_i \leqq \delta_i$ *such that for any* $B \equiv (b_1, \cdots, b_n)^t \in E^n$ *with* $\alpha_i \leqq b_i \leqq \beta_i$ *for* $i = 1, \cdots, n$, *if* x *satisfies equation* (1) *then* $\gamma_i \leqq x_i \leqq \delta_i$ *for* $i = 1, \cdots, n$.

Proof of Theorem 5: We first prove a useful lemma.

Lemma 1. Let f *be a strictly monotone increasing mapping of* E^1 *onto itself. Let* x, b, α, β *be real numbers such that* $xf(x) \leqq xb$ *with* $\alpha \leqq b \leqq \beta$. *Then* $\gamma \leqq x \leqq \delta$, *where* $\gamma = \min \{f^{-1}(\alpha), 0\}$ *and* $\delta = \max \{f^{-1}(\beta), 0\}$.

Proof: Let $\alpha \leqq b \leqq \beta$ and define $\gamma = \min \{f^{-1}(\alpha), 0\}$ and $\delta = \max \{f^{-1}(\beta), 0\}$. Let x satisfy $xf(x) \leqq xb$. Then $x(f(x) - b) \leqq 0$. Clearly, $\gamma \leqq 0 \leqq \delta$ and hence if $x = 0$ then $\gamma \leqq x \leqq \delta$. If $x > 0$ then $f(x) \leqq b \leqq \beta$ which implies $x \leqq f^{-1}(\beta) \leqq \delta$ and hence $\gamma \leqq 0 < x \leqq \delta$. If $x < 0$, then $f(x) \geqq b \geqq \alpha$ which implies $x \geqq f^{-1}(\alpha) \geqq \gamma$ and hence $\gamma \leqq x < 0 \leqq \delta$. □

(*Proof of Theorem 5*) Since $A \in P_0$ there exists $k_1 \in \{1, \cdots, n\}$ such that $x_{k_1}(Ax)_{k_1} \geqq 0$ and hence,

$$x_{k_1} b_{k_1} = x_{k_1} f_{k_1}(x_{k_1}) + x_{k_1}(Ax)_{k_1} \geqq x_{k_1} f_{k_1}(x_{k_1}).$$

Thus, by Lemma 1, there exist $\gamma_{k_1}^{(1)} = \gamma_{k_1}^{(1)}(f_{k_1}, \alpha_{k_1})$ and $\delta_{k_1}^{(1)} = \delta_{k_1}^{(1)}(f_{k_1}, \beta_{k_1})$

such that $\gamma_{k_1}^{(1)} \leqq x_{k_1} \leqq \delta_{k_1}^{(1)}$. Now if F_{n-1} denotes the mapping of E^{n-1} onto E^{n-1} defined by

$$F_{n-1} \equiv (f_1(\cdot), \cdots, f_{k_1-1}(\cdot), f_{k_1+1}(\cdot), \cdots, f_n(\cdot))^t,$$

if A_{n-1} denotes the $(n - 1) \times (n - 1)$ matrix obtained from A by deleting the k_1-st row and column (note that $A_{n-1} \varepsilon P_0$), if

$$a_{n-1} = (a_{1.k_1}, \cdots, a_{k_1-1.k_1}, a_{k_1+1.k_1}, \cdots, a_{n.k_1})^t,$$

and if

$$B_{n-1} = (b_1, \cdots, b_{k_1-1}, b_{k_1+1}, \cdots, b_n)^t,$$

then

$$F_{n-1}(x) + A_{n-1}x = B_{n-1} - a_{n-1}x_{k_1}.^*$$

Since $A_{n-1} \varepsilon P_0$, there is a $k_2 \varepsilon \{1, \cdots, k_1-1, k_1+1, \cdots, n\}$ such that $x_{k_2}(A_{n-1}x)_{k_2} \geqq 0$ and hence, as before,

$$x_{k_2}(b_{k_2} - a_{k_2.k_1}x_{k_1}) \geqq x_{k_2}f_{k_2}(x_{k_2}).$$

But, if $a_{k_2.k_1} \leqq 0$, then

$$\alpha_{k_2} - a_{k_2.k_1}\gamma_{k_1}^{(1)} \leqq b_{k_2} - a_{k_2.k_1}x_{k_1} \leqq \beta_{k_2} - a_{k_2.k_1}\delta_{k_1}^{(1)},$$

and if $a_{k_2.k_1} > 0$, then

$$\alpha_{k_2} - a_{k_2.k_1}\delta_{k_1}^{(1)} \leqq b_{k_2} - a_{k_2.k_1}x_{k_1} \leqq \beta_{k_2} - a_{k_2.k_1}\gamma_{k_1}^{(1)}.$$

Therefore, by Lemma 1, there is a $\gamma_{k_2}^{(1)} = \gamma_{k_2}^{(1)}(f_{k_2}, \alpha_{k_2} - a_{k_2.k_1}\gamma_{k_1}^{(1)})$ and $\delta_{k_2}^{(1)} = \delta_{k_2}^{(1)}(f_{k_2}, \beta_{k_2} - a_{k_2.k_1}\delta_{k_1}^{(1)})$ such that $\gamma_{k_2}^{(1)} \leqq x_{k_2} \leqq \delta_{k_2}^{(1)}$ if $a_{k_2.k_1} \leqq 0$, and similarly for $a_{k_2\,k_1} > 0$.

The above process may be repeated successively until the n pairs of real numbers $\gamma_{k_i}^{(1)}$, $\delta_{k_i}^{(1)}$, $(i = 1, \cdots, n)$ have been obtained. Thus, for any given B with $\alpha_i \leqq b_i \leqq \beta_i$ for $i = 1, \cdots, n$, the components of the solution x of equation (1) will be bounded by these pairs of numbers, provided it is known at each step which coordinate k_i to choose. The appropriate coordinate choice, however, will in general depend on the particular solution x which is associated with the given B. For different input vectors B the appropriate choice will in general be different. Therefore, in order to obtain bounds on x which are valid for all B with $\alpha_i \leqq b_i \leqq \beta_i$ $(i = 1, \cdots, n)$ we must consider each of the $n!$ permutations of the coordinates $\{1, \cdots, n\}$ and, for each one, generate the set of bounds $\{\gamma_{k_i}^{(\nu)}, \delta_{k_i}^{(\nu)}: i = 1, \cdots, n\}$ for $\nu = 1, \cdots, n!$. We then define $\gamma_i =$

* In this equation x is understood to be $(x_1, \cdots, x_{k_1-1}, x_{k_1+1}, \cdots, x_n)^t$.

$\min \{\gamma_i^{(\nu)}: \nu = 1, \cdots, n!\}$ and $\delta_i = \max \{\delta_i^{(\nu)}: \nu = 1, \cdots, n!\}$ for $i = 1, \cdots, n$. Then, for each B with $\alpha_i \leqq b_i \leqq \beta_i$ for $i = 1, \cdots, n$, we have that $\gamma_i \leqq x_i \leqq \delta_i$ for $i = 1, \cdots, n$, since at least one of the sets of bounds $\{\gamma_{k_i}^{(\nu)}, \delta_{k_i}^{(\nu)}: i = 1, \cdots, n\}$ must always apply. □

If the matrix A of Theorem 5 satisfies a stronger condition than $A \in P_0$ (that is, if A satisfies a weak row-sum dominance condition),

$$a_{ii} \geqq \sum_{j \neq i} | a_{ij} |, \qquad \text{for} \quad i = 1, \cdots, n,$$

it is possible to use a method that requires much less computational effort than that of Theorem 5 to compute the vectors γ and δ whose components bound the corresponding components of the solution of equation (1). This method of computing the bounds, a straightforward generalization of an idea presented in Ref. 3, is explained in Appendix B.

From Theorems 3 and 4 we now have the result: *Every bounded input sequence* B^1, B^2, B^3, $\cdots$ *is mapped by equation* (1) *into a bounded output sequence* x^1, x^2, x^3, $\cdots$, *for each* $F \in \mathfrak{F}^n$, *if and only if* $A \in P_0$.

In the proof of Theorem 5, the number of real numbers $\gamma_{k_i}^{(\nu)}$, $\delta_{k_i}^{(\nu)}$ which must be computed, in order to determine bounds for x, is $2n \times (n!)$. At the expense of obtaining poorer bounds it is easy to reduce this number to $2n^2$. Suppose we compute, at the first step, the $2n$ numbers $\gamma_1^{(1)}, \delta_1^{(1)}, \cdots, \gamma_n^{(1)}, \delta_n^{(1)}$ and set $\lambda_1 = \min \{\gamma_1^{(1)}, \cdots, \gamma_n^{(1)}\}$, $\mu_1 = \max \{\delta_1^{(1)}, \cdots, \delta_n^{(1)}\}$. Then, for each B with $\alpha_i \leqq b_i \leqq \beta_i$ for $i = 1, \cdots, n$, one of the components of the corresponding x will be bounded by λ_1 (from below) and μ_1 (from above). We next compute the $2n$ numbers $\gamma_i^{(2)} = \gamma_i^{(2)}(f_i, \alpha_i - p_i^{(1)})$, $\delta_i^{(2)} = \delta_i^{(2)}(f_i, \beta_i - q_i^{(1)})$, where $p_i^{(1)} = \max \{a_{ij}\lambda_1, a_{ij}\mu_1 : j \neq i\}$, $q_i^{(1)} = \min \{a_{ij}\lambda_1, a_{ij}\mu_1 : j \neq i\}$, and denote the smallest $\gamma_i^{(2)}$ by λ_2 and the largest $\delta_i^{(2)}$ by μ_2. Then we have bounds which apply for two of the components of the x which corresponds to any B with $\alpha_i \leqq b_i \leqq \beta_i$ for $i = 1, \cdots, n$. By computing $\gamma_i^{(3)} = \gamma_i^{(3)}(f_i, \alpha_i - p_i^{(1)} - p_i^{(2)})$, $\delta_i^{(3)} = \delta_i^{(3)}(f_i, \beta_i - q_i^{(1)} - q_i^{(2)})$, etc., the above process may be continued to obtain the numbers $\lambda_1, \cdots, \lambda_n$, $\mu_1, \cdots, \mu_n$. Each component of the x corresponding to any B with $\alpha_i \leqq b_i \leqq \beta_i$ for $i = 1, \cdots, n$ will be bounded by $\lambda = \min \{\lambda_1, \cdots, \lambda_n\}$ (from below) and $\mu = \max \{\mu_1, \cdots, \mu_n\}$ (from above).

A matter that is closely related to the proofs of the above theorems on the boundedness of solutions of equation (1) is that of proving: *For each* $F \in \mathfrak{F}^n$ *and each* $A \in P_0$ *the solution* x *of equation* (1) *is a continuous function of the vector* B. It is obvious that it will suffice to prove that for each $F \in \mathfrak{F}^n$ with $F(\theta) = \theta$, and for each $A \in P_0$, the solution x of equation (1) is continuous in B at $B = \theta$. We then note that if f

satisfies the hypotheses of Lemma 1 and, in addition, if $f(0) = 0$ then, due to the continuity of f^{-1}, for every $\epsilon > 0$ there exists $\zeta > 0$ such that if α, β in Lemma 1 satisfy $-\zeta < \alpha \leqq b \leqq \beta < \zeta$ then γ, δ in Lemma 1 satisfy $-\epsilon < \gamma \leqq x \leqq \delta < \epsilon$. This observation may be used to incorporate a simple "ϵ-δ argument" into the steps of the previous paragraph to show that when $F(\theta) = \theta$ then for arbitrary $\epsilon > 0$, one can determine $\zeta > 0$ such that $\| B \| < \zeta$ implies $\| x \| < \epsilon$.

At this point we return to the matter of the existence and uniqueness of solutions of equation (1). We state first a theorem of R. S. Palais (Ref. 7—see also the Appendix of Ref. 8) which shows the connection between the concepts of existence and uniqueness of solutions and the boundedness of solutions.

Palais' Theorem. Let $f_1, \cdots, f_n$ be n continuously differentiable real valued functions of n real variables. Necessary and sufficient conditions that the mapping $f : E^n \rightarrow E^n$ defined by $f(x) = (f_1(x), \cdots, f_n(x))^t$ be a diffeomorphism of E^n onto itself are:

(*i*) $\det [\partial f_i/\partial x_j]$ *never vanishes.*
(*ii*) $\lim_{\|x\|\rightarrow\infty} \| f(x) \| = \infty$.

Palais' Theorem may be used to prove a result which is almost equivalent to our Theorem 3, that is:

Theorem 6. If A is an $n \times n$ matrix then there exists a unique solution of equation (1) *for each $F \equiv (f_1(x_1), \cdots, f_n(x_n))^t$ with continuously differentiable, strictly monotone increasing functions f_i which map E^1 onto itself, and whose slopes are everywhere positive, and for each $B \varepsilon E^n$, if and only if $A \varepsilon P_0$.*

A proof of Theorem 6 which is independent of our Theorem 3 is easy to construct: For all $A \varepsilon P_0$, the rather trivial Theorem 1 guarantees that condition (*i*) of Palais' Theorem is satisfied, and Theorem 5 guarantees that condition (*ii*) is satisfied. If $A \not\varepsilon P_0$ then a choice of F such as is specified in the "*if*" part of the proof of Theorem 4 provides a case in which condition (*ii*) of Palais' Theorem is violated.

VI. SUFFICIENT CONDITIONS FOR $A \varepsilon P_0$ OR P

For a given matrix A, it is not in general an easy task to determine whether or not A satisfies any one of the four equivalent conditions of Fiedler and Pták which are given in Section III and which serve to define the class of matrices P_0 (or the conditions which define P). This is particularly true when the order of A is large. For this reason, we now give several conditions which are sufficient to insure that a matrix A is in P_0 or P (and which are not so difficult to verify).

Suppose it were known that every eigenvalue of A as well as every eigenvalue of each principal submatrix of A had a nonnegative (positive) real part. Then this would guarantee that $A \varepsilon P_0$ (P). This is the main idea involved in the following theorem.

Theorem 7. If any one of the following inequalities is satisfied by the elements a_{ij} of the matrix A, for all $i = 1, \cdots, n$, then $A \varepsilon P_0$.

$$(i) \qquad a_{ii} \geqq \Big(\sum_{j \neq i} | a_{ij} |\Big)^{\alpha} \Big(\sum_{k \neq i} | a_{ki} |\Big)^{1-\alpha}, \qquad 0 \leqq \alpha \leqq 1;$$

$$(ii) \qquad a_{ii} \geqq \alpha_i^{1/q} \Big(\sum_{j \neq i} | a_{ij} |^{p}\Big)^{1/p}, \qquad p \geqq 1, \qquad p^{-1} + q^{-1} = 1,$$

α_i positive numbers satisfying $\sum_{i=1}^{n} (1 + \alpha_i)^{-1} \leqq 1$;

$$(iii) \qquad a_{ii} \geqq \alpha \max_{j \neq i} | a_{ij} |, \ \alpha \text{ positive satisfying}$$

$$\sum_{i=1}^{n} \Big\{ \sum_{j \neq i} | a_{ij} | \big(\max_{j \neq i} | a_{ij} |\big)^{-1} \Big\} \leqq \alpha(1 + \alpha), \qquad (0/0 = 0).$$

If any one of the above inequalities with $\geqq$ replaced by $>$ is satisfied for $i = 1, \cdots, n$, then $A \varepsilon P$.

Proof: If the right-hand side of any of the above inequalities is denoted by the nonnegative number r_i then it is well known that all of the eigenvalues of the matrix A are contained in the union $\bigcup \{C_i : i = 1, \cdots, n\}$ of the disks $C_i = \{z : | z - a_{ii} | \leqq r_i\}$.[9] But the condition $a_{ii} \geqq (>)\ r_i$ guarantees that if $z \varepsilon C_i$ then $\mathrm{Re}(z) \geqq (>)\ 0$. Thus, each of the eigenvalues of the matrix A has a nonnegative (positive) real part. The same is true of each eigenvalue of every principal submatrix of A, for if one of the above inequalities is satisfied by the elements of A it is also satisfied by the elements of any principal submatrix. □

VII. COMPUTATION OF THE SOLUTION

At present, the authors know of no single computational algorithm which is guaranteed to yield the solution of equation (1) for all $F \varepsilon \mathfrak{F}^n$, $A \varepsilon P_0$, $B \varepsilon E^n$. However, there are several ways that the solution may be computed for large classes of such equations.

If, for example, the matrix A satisfies either a weak row-sum or weak column-sum dominance condition (inequality (*i*) of Theorem 7 with either $\alpha = 1$ or $\alpha = 0$) and if $F \in \mathfrak{F}^n$ with, roughly speaking, the slopes of each f_i bounded from below by some positive constant, then it can be shown (see Appendix A) that an algorithm for computing the solution can be obtained by the use of Banach's contraction-mapping fixed point theorem.

If the matrix A is positive semidefinite then, as mentioned in Section I, the existence of a unique solution of equation (1) for all $F \in \mathfrak{F}^n$ follows from the earlier work of Sandberg and Minty. If, in addition, there exists for $i = 1, \cdots, n$, positive constants α_i and β_i such that

$$\alpha_i \leqq \frac{f_i(u) - f_i(v)}{u - v} \leqq \beta_i$$

for all $u \neq v$, then Sandberg's iteration scheme (also resulting from an application of the contraction-mapping fixed point theorem) can be used to compute the solution.[1] In this regard, if the techniques of Section V are first used to obtain bounds on the location of the solution then one could modify equation (1) by changing the nature of the functions f_i outside the domain in which the solution is known to lie (but still keeping the f_i strictly monotone increasing from E^1 onto E^1) and obtain a new equation which has the same solution as the original equation. By doing this, the functions f_i in the new equation might be made to satisfy the above inequalities in cases where this was impossible for the original f_i. Also, even if these inequalities could be satisfied for the original equation, larger values of α_i and smaller values of β_i might be used for the modified equation. This can result in a more rapidly converging iteration process (see Section VII of Ref. 3). Similarly, the bounds can be used to improve the performance of other iteration schemes.

In case $A \in P_0$ is not positive semidefinite, it might be that there exist diagonal matrices $\Delta_1, \Delta_2 > 0$ such that $\Delta_1 A \Delta_2$ is positive semidefinite. If such matrices can be found, then Sandberg's iteration scheme could be used to compute the solution of the equation

$$\Delta_1 F(\Delta_2 x) + \Delta_1 A \Delta_2 x = \Delta_1 B,$$

from which the solution of equation (1) may be obtained directly. Unfortunately, it is not the case that such $\Delta_1, \Delta_2 > 0$ exist for all $A \in P_0$. For example, it is quite easily verified that for

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix},$$

even though $A \in P_0$, the matrix $\Delta_1 A \Delta_2$ is not positive semidefinite for any choice of $\Delta_1, \Delta_2 > 0$.

It is easily verified, however, that appropriate $\Delta_1, \Delta_2 > 0$ can be found for all 2×2 matrices $A \in P_0$ except those for which

(*i*) $a_{11}a_{22} = 0,$

and

(*ii*) $a_{12}a_{21} = 0,$

and

(*iii*) either $a_{12} \neq 0$ or $a_{21} \neq 0.$

In particular, for all nonsingular 2×2 matrices $A \in P_0$, appropriate Δ_1, Δ_2 ($\Delta_2 = I$) can be found. Thus, Sandberg's iteration scheme could be used, for example, to compute the solution of the example problem of Section II which was taken from Calahan's book.

VIII. APPLICATION TO EQUATIONS FOR TRANSISTOR NETWORKS

In this section some of the above theory is applied to the equations which describe the behavior of electrical networks containing transistors. By the word transistor we refer to the three-terminal device whose equivalent circuit is shown in Fig. 2.* Considering the transistor as a nonlinear two-port network, the following equations which express the port currents in terms of the port voltages follow immediately from inspection of Fig. 2:

$$\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} 1 & -\alpha_{12} \\ -\alpha_{21} & 1 \end{bmatrix} \begin{bmatrix} f_1(v_1) \\ f_2(v_2) \end{bmatrix}.$$

We assume, as is the case for the usual large-signal model of a physical transistor, that $0 < \alpha_{12} < 1$, $0 < \alpha_{21} < 1$, and that both of the functions f_1 and f_2 are continuous and strictly monotone increasing. The character of the functions f_1 and f_2 which describe the transistor's nonlinear conductances will depend on whether the transistor is designated as NPN or PNP. We shall, however, have no occasion to distinguish between these two cases in what is to follow.

Suppose an electrical network is synthesized by connecting together,

* In some respects this equivalent circuit is an *ideal* model of a transistor. Nevertheless, since this model is often used in the design and the computer analysis of transistor networks, consideration of it is important. The presence of series resistance at the base, emitter, and collector terminals of a transistor will be considered by the authors in another paper.

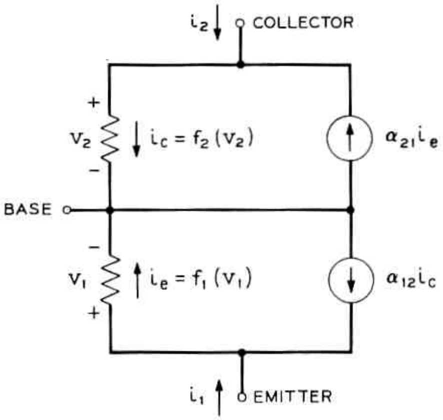


Fig. 2 — The equivalent circuit of a transistor.

in an arbitrary manner, any (finite) number of transistors, resistors (that is, linear resistors with nonnegative resistance), voltage sources, current sources, and nonlinear resistors which are described by strictly monotone increasing conductance functions (and which we shall henceforth refer to as "diodes"). Suppose the network contains n transistors and d diodes. For $k = 1, \cdots, n$, let x_{2k-1}, x_{2k}, y_{2k-1}, and y_{2k} denote the voltage and current variables v_1, v_2, i_1, and i_2, respectively, for the k-th transistor. For $k = 1, \cdots, d$, let x_{2n+k} and y_{2n+k} denote the voltage across, and the current through, the kth diode. Let these variables be related by $y_{2n+k} = f_{2n+k}(x_{2n+k})$. Then, if $x = (x_1, \cdots, x_{2n+d})^t$ and $y = (y_1, \cdots, y_{2n+d})^t$, we have

$$y = TF(x), \tag{9}$$

where $T = \operatorname{diag}(T_1, T_2)$, with T_1 a block diagonal matrix with n 2×2 diagonal blocks of the form

$$\begin{bmatrix} 1 & -\alpha_{12}^{(k)} \\ -\alpha_{21}^{(k)} & 1 \end{bmatrix},$$

and T_2 a $d \times d$ identity matrix. The nonlinear function F has the form $F(x) \equiv (f_1(x_1), \cdots, f_{2n+d}(x_{2n+d}))^t$.

Consider now the $(2n+d)$-port network of resistors and independent sources which is formed from the original network by removing the transistors and diodes. If the y-parameter matrix G of this $(2n+d)$-port exists then we have the additional equation relating the vectors

x and y:

$$y = -Gx + u, \tag{10}$$

where u is some vector of constants which is, in general, nonzero since sources are present in the $(2n+d)$-port.

Combining equations (9) and (10) we obtain

$$TF(x) + Gx = u. \tag{11}$$

Now T is a nonsingular matrix and hence, if equation (11) is multiplied by T^{-1}, we obtain an equation having the form of equation (1). If the matrix $T^{-1}G \varepsilon P_0$ then, by Corollary 1, there exists at most one set of transistor and diode voltages satisfying equation (11). Moreover, if each of the nonlinear functions describing the transistors and diodes in our network maps E^1 *onto* E^1, or if $T^{-1}G \varepsilon P$, then Theorem 3, or Corollary 3, guarantees the *existence* of a unique solution of equation (11).

We have been careful to distinguish between the case when our theory guarantees only the uniqueness of a solution and the case when it guarantees both the solution's existence and its uniqueness for the following reason: In the analysis of transistor networks the nonlinear functions which are used to describe diodes or to describe the nonlinear conductances in the equivalent circuit of a transistor are often taken to be of the form

$$f(x) = I_0(e^{\lambda x} - 1),$$

where I_0 and λ are constants. The range of such a function is not the entire real line. Presumably, therefore, one can construct transistor networks having the property that if functions of the above type are used in a transistor's equivalent circuit then the network admits no solution. We now give a simple example of such a network. We wish to emphasize, though, that even for these networks whose equations may sometimes have no solution, our theory still guarantees that if $T^{-1}G \varepsilon P_0$ and if a solution of equation (11) exists, then it is unique.

Consider the network of Fig. 3. For this network, equation (11) becomes

$$\begin{bmatrix} 1 & -\alpha_{12} \\ -\alpha_{21} & 1 \end{bmatrix} \begin{bmatrix} f_1(v_1) \\ f_2(v_2) \end{bmatrix} + \begin{bmatrix} g & -g \\ -g & g \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} I_a \\ I_b \end{bmatrix}.$$

Suppose $\alpha_{12} = 0.5$, $\alpha_{21} = 0.9$, and $g = 5.5$ mhos. Then, the above equation is equivalent to

$$\begin{bmatrix} f_1(v_1) \\ f_2(v_2) \end{bmatrix} + \begin{bmatrix} 5 & -5 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \frac{1}{11} \begin{bmatrix} 20 & 10 \\ 18 & 20 \end{bmatrix} \begin{bmatrix} I_a \\ I_b \end{bmatrix}.$$

Hence, v_1 and v_2 must satisfy

$$f_1(v_1) + 5f_2(v_2) = 10(I_a + I_b).$$

If we now assume that the transistor's nonlinear conductances are described by the functions

$$f_1(v_1) = -I_e(e^{-\lambda_e v_1} - 1),$$
$$f_2(v_2) = -I_c(e^{-\lambda_c v_2} - 1),$$

where the parameters I_e, I_c, λ_e, and λ_c are each positive, then for all v_1, v_2 we have

$$f_1(v_1) + 5f_2(v_2) < I_e + 5I_c .$$

Hence, if the values of the independent current sources of Fig. 3 are chosen such that

$$I_a + I_b \geqq \tfrac{1}{10}I_e + \tfrac{1}{2}I_c ,$$

then the equation for this network has no solution.

Let us now consider the problem of determining whether or not, for a given network, the matrices T and G in equation (11) satisfy the condition $T^{-1}G \varepsilon P_0$ (or $T^{-1}G \varepsilon P$). (The existence of many transistor bistable circuits assures us that this condition is not always satisfied.)

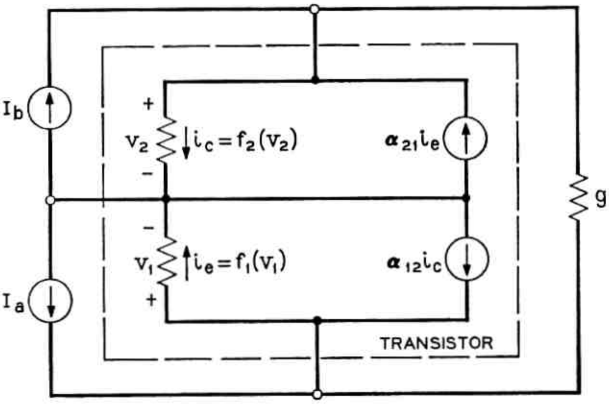


Fig. 3 — A transistor network whose equations may have no solution.

There is a large class of networks for which this condition is satisfied, and for which a simple inspection of the G matrix suffices to identify a member of the class.

Since the matrix T satisfies a strong column-sum dominance condition, that is, since

$$t_{ii} > \sum_{j \neq i} | t_{ji} | \qquad \text{for} \quad i = 1, \cdots , 2n + d,$$

the following theorem guarantees that if the matrix G also satisfies a strong column-sum dominance condition, then $T^{-1}G \varepsilon P$, and that if the matrix G satisfies a weak column-sum dominance condition,

$$g_{ii} \geqq \sum_{j \neq i} | g_{ji} |,$$

then $T^{-1}G \varepsilon P_0$ and, hence, the above conclusions concerning the existence and the uniqueness of a solution follow.

Theorem 8. If the square matrix A satisfies a strong column-sum dominance condition and if the square matrix B satisfies a weak (strong) column-sum dominance condition, then $A^{-1}B \varepsilon P_0$ (P).

Proof: Suppose $A^{-1}B \not\varepsilon P_0$. Then, by the main result of the "only if" part of the proof of Theorem 3, there exists some diagonal matrix $D > 0$ such that $\det (D + A^{-1}B) = 0$. But $\det (D + A^{-1}B) = \det (A^{-1}) \cdot \det (AD + B)$, and $\det (A^{-1}) \neq 0$. Likewise, $\det (AD + B) \neq 0$ since $AD + B$ satisfies a strong column-sum dominance condition. Hence, $A^{-1}B \varepsilon P_0$.

With B strongly column-sum dominant, let $\delta > 0$ be such that $B - \delta A$ also possesses the strong dominance property. Suppose that $A^{-1}B - \delta I \not\varepsilon P_0$. Then, as above, there is a $D > 0$ such that $A^{-1}B - \delta I + D = A^{-1}[B - \delta A + AD]$ is singular, which is a contradiction. Therefore $A^{-1}B - \delta I \varepsilon P_0$, and, by Theorem 1, $A^{-1}B \varepsilon P$. □

IX. COMMON-BASE TRANSISTOR NETWORKS

We now consider a special class of the networks which are comprised of transistors, resistors, diodes, and independent sources. We consider the class of all such networks for which there is a single node (called ground) to which the base terminal of each transistor is connected. Let us first consider a subclass of this class of networks; that is, let us temporarily assume that no diodes are present. For all networks in this subclass it is easily verified that when the G matrix for equation (11) exists, then it satisfies the above weak column-sum dominance condition and hence, by Theorem 8, $T^{-1}G \varepsilon P_0$. This fact is made evident if we consider the network of resistors which is described by G (that is, the linear multiport to which the transistors are connected, with all sources removed) and first simplify this network by using the star-mesh transformation to remove all internal nodes. Of course for many networks of this sublcass G is strongly column-sum dominant, in which case $T^{-1}G \varepsilon P$.

It is clear that the networks for which the G matrix fails to exist are exactly those networks in which either one or more of the collector or emitter terminals are connected, through the resistor network, directly to ground (that is, through a branch having infinite conductance), or else two (or more) of the transistors' collector or emitter terminals are connected directly together (through a branch of the resistor network having infinite conductance). These direct connections can exist in the resistor network either because of corresponding short-circuits in the original linear multiport, or because of corresponding connections involving branches which contain only ideal voltage sources.

If one assumes that each transistor in the network has a nonzero series resistance associated with both its emitter and its collector terminals (this assumption certainly being consistent with physical reality) then one need not be concerned about the possibility of the nonexistence of the G matrix since the situations mentioned in the previous paragraph cannot occur. We now show, however, that one need not rely upon this assumption in order to prove the uniqueness of the solution of the equations which describe the networks that we are considering.

We have observed that the matrix G will not exist if and only if the linear multiport has fewer independent port voltages than it has ports. In this case we modify the nonlinear multiport in such a manner that we can break some of the connections to the linear multiport so that it then possesses a G matrix and hence can be described by an equation having the form of equation (10). The modifications to the nonlinear multiport which are called for are obviously the addition of voltage sources between certain nodes, the values of these sources being the same as those of the voltage sources connecting the corresponding nodes in the linear multiport. This simple concept is illustrated in Fig. 4. Here, the network of Fig. 4a, containing a linear 6-port, has been replaced by the "equivalent" network of Fig. 4b containing a linear 3-port. Although the G matrix of the 6-port does not

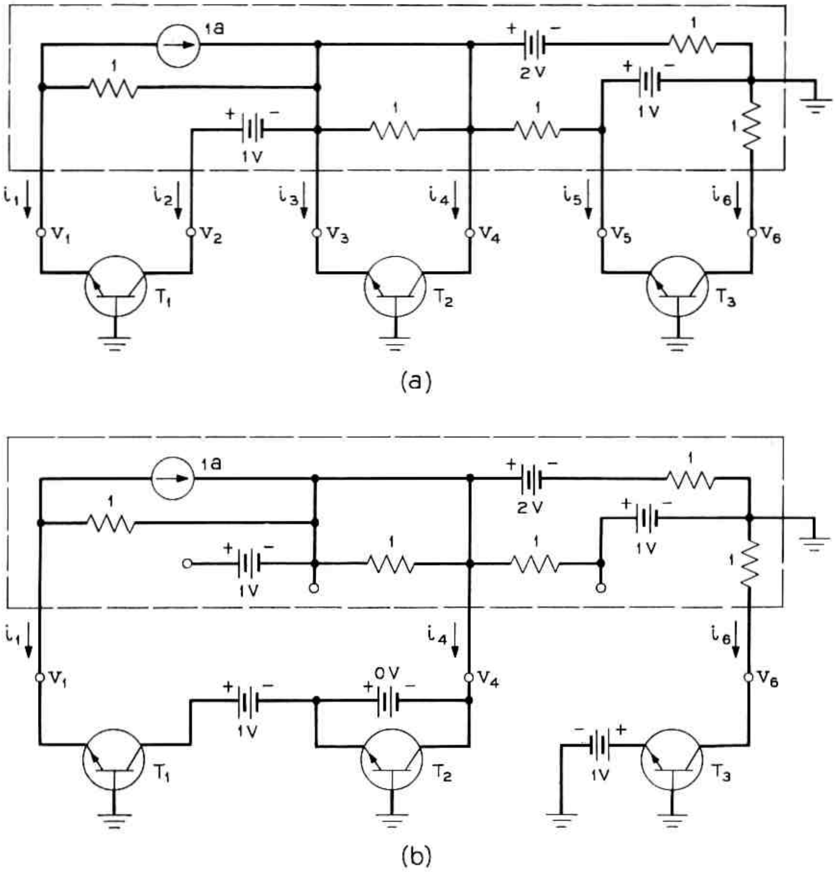


Fig. 4 — Example of a grounded-base transistor network.

exist, it does exist for the 3-port which can be described by

$$\begin{bmatrix} i_1 \\ i_4 \\ i_6 \end{bmatrix} = -\begin{bmatrix} 1 & -1 & 0 \\ -1 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_4 \\ v_6 \end{bmatrix} + \begin{bmatrix} -1 \\ 4 \\ 0 \end{bmatrix}.$$

We have shown that the above artifice allows an equation having the form of equation (10) to always be written to describe the linear multiport contained in our network. We now show that an equation like equation (9) can be written to describe the nonlinear part of our modified network. The equation which we obtain is of the form $y = PTF(P^t x + C)$ with P an $m \times 2n$ matrix $(m < 2n)$ and C a $2n$-vector.

Consider the equation which describes the nonlinear part of a common-base transistor network before any of the above-mentioned modifications (that is, the addition of voltage sources) are made. This equation has the form of equation (9) with T being a $2n \times 2n$ block diagonal matrix (recall that n is the number of transistors present). Let us consider the effect on this equation of the modification of the network by adding voltage sources, one at a time. There are two different ways of adding voltage sources that must be considered.

Suppose a voltage source of voltage E is connected between nodes j and k (with plus reference at node j), and suppose the connections between node j and the linear multiport are then open-circuited. This situation is illustrated in Fig. 5. Using the notation indicated in this figure, we have

$$\begin{aligned} i' &= TF(v), \\ i_\nu &= i'_\nu \qquad \text{for} \quad \nu \neq j, k, \\ i_j &= 0, \\ i_k &= i'_j + i'_k\,, \\ v_j &= v_k + E. \end{aligned}$$

Let us now define the vectors v^* and i^* to be the $(2n-1)$-vectors obtained from v and i, respectively, by deleting the v_j and i_j elements. Then, if $F^*(v^*)$ is the $2n$-vector obtained from $F(v)$ by replacing the

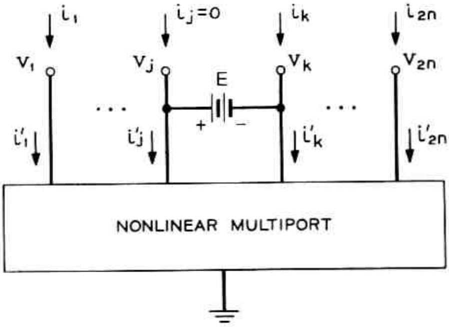


Fig. 5 — Typical modification of the nonlinear multiport network.

argument v_j by $v_k + E$, we then have that

$$i^* = T^*F^*(v^*), \tag{12}$$

where the $(2n-1) \times 2n$ matrix T^* is obtained from T by adding the j-th row to the k-th row and then deleting the j-th row. Observe that $T^*F^*(v^*)$ can be written as $QTF(Q^tv^* + R)$ in which the j-th element of the $2n$-vector R is E, all other elements of R are zero, and Q is obtained from the identity matrix of order $2n$ by adding the j-th row to the k-th row and then deleting the j-th row.

In case a voltage source of voltage E is connected between node j and ground (with the plus reference at node j) and all connections between node j and the linear multiport are open-circuited, then we can again form equation (12) from equation (9) by simply replacing v_j by E wherever it appears in the argument of F, to form $F^*(v^*)$, and deleting the j-th row of the matrix T, to form T^*. In this case $T^*F^*(v^*)$ can be written as $QTF(Q^tv^* + R)$ in which R is as defined earlier, but in this case Q is obtained from the identity matrix of order $2n$ by simply deleting the j-th row.

The above processes can be applied repeatedly to account for the addition of an arbitrary number of voltage sources to the nonlinear multiport. The resulting equation which describes the multiport will have the form

$$\begin{aligned} y &= Q_p \cdots Q_2Q_1TF(Q_1^tQ_2^t \cdots Q_p^tx + C) \\ &\equiv \tilde{T}\tilde{F}(x) \end{aligned}$$

with C some constant $2n$-vector and each of the Q_i obtained from the identity matrix of the appropriate order in one of the two ways described above.

Consider equation (9) in which T is a square matrix. Due to the strict monotonicity of each component function of F, the mapping $TF(x)$ has the following property: If p, q are arbitrary $2n$-vectors then there is a diagonal matrix $D > 0$ such that

$$TF(p) - TF(q) = TD(p - q), \tag{13}$$

and furthermore, the matrix TD is strongly column-sum dominant (since T is strongly column-sum dominant). We now wish to show that a similar fact is true in the more general case.

With m the number of rows of Q_p, let p and q denote arbitrary m-vectors. Then since there is a diagonal $D > 0$ such that

$$F(Q_1^tQ_2^t \cdots Q_p^tp + C) - F(Q_1^tQ_2^t \cdots Q_p^tq + C) = DQ_1^tQ_2^t \cdots Q_p^t(p - q),$$

we have

$$\tilde{T}\tilde{F}(p) - \tilde{T}\tilde{F}(q) = Q_p \cdots Q_2Q_1TDQ_1^tQ_2^t \cdots Q_p^t(p - q).$$

The fact that $Q_p \cdots Q_2Q_1TDQ_1^tQ_2^t \cdots Q_p^t$ is strongly column-sum dominant follows from the very easily verified proposition that the product $Q_kMQ_k^t$ $(k = 1, 2, \cdots, p)$ possesses that property whenever M does.

Therefore if x^1 and x^2 denote two solutions of the "generalized equation (11)," then $[\tilde{T}\tilde{D} + G](x^1 - x^2) = \theta$ in which

$$\tilde{T} = Q_p \cdots Q_2Q_1T \qquad \text{and} \qquad \tilde{D} = DQ_1^tQ_2^t \cdots Q_p^t .$$

But $\tilde{T}\tilde{D}$, and hence $\tilde{T}\tilde{D} + G$, is strongly column-sum dominant and hence, nonsingular. This implies that $x^1 = x^2$.

We have now shown that in any network constructed from resistors, independent sources, and transistors having a common-base connection, the transistors' base-emitter and base-collector voltages are unique. It is a trivial matter to show that the same result applies when diodes are also allowed to be present in the network.

Suppose the result was not true for some network containing at least one diode. Then there would be two different sets of voltages and currents which satisfy Kirchoff's laws. Thus for each diode in the network there would be two (not necessarily distinct) pairs of points $(v_d^{(1)}, i_d^{(1)})$, $(v_d^{(2)}, i_d^{(2)})$ at which the diode is biased, corresponding to each solution. Letting f denote the strictly monotone increasing function which characterizes the diode we have $i_d^{(1)} = f(v_d^{(1)})$ and $i_d^{(2)} = f(v_d^{(2)})$. But then, suppose the diode is replaced by the series combination of a resistor r and a voltage source E whose values are chosen so that the line $i_d = (1/r)v_d - E/r$ passes through the points $(v_d^{(1)}, i_d^{(1)})$ and $(v_d^{(2)}, i_d^{(2)})$. (Due to the strict monotonicity of f, this can certainly be done with some positive choice of r.) Performing the above type of substitution for each diode in the network, we obtain a new network of the type already considered. This new network would possess two different sets of transistor base-emitter and base-collector voltages (the same as before). This contradicts our previous result, and hence the previous result must apply, even when diodes are present in the network.

To determine the equilibrium solutions of the differential equations which describe a network containing inductors and capacitors as well as the elements mentioned above, one must determine the solutions of a dc equation for a network of the above class. Thus, in summary, what we have shown is: *One cannot synthesize a bistable network which consists of resistors, inductors, capacitors, diodes, independent* *voltage and current sources, and an arbitrary number of (Fig. 2) transistors having a common base connection (or, in particular, only one transistor).*

X. ACKNOWLEDGMENT

The authors would like to acknowledge the helpful conversations with their colleague H. C. So.

APPENDIX A

Algorithms for Computing Solutions of Equation (1)

In this appendix two algorithms for computing the solution of equation (1) are presented. It is proved that one of the algorithms will always converge to the solution of equation (1) if the matrix A satisfies either a weak row-sum or column-sum dominance condition (inequality (i) of Theorem 7 with either $\alpha = 1$ or $\alpha = 0$) and if, roughly speaking, the slopes of each f_i are bounded from below by some positive constant. In each case the proof of convergence relies upon Banach's contraction-mapping fixed point theorem, and therefore also represents an independent proof of the existence and uniqueness of a solution of equation (1) for the conditions stated above.

The following notation will be used: For fixed $F \in \mathfrak{F}^n$, $B \in E^n$, let $f(x) \equiv F(x) - B$; also, if A is a given $n \times n$ matrix with elements a_{ij}, we define the diagonal matrix D by $D = \text{diag}\,[a_{11}, a_{22}, \cdots, a_{nn}]$, and let $\Delta = A - D$.

Theorem A. If the $n \times n$ matrix A satisfies

$$a_{ii} \geqq \sum_{j \neq i} | a_{ij} |, \qquad \text{for} \quad i = 1, \cdots, n,$$

and if $F \in \mathfrak{F}^n$, $B \in E^n$, and if there exists some $\epsilon > 0$ such that for each $\alpha, \beta \in E^1$, $\epsilon \,| \alpha - \beta | \leqq | f_i(\alpha) - f_i(\beta) |$ for $i = 1, \cdots, n$, then equation (1) possesses a unique solution, and if x^0 is an arbitrary point in E^n, the sequence $x^0, x^1, x^2, \cdots$ defined by

$$x^{k+1} = (f + D)^{-1}(-\Delta)x^k$$

converges to the solution.

Proof: Equation (1) may be rewritten as

$$f(x) + Dx + \Delta x = \theta.$$

Hence, if the operation $T: E^n \rightarrow E^n$ is defined by $T = (f + D)^{-1}(-\Delta)$, then the solution of the equation $x = Tx$ is identical to the solution of equation (1). We now prove that the sequence $x^0, x^1, x^2, \cdots$ converges to this solution by proving that T is a contraction.

Let x and y be arbitrary points in E^n and let $g = Tx$, $h = Ty$. Then, $f(g) + Dg = -\Delta x$ and $f(h) + Dh = -\Delta y$. Thus, for $i = 1, \cdots, n$,

$$f_i(g_i) - b_i + a_{ii}g_i = -(\Delta x)_i ,$$

and

$$f_i(h_i) - b_i + a_{ii}h_i = -(\Delta y)_i .$$

Subtracting, we obtain

$$f_i(g_i) - f_i(h_i) + a_{ii}(g_i - h_i) = (\Delta y)_i - (\Delta x)_i .$$

Since f_i is strictly monotone increasing, we have

$$| f_i(g_i) - f_i(h_i) | + a_{ii} | g_i - h_i | = | (\Delta x)_i - (\Delta y)_i |,$$

and hence, since $\epsilon + a_{ii} > 0$,

$$| g_i - h_i | \leqq \frac{1}{a_{ii} + \epsilon} | (\Delta x)_i - (\Delta y)_i |.$$

Now,

$$\begin{aligned} | (\Delta x)_i - (\Delta y)_i | &= | \sum_{j \neq i} a_{ij}(x_j - y_j) | \\ &\leqq \sum_{j \neq i} (| a_{ij} | \cdot | x_j - y_j |) \\ &\leqq (\sum_{j \neq i} | a_{ij} |) \cdot \max_j | x_j - y_j |. \end{aligned}$$

Thus, defining the metric ρ on E^n by $\rho(x, y) = \max_j | x_j - y_j |$, we have, for $i = 1, \cdots, n$,

$$| g_i - h_i | \leqq \frac{1}{a_{ii} + \epsilon} (\sum_{j \neq i} | a_{ij} |) \cdot \rho(x, y).$$

But, since $0 \leqq \sum_{j \neq i} | a_{ij} | < a_{ii} + \epsilon$, there exists K, $0 \leqq K < 1$, such that $| g_i - h_i | \leqq K \cdot \rho(x, y)$ for $i = 1, \cdots, n$, and in particular, $\rho(Tx, Ty) = \max_i | g_i - h_i | \leqq K \cdot \rho(x, y)$. Hence T is a contraction. □

Theorem B. If the $n \times n$ *matrix* A *satisfies*

$$a_{ii} \geqq \sum_{j \neq i} | a_{ji} |, \qquad \text{for} \quad i = 1, \cdots, n,$$

and if $F \in \mathfrak{F}^n$, $B \in E^n$, *and if there exists some* $\epsilon > 0$ *such that for each*

$\alpha, \beta \in E^1$, $\epsilon \mid \alpha - \beta \mid \leqq \mid f_i(\alpha) - f_i(\beta) \mid$ for $i = 1, \cdots, n$, then equation (1) possesses a unique solution, and if z^0 is an arbitrary point in E^n, the sequence $z^0, z^1, z^2, \cdots$ defined by

$$z^{k+1} = -\Delta(f + D)^{-1}z^k$$

converges to some point z^* and the solution of equation (1) *is given by*

$$x^* = (f + D)^{-1}z^*.$$

Proof: As in Theorem A, the solution of equation (1) is also the solution of $x = (f + D)^{-1}(-\Delta)x$. For each $x \in E^n$, let $z = (f + D)x$ and hence $x = (f + D)^{-1}z$. Thus, x^* is the solution of equation (1) if $x^* = (f + D)^{-1}z^*$, where z^* is the solution of $z = -\Delta(f + D)^{-1}z$. The theorem is thus proved if it is proved that the operator $T \equiv -\Delta(f + D)^{-1}$ is a contraction.

Let P denote the operator $(f + D)^{-1}$, and let x and y be arbitrary points in E^n. Then, proceeding as in the proof of Theorem A, we obtain

$$\mid (Px)_j - (Py)_j \mid \leqq \frac{1}{a_{jj} + \epsilon} \mid x_j - y_j \mid, \qquad \text{for} \quad j = 1, \cdots, n.$$

Thus, if $g = Tx$ and $h = Ty$, then for $i = 1, \cdots, n$,

$$g_i = -\sum_{j \neq i} a_{ij}(Px)_j \qquad \text{and} \qquad h_i = -\sum_{j \neq i} a_{ij}(Py)_j \ .$$

Hence

$$\begin{aligned} \mid g_i - h_i \mid &= \mid \sum_{j \neq i} a_{ij}((Px)_j - (Py)_j) \mid \\ &\leqq \sum_{j \neq i} (\mid a_{ij} \mid \cdot \mid (Px)_j - (Py)_j \mid) \\ &\leqq \sum_{j \neq i} \left(\mid a_{ij} \mid \cdot \frac{1}{a_{jj} + \epsilon} \cdot \mid x_j - y_j \mid \right) \cdot \end{aligned}$$

Therefore,

$$\begin{aligned} \sum_{i=1}^{n} \mid g_i - h_i \mid &\leqq \sum_{i=1}^{n} \sum_{j \neq i} \frac{\mid a_{ij} \mid}{a_{jj} + \epsilon} \mid x_j - y_j \mid \\ &= \sum_{j=1}^{n} \left(\sum_{i \neq j} \frac{\mid a_{ij} \mid}{a_{jj} + \epsilon} \right) \mid x_j - y_j \mid. \end{aligned}$$

But, there exists K, $0 \leqq K < 1$, such that, for $j = 1, \cdots, n$,

$$\sum_{i \neq j} \frac{\mid a_{ij} \mid}{a_{jj} + \epsilon} \leqq K,$$

and hence

$$\sum_{i=1}^{n} | g_i - h_i | \leqq K \sum_{j=1}^{n} | x_j - y_j |.$$

Defining the metric ρ on E^n by

$$\rho(x, y) = \sum_{i=1}^{n} | x_i - y_i |,$$

we therefore have

$$\rho(Tx, Ty) = \sum_{i=1}^{n} | g_i - h_i | \leqq K \cdot \rho(x, y),$$

and hence T is a contraction. □

APPENDIX B

Determination of Bounds on the Solution of Equation (1)

In this appendix we present a method for determining bounds on the solution of equation (1) when $F \varepsilon \mathfrak{F}^n$, A is weakly row-sum dominant, and (for given $\alpha \equiv (\alpha_1, \cdots, \alpha_n)^t$, $\beta \equiv (\beta_1, \cdots, \beta_n)^t \varepsilon E^n$) $B \equiv (b_1, \cdots, b_n)^t$ satisfies $\alpha_i \leqq b_i \leqq \beta_i$ for $i = 1, \cdots, n$. The solution bounds are, in general, easier to compute than those of Theorem 5. The method presented here is a generalization of an idea presented in Ref. 3.

The computation of the solution bounds proceeds in two steps. First, one solves each of the equations

$$F(x) = \alpha \tag{14a}$$

and

$$F(x) = \beta. \tag{14b}$$

Denoting the solutions of equations (14a) and (14b) by $\mu \equiv (\mu_1, \cdots, \mu_n)^t$ and $\nu \equiv (\nu_1, \cdots, \nu_n)^t$, respectively, and defining

$$\lambda = \max\{| \mu_1 |, \cdots, | \mu_n |, | \nu_1 |, \cdots, | \nu_n |\},$$

and

$$B' = (\sum_{j \neq 1} | a_{1j} |, \cdots, \sum_{j \neq n} | a_{nj} |)^t,$$

one then solves each of the equations

$$F(x) + \text{diag}\,[a_{11}, \cdots, a_{nn}]x = \alpha - \lambda B', \tag{15a}$$

$$F(x) + \text{diag}\,[a_{11}, \cdots, a_{nn}]x = \beta + \lambda B'. \tag{15b}$$

Denoting the solutions of equations (15a) and (15b) by $\gamma \equiv (\gamma_1, \cdots, \gamma_n)^t$ and $\delta \equiv (\delta_1, \cdots, \delta_n)^t$, respectively, one has $\gamma_i \leqq x_i^0 \leqq \delta_i$ for $i = 1, \cdots, n$, where x^0 is the solution of equation (1) that corresponds to any B satisfying $\alpha_i \leqq b_i \leqq \beta_i$ for $i = 1, \cdots, n$.

To prove that the components of the vectors γ and δ, determined by the above procedure, are indeed bounds for the corresponding components of the solution x^0 involves no more than a word-for-word repetition of the proof of Theorem 2 of Ref. 3, with several quite obvious modifications. We omit the details.

REFERENCES

1. Sandberg, I. W., "On the Properties of Some Systems that Distort Signals-I," B.S.T.J., *42,* No. 5 (September 1963), pp. 2033–2046.
2. Minty, G. J., "Two Theorems on Nonlinear Functional Equations in Hilbert Space," Bull. Amer. Math. Soc., *69,* No. 5 (September 1963), pp. 691–692.
3. Willson, Jr., A. N., "On the Solutions of Equations for Nonlinear Resistive Networks," B.S.T.J., *47,* No. 8 (October 1968), pp. 1755–1773.
4. Calahan, D. A., *Computer-Aided Network Design* (Preliminary Ed.), New York: McGraw-Hill, 1968.
5. Fiedler, M. and Pták, V., "On Matrices with Non-Positive Off-Diagonal Elements and Positive Principal Minors," Czech. Math. J., *12,* No. 3 (1962), pp. 382–400.
6. Fiedler, M. and Pták, V., "Some Generalizations of Positive Definiteness and Monotonicity," Numer. Math., *9,* No. 2 (1966), pp. 163–172.
7. Palais, R. S., "Natural Operations on Differential Forms," Trans. Amer. Math. Soc., *92,* No. 1 (1959), pp. 125–141.
8. Holzmann, C. A. and Liu, R., "On the Dynamical Equations of Nonlinear Networks with n-Coupled Elements," Proc. Third Ann. Allerton Conf. on Circuit and System Theory., (U. of Illinois, 1965), pp. 536–545.
9. Marcus, M., *Basic Theorems in Matrix Theory,* Washington, D. C., National Bureau of Standards Applied Mathematics Series, *57* (1960).

Some Theorems on the Dynamic Response of Nonlinear Transistor Networks

By I. W. SANDBERG

(Manuscript received July 16, 1968)

Relative to the huge body of theory of linear time-invariant systems, very little of a general and precise nature is known about the network-theoretic properties of transistor circuits operating under large-signal conditions. One basic property P which a transitor network might have is that if the input approaches a constant, then the output approaches a constant which is independent of the initial conditions. In this paper we prove a stability theorem concerning a nonlinear differential equation that governs the behavior of a large class of networks. A corollary of this theorem asserts that if a certain condition is satisfied, then property P holds.

We consider also the problem of estimating the rate of decay of transients in transistor networks and we prove theorems which allow us to make some often quite conservative, but definite, statements concerning limitations on switching speeds. A practical example considered shows that in some cases the bounds, which are frequently very easy to evaluate, can be quite useful.

The proofs depend in an interesting way on the relationship between the static diode characteristic and the nonlinear capacitance associated with a semiconductor junction.

I. INTRODUCTION AND DERIVATION OF THE DIFFERENTIAL EQUATION

We initially consider the network of Fig. 1, which contains transistors, linear resistors, voltage sources, and current sources. Each transistor is represented by a model of the type shown in Fig. 2 (see Gummel[1] and Koehler[2]) which takes into account nonlinear dc properties as well as the presence of nonlinear junction capacitances. Associated with this model are six parameters: α_f, α_r, τ_e, τ_c, c_e, and c_c (all positive constants; $\alpha_f < 1$, $\alpha_r < 1$) and two nonlinear functions $f_e(\cdot)$ and $f_c(\cdot)$.

Concerning $f_e(\cdot)$ and $f_c(\cdot)$, for our purposes it is necessary to assume only that

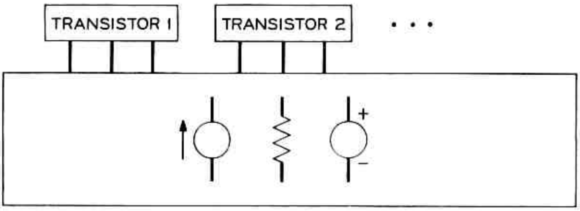


Fig. 1 — General network containing transistors, sources, and resistors.

Assumption 1: For each transistor: $f_e(\cdot)$ and $f_c(\cdot)$ are strictly-monotone increasing mappings of the real interval $(-\infty, \infty)$ into itself; $f_e(0) = f_c(0) = 0$, and $f_e(\cdot)$ and $f_c(\cdot)$ are continuously differentiable on $(-\infty, \infty)$.

The functions $f_e(\cdot)$ and $f_c(\cdot)$ of Gummel's model[1] are of simple exponential type and satisfy Assumption 1.

From Fig. 2:

$$i_e = \frac{d}{dt}\left[c_e v_e + \tau_e f_e(v_e)\right] + f_e(v_e) - \alpha_r f_c(v_c),$$

$$i_c = \frac{d}{dt}\left[c_c v_c + \tau_c f_c(v_c)\right] - \alpha_f f_e(v_e) + f_c(v_c).$$

Suppose that the network of Fig. 1 contains p transistors; for $k = 1, 2, \cdots, p$, let v_{2k-1} and v_{2k}, respectively, denote the emitter to base voltage and the collector to base voltage of the kth transistor. Simi-

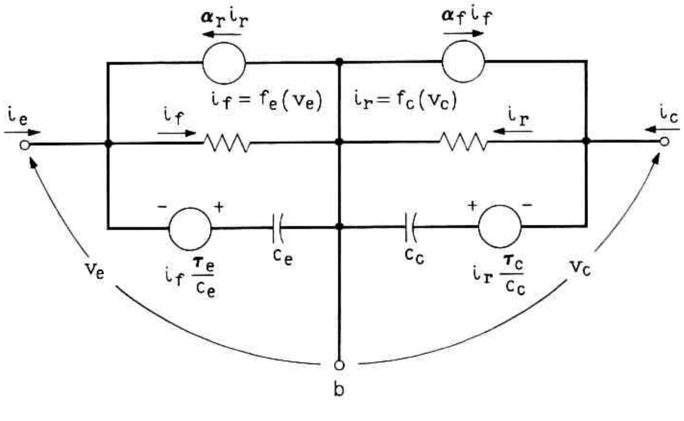


Fig. 2 — Transistor model.

larly, for $k = 1, 2, \cdots, p$, let i_{2k-1} and i_{2k}, respectively, denote the emitter current and the collector current of the kth transistor (with reference polarities as indicated in Fig. 2). Then, with $v = (v_1, v_2, \cdots, v_{2p})^{tr}$, $i = (i_1, i_2, \cdots, i_{2p})^{tr}$, $f_{2k-1}(\cdot)$ and c_{2k-1} the $f_e(\cdot)$ and c_e of the kth transistor, and $f_{2k}(\cdot)$ and c_{2k} the $f_c(\cdot)$ and c_c of the kth transistor,

$$i = \frac{d}{dt}[C(v)] + TF(v) \tag{1}$$

where, for $j = 1, 2, \ldots, 2p$,

$$[C(v)]_j = c_j v_j + \tau_j f_j(v_j) \tag{2}$$

$$[F(v)]_j = f_j(v_j), \tag{3}$$

and $T = T_1 \oplus T_2 \oplus \cdots \oplus T_p$, the direct sum of p 2×2 matrices T_k in which

$$T_k = \begin{bmatrix} 1 & -\alpha_r^{(k)} \\ -\alpha_f^{(k)} & 1 \end{bmatrix}$$

for $k = 1, 2, \cdots, p$.

We assume that the linear resistive portion of the structure of Fig. 1 introduces the constraint

$$i = -Gv + B \tag{4}$$

in which G is a conductance matrix and B is an element of the set $\mathcal{B}$ of all real bounded continuous $2p$-vector-valued functions of t on $[0, \infty)$.

From equations (1) and (4)

$$\frac{d}{dt}[C(v)] + TF(v) + Gv = B. \tag{5}$$

Let $u = C(v)$. Since all of the c_j and τ_j are positive, and each of the $f_j(\cdot)$ is continuous and monotone increasing, there exists a $C^{-1}(\cdot)$ such that $v = C^{-1}(u)$. Thus,

$$\frac{du}{dt} + TF[C^{-1}(u)] + GC^{-1}(u) = B. \tag{6}$$

The Jacobian matrix J_u of $TF[C^{-1}(u)] + GC^{-1}(u)$ is

$$T \operatorname{diag}\left\{\frac{f_j'[g_j(u_j)]}{c_j + \tau_j f_j'[g_j(u_j)]}\right\} + G \operatorname{diag}\left\{\frac{1}{c_j + \tau_j f_j'[g_j(u_j)]}\right\}$$

in which for all $j = 1, 2, \cdots, 2p$

$$g_j(u_j) = [C^{-1}(u)]_j$$

with each of the $g_j(\cdot)$ continuously differentiable.

Since J_u is continuously dependent on u, and $|| J_u ||$ ($|| \cdot ||$ any norm) is bounded from above uniformly in u, it follows that there exists a constant L such that

$$|| TF[C^{-1}(u_a)] + GC^{-1}(u_a) - TF[C^{-1}(u_b)] - GC^{-1}(u_b) || \leqq L || u_a - u_b || \qquad (7)$$

for all u_a and u_b belonging to real Euclidean $2p$-space E^{2p}. In particular, we have

$$|| TF[C^{-1}(u)] + GC^{-1}(u) - B || \leqq L || u || + ||B|| \qquad (8)$$

for all $t \geqq 0$ and all $u \in E^{2p}$. Therefore (see, for example, Nemytskii and Stepanov[3]), for any initial condition $u_0 \in E^{2p}$, there exists a unique continuous $2p$-vector-valued function $u(\cdot)$ such that $u(0) = u_0$ and equation (6) is satisfied for all $t > 0$. In other words, under the assumptions we have introduced, it makes sense to study the properties of the solution of the equation

$$\frac{du}{dt} + TF[C^{-1}(u)] + G[C^{-1}(u)] = B, \qquad t \geqq 0 \quad [u(0) = u_0] \qquad (9)$$

II. STATEMENT OF RESULTS, AND EXAMPLES

We need the following definitions.

Definition 1: A real matrix M of arbitrary order n is *strongly column-sum dominant* if and only if for all $j = 1, 2, \ldots, n$

$$m_{jj} - \sum_{i \neq j} | m_{ij} | > 0.$$

An important property of T is that it is strongly column-sum dominant.

Definition 2: We shall say that a real matrix M of order $2p$ is an element of $\mathfrak{D}$ if and only if there exists a diagonal matrix diag $(d_1, d_2, \cdots, d_{2p})$ with each $d_j > 0$ such that

$$\alpha_f^{(k)} < \frac{d_{2k-1}}{d_{2k}} < \frac{1}{\alpha_r^{(k)}}$$

for $k = 1, 2, \cdots, p$, and diag $(d_1, d_2, \cdots, d_{2p})$ M is strongly column-sum dominant.

Our main result* concerning equation (9) is:

* Proofs of all results in this section are given in Section III.

Theorem 1: If $G \varepsilon \mathfrak{D}$, and $u_a(\cdot)$ and $u_b(\cdot)$ satisfy

$$\frac{du_a}{dt} + TF[C^{-1}(u_a)] + G[C^{-1}(u_a)] = B_a(t), \qquad t \geqq 0 \tag{10}$$

$$\frac{du_b}{dt} + TF[C^{-1}(u_b)] + G[C^{-1}(u_b)] = B_b(t), \qquad t \geqq 0 \tag{11}$$

with $B_a \varepsilon \mathcal{B}$ and $B_b \varepsilon \mathcal{B}$, and if $[B_a(t) - B_b(t)] \rightarrow \theta$ (the zero vector of E^{2p}) as $t \rightarrow \infty$, then $[u_a(t) - u_b(t)] \rightarrow \theta$ as $t \rightarrow \infty$.

An interesting corollary of Theorem 1 is

Corollary 1: Referring to equation (9), if $G \varepsilon \mathfrak{D}$, and if there exists a constant vector B_∞ such that $[B(t) - B_\infty] \rightarrow \theta$ as $t \rightarrow \infty$, then there exists a constant vector u_∞ such that $[u(t) - u_\infty] \rightarrow \theta$ as $t \rightarrow \infty$, and u_∞ is independent of the initial condition u_0. In particular, if $B_\infty = \theta$, then $u_\infty = \theta$.

It is interesting to observe that $G \varepsilon \mathfrak{D}$ whenever the base leads of all transistors are connected together and there is a resistor between the emitter and base, and between the collector and base, of every transistor, for then G is strongly column-sum dominant. Also it is easy to give examples of conductance matrices which are not strongly column-sum dominant, and which belong to $\mathfrak{D}$. For instance, for the network of Fig. 3.

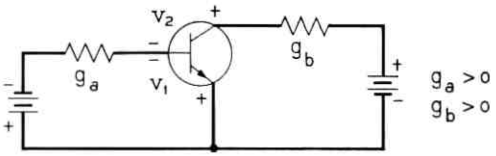


Fig. 3 — Single-transistor network.

$$G = \begin{pmatrix} g_a + g_b & -g_b \\ -g_b & g_b \end{pmatrix}$$

and diag $(d_1, d_2)G$ is strongly column-sum dominant for $d_2 = 1$ and some d_1 such that

$$\alpha_f < d_1 < \frac{1}{\alpha_r}.^*$$

* More generally, G of order $2p$ with positive diagonal elements belongs to $\mathfrak{D}$ whenever it is possible to obtain a strongly column-sum dominant matrix from G by adding an arbitrarily small positive quantity to a single diagonal element.

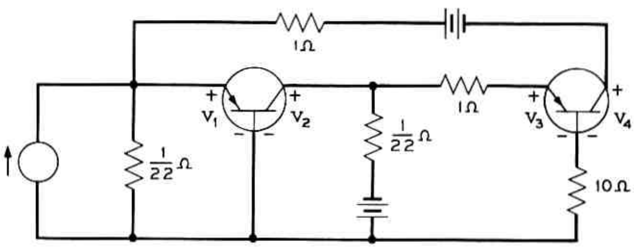


Fig. 4 — A two-transistor circuit.

As another example, consider the circuit of Fig. 4, for which

$$G = \frac{1}{21}\begin{bmatrix} 473 & -10 & 10 & -11 \\ -10 & 473 & -11 & 10 \\ 10 & -11 & 11 & -10 \\ -11 & 10 & -10 & 11 \end{bmatrix}.$$

Since diag (1, 1, 22, 22)G is strongly column-sum dominant, $G \in \mathfrak{D}$.

Finally, for the network shown in Fig. 5,

$$G = \frac{1}{21}\begin{bmatrix} 11 & -10 & 10 & -11 \\ -10 & 11 & -11 & 10 \\ 10 & -11 & 11 & -10 \\ -11 & 10 & -10 & 11 \end{bmatrix}.$$

In this case, G is obviously singular and hence does not belong to $\mathfrak{D}$. Suppose that the source current of Fig. 5 $i_0(t)$ is a constant and that the transistor functions $f_1(\cdot)$, $f_2(\cdot)$, $f_3(\cdot)$, and $f_4(\cdot)$ are all bounded from below by the constant b (this is certainly an assumption consistent with our earlier assumptions and with the character of transistor models

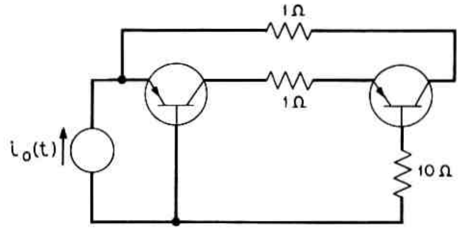


Fig. 5 — Transistor circuit for which the dc equations may have no solution.

ordinarily used.) We wish to show that here for sufficiently small i_0, there does not exist a constant vector u_∞ such that $[u(t) - u_\infty] \to \theta$ as $t \to \infty$.

Suppose that $u(t) \to u_\infty$, a constant vector, as $t \to \infty$. Then there would exist a $2p$-vector v_∞ such that $u_\infty = C(v_\infty)$ and

$$TF(v_\infty) + Gv_\infty = B$$

with $B = (i_0, 0, 0, \cdots, 0)^{tr}$. Let η denote the $2p$-row-vector $(1, 1, 1, \cdots, 1)$. Then

$$\eta TF(v_\infty) + \eta Gv_\infty = \eta B.$$

But $\eta Gv_\infty = 0$, and hence

$$i_0 = \sum_{k=1}^{p} [1 - \alpha_f^{(k)}] f_{2k-1}(v_{\infty 2k-1}) + \sum_{k=1}^{p} [1 - \alpha_r^{(k)}] f_{2k}(v_{\infty 2k})$$

which does not possess a solution v_∞ if

$$i_0 < b \sum_{k=1}^{p} [1 - \alpha_f^{(k)}] + [1 - \alpha_r^{(k)}].$$

2.1 *Estimation of the Rate of Decay of Transients*

Theorem 2: If the hypotheses of Corollary 1 are satisfied with $B(t) = B_\infty$ for $t \geqq 0$, then

$$\sum_{j=1}^{2p} d_j \mid u_j(t) - u_{\infty j} \mid \leqq \exp(-\underline{K}t) \sum_{j=1}^{2p} d_j \mid u_j(0) - u_{\infty j} \mid, \qquad t \geqq 0$$

for every set of positive constants $d_1, d_2, \cdots, d_{2p}$ such that

$$0 < \underline{K} \triangleq \min_j \min \left\{ \frac{1}{\tau_j} (1 - \tilde{d}_j d_j^{-1} \alpha_j), \frac{1}{c_j} (g_{jj} - \sum_{i \neq j} d_i d_j^{-1} \mid g_{ij} \mid) \right\}$$

in which $-\alpha_j$ is the nonzero off-diagonal term in the jth column of T, and $\tilde{d}_j = d_{j+1}$ for j odd and $\tilde{d}_j = d_{j-1}$ for j even.

It is easy to show that $G \varepsilon \mathfrak{D}$ implies that there are positive constants d_j, $j = 1, 2, \cdots, 2p$, such that $\underline{K} > 0$.

As an example of the application of Theorem 2, consider the problem of estimating the switching time of the single-transistor inverter circuit of Fig. 6 in which $\alpha_f = 0.968$, $c_e = 2 \times 10^{-12}$ fd, $\tau_e = 1.7 \times 10^{-10}$ second, $\alpha_r = 0.583$, $c_c = 1.7 \times 10^{-12}$ fd, and $\tau_c = 2.62 \times 10^{-8}$ second. Here (in mhos)

$$G = \begin{bmatrix} 1.1886 \times 10^{-3} & -1.01215 \times 10^{-3} \\ -1.01215 \times 10^{-3} & 1.01215 \times 10^{-3} \end{bmatrix}$$

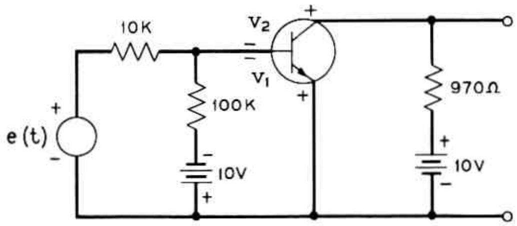


Fig. 6 — Practical logical-inverter circuit.

which takes into account a bulk base resistance of 280 ohms and a bulk collector resistance of 18 ohms. The circuit is initially at steady state with $e(t) = 0.3$ volt for $t < 0$. For $t \geqq 0$, $e(t) = 10$ volts, and as $t \rightarrow \infty$, $u(t) \rightarrow u_\infty$, some constant vector. With $d_2 = 1$, the number $\underline{K}$ is the smallest of the four quantities: $0.58(1 - 0.968d_1^{-1}) \times 10^{10}$, $0.5(1.1886 - 1.01215d_1^{-1}) \times 10^9$, $0.3815(1 - 0.583\mathrm{d}_1) \times 10^8$, and $0.58(1.01215)(1 - d_1) \times 10^9$.

It is clear that d_1 must satisfy $0.968 < d_1 < 1$ in order that $\underline{K} > 0$. Then optimal choice of d_1 (that is, the choice that yields the largest value of $\underline{K}$) is approximately 0.9709. For $d_1 = 0.9709$, $\underline{K} = 1.66 \times 10^7$. Let the "charge switching time" t_s denote the smallest value of t such that $\sum_{j=1}^{2} | u_j(t) - u_{\infty j} |$ is less than or equal to two percent of $\sum_{j=1}^{2} | u_j(0) - u_{\infty j} |$ for all $t \geqq t_s$. Then our upper bound on t_s is approximately $4 \times (1.66)^{-1} \times 10^{-7} \approx 241$ nanoseconds. The actual value of t_s, as determined by numerically integrating the system of two nonlinear differential equations is approximately 57 nanoseconds. Thus, for this circuit, Theorem 2 provides a very easily evaluated and useful upper bound on t_s.

Finally, we state a result which provides an often rather conservative but easily evaluated *lower bound* on the rate of decay of transients.

Theorem 3: With B a constant real 2p-vector, let

$$\frac{du}{dt} + TF[C^{-1}(u)] + GC^{-1}(u) = B, \qquad t \geqq 0.$$

If there exists a constant 2p-vector u_∞ such that $[u(t) - u_\infty] \rightarrow \theta$ as $t \rightarrow \infty$, then for any choice of positive constants d_j, $j = 1, 2, \cdots, 2p$:

$$\sum_{j=1}^{2p} d_j \, | u_j(t) - u_{\infty j} | \geqq \exp(-\bar{K}t) \sum_{j=1}^{2p} d_j \, | u_j(0) - u_{\infty j} |, \qquad t \geqq 0$$

in which

$$\bar{K} = \max_j \max \left\{ \frac{1}{\tau_j} (1 + \tilde{d}_j d_j^{-1} \alpha_j), \frac{1}{c_j} \sum_{i=1}^{2p} d_i d_j^{-1} \mid g_{ij} \mid \right\}$$

where $-\alpha_j$ is the nonzero off-diagonal element in the jth column of T, and $\tilde{d}_j = d_{j+1}$ for j odd, and $\tilde{d}_j = d_{j-1}$ for j even.

The arguments used to prove the results stated in this section can be modified in a straightforward manner to prove far more general results concerning networks that contain diodes, capacitors, and inductors, in addition to the elements of the structure of Fig. 1. Some of these more general results are described in Section IV.

III. PROOFS

3.1 *Proof of Theorem 1*

We first show that

$$F[C^{-1}(u_a)] - F[C^{-1}(u_b)] = D_1(u_a - u_b), \; t \geqq 0 \tag{12}$$

and

$$C^{-1}(u_a) - C^{-1}(u_b) = D_2(u_a - u_b), \; t \geqq 0 \tag{13}$$

with D_1 and D_2 diagonal matrices dependent on t and possessing some special properties.

For $j = 1, 2, \cdots, 2p$, let $g_j(u_{aj}) = [C^{-1}(u_a)]_j$ and $g_j(u_{bj}) = [C^{-1}(u_b)]_j$. Then, using equation (2),

$$u_{aj} - u_{bj} = c_j[g_j(u_{aj}) - g_j(u_{bj})] + \tau_j\{f_j[g_j(u_{aj})] - f_j[g_j(u_{bj})]\}.$$

Thus if $u_{aj} \neq u_{bj}$,

$$\frac{f_j[g_j(u_{aj})] - f_j[g_j(u_{bj})]}{u_{aj} - u_{bj}} = \frac{r_j(u_{aj}, u_{bj})}{c_j + \tau_j r_j(u_{aj}, u_{bj})},$$

in which (for $u_{aj} \neq u_{bj}$)

$$r_j(u_{aj}, u_{bj}) = \frac{f_j[g_j(u_{aj})] - f_j[g_j(u_{bj})]}{g_j(u_{aj}) - g_j(u_{bj})}.$$

In a similar manner we find that for all $u_{aj} \neq u_{bj}$:

$$\frac{g_j(u_{aj}) - g_j(u_{bj})}{u_{aj} - u_{bj}} = \frac{1}{c_j + \tau_j r(u_{aj}, u_{bj})}.$$

Now, let us define for $j = 1, 2, \cdots, 2p$

$$r_j(u_{aj}, u_{bj}) = f_j'[g_j(u_{aj})]$$

when $u_{aj} = u_{bj}$. Then since u_{aj} and u_{bj} are continuous on $[0, \infty)$, it follows (see Appendix A) that $r_j(u_{aj}, u_{bj})$ is continuous on $[0, \infty)$. Since $r_j(u_{aj}, u_{bj})$ is nonnegative, it is clear that both

$$\frac{r_j(u_{aj}, u_{bj})}{c_j + \tau_j r_j(u_{aj}, u_{bj})}$$

and

$$\frac{1}{c_j + \tau_j r_j(u_{aj}, u_{bj})}$$

are continuous on $[0, \infty)$. Moreover equations (12) and (13) are satisfied with

$$D_1 = \text{diag}\left\{\frac{r_j(u_{aj}, u_{bj})}{c_j + \tau_j r_j(u_{aj}, u_{bj})}\right\} \tag{14}$$

$$D_2 = \text{diag}\left\{\frac{1}{c_j + \tau_j r_j(u_{aj}, u_{bj})}\right\}. \tag{15}$$

At this point we have

$$\frac{d}{dt}(u_a - u_b) + (TD_1 + GD_2)(u_a - u_b) = B_a - B_b, \qquad t \geqq 0 \tag{16}$$

with $TD_1 + GD_2$ continuous on $[0, \infty)$.

We need the following lemma.

Lemma 1: Let $M(\cdot)$ be a continuous real $n \times n$ matrix-valued function of t defined on $[0, \infty)$ such that there exist positive constants ϵ and $c_1, c_2, \cdots, c_n$, with the property that for $j = 1, 2, \cdots, n$ and all $t \geqq 0$*

$$m_{jj} - \sum_{i \neq j} c_i c_j^{-1} \mid m_{ij} \mid \geqq \epsilon.$$

Let x be a differentiable real n-vector-valued function on $[0, \infty)$ such that

$$\frac{dx}{dt} + Mx = 0, \qquad t \geqq 0.$$

Then there exists a constant k such that for $i = 1, 2, \cdots, n$, and all $t \geqq 0$

$$\mid x_i(t) \mid \leqq k \exp(-\epsilon t).$$

Moreover, k depends only on the c_i and the initial values $x_i(0)$.

* In Ref. 4, Rosenbrock states a similar result, but does not give a rigorous proof. He considers the case in which $c_j = 1$ for $j = 1, 2, \cdots, n$.

Proof of Lemma 1: Let the functional s be defined in terms of an arbitrary continuously differentiable scalar function $\varphi(\cdot)$ by

$$\begin{aligned} s(\varphi)(t) &= 1 \quad \text{if} \quad \varphi(t) > 0 \quad \text{or} \quad \text{if} \quad \varphi(t) = 0 \quad \text{and} \quad \varphi'(t) > 0 \\ &= -1 \quad \text{if} \quad \varphi(t) < 0 \quad \text{or} \quad \text{if} \quad \varphi(t) = 0 \quad \text{and} \quad \varphi'(t) < 0 \\ &= 0 \quad \text{if} \quad \varphi(t) = 0 \quad \text{and} \quad \varphi'(t) = 0. \end{aligned}$$

Then for $t \geqq 0$,

$$\begin{aligned} \sum_i c_i s(x_i)(t) x_i'(t) &= -\sum_i c_i s(x_i)(t) \sum_j m_{ij} x_j \\ &= -\sum_j x_j \sum_i c_i s(x_i)(t) m_{ij} \\ &= -\sum_j x_j c_j s(x_j)(t) m_{jj} - \sum_j x_j \sum_{i \neq j} c_i s(x_i)(t) m_{ij} \\ &\leqq -\sum_j c_j m_{jj} \mid x_j \mid + \sum_j \mid x_j \mid \sum_{i \neq j} c_i \mid m_{ij} \mid \\ &\leqq -\epsilon \sum_j \mid c_j x_j \mid. \end{aligned}$$

But $\sum_i c_i s(x_i)(t) x_i'$ is equal to $\frac{d^+}{dt} \sum_i \mid c_j x_j \mid$, the right-hand derivative of $\sum_i \mid c_j x_j \mid$ [see Appendix B; the derivative of $\mid x_i \mid$ need not exist at points t at which $x_j(t) = 0$]. Therefore

$$\frac{d^+}{dt} \sum_j \mid c_j x_j \mid \leqq -\epsilon \sum_j \mid c_j x_j \mid, \qquad t \geqq 0$$

from which it follows that

$$\sum_j \mid c_j x_j(t) \mid \leqq \exp(-\epsilon t) \sum_j \mid c_j x_j(0) \mid, \qquad t \geqq 0. \ \square$$

If $M(\cdot)$ satisfies the conditions of Lemma 1, then it is easy to show that the unique continuously differentiable $n \times n$ matrix-valued function X defined on $[0, \infty)$ which satisfies

$$\frac{dX}{dt} + MX = 0, \qquad t \geqq 0 \quad [X(0) = I]$$

possesses the property that (for any norm $\|\cdot\|$ on E^n) there exists a constant K_1 such that

$$\| X(t) X(\tau)^{-1} \| \leqq K_1 \exp[-\epsilon(t - \tau)]$$

for all $t \geqq \tau$.

Returning now to equation (16), assume that $[TD_1 + GD_2]$ satisfies the conditions on $M(\cdot)$ of Lemma 1. Then with Y the solution of

$$\frac{dY}{dt} + [TD_1 + GD_2]Y = 0, \qquad t \geqq 0 \quad [Y(0) = I]$$

we have

$$u_a(t) - u_b(t) = Y(t)[u_a(0) - u_b(0)] + \int_0^t Y(t)Y(\tau)^{-1}[B_a(\tau) - B_b(\tau)]\, d\tau, \qquad t \geqq 0.$$

Therefore, for $t \geqq 0$

$$\begin{aligned} \| u_a(t) - u_b(t) \| &\leqq \| Y(t)[u_a(0) - u_b(0)]\| \\ &\quad + \int_0^t \| Y(t)Y(\tau)^{-1} \| \cdot \| B_a(\tau) - B_b(\tau) \| \, d\tau \\ &\leqq \| Y(t)[u_a(0) - u_b(0)] \| \\ &\quad + K_2 \int_0^t \exp\,[-\epsilon(t - \tau)] \, \| B_a(\tau) - B_b(\tau) \| \, d\tau \end{aligned}$$

for some positive constant K_2. Since $\| B_a(\tau) - B_b(\tau) \| \to 0$ as $\tau \to \infty$, it follows that $\| u_a(t) - u_b(t) \| \to 0$ as $t \to \infty$.

It remains only to prove that $[TD_1 + GD_2]$ meets the conditions imposed on $M(\cdot)$ of Lemma 1. Since $G \in \mathfrak{D}$, there exists a diagonal matrix diag $(d_1, d_2, \cdots, d_{2p})$ with $d_j > 0$ for $j = 1, 2, \cdots, 2p$ and

$$\alpha_f^{(k)} < \frac{d_{2k-1}}{d_{2k}} < \frac{1}{\alpha_r^{(k)}}$$

for $k = 1, 2, \cdots, p$ such that both

$$\text{diag}\,(d_1, d_2, \cdots, d_{2p})G$$

and

$$\text{diag}\,(d_1, d_2, \cdots, d_{2p})T$$

are strongly column-sum dominant. Thus for $j = 1, 2, \cdots, 2p$

$$t_{jj} - \sum_{i \neq j} d_i d_j^{-1} \mid t_{ij} \mid > 0$$

$$g_{jj} - \sum_{i \neq j} d_i d_j^{-1} \mid g_{ij} \mid > 0.$$

Let $W = TD_1 + GD_2$. Then, for $j = 1, 2, \cdots, 2p$,

$$w_{jj} = t_{jj} \frac{r_j}{c_j + \tau_j r_j} + g_{jj} \frac{1}{c_j + \tau_j r_j}$$

and

$$\sum_{i \neq j} d_i d_j^{-1} \mid w_{ij} \mid = \sum_{i \neq j} d_i d_j^{-1} \left| t_{ij} \frac{r_j}{c_j + \tau_j r_j} + g_{ij} \frac{1}{c_j + \tau_j r_j} \right|.$$

Therefore

$$w_{jj} - \sum_{i \neq j} d_i d_j^{-1} \mid w_{ij} \mid \geqq \frac{r_j}{c_j + \tau_j r_j} (t_{jj} - \sum_{i \neq j} d_i d_j^{-1} \mid t_{ij} \mid) + \frac{1}{c_j + \tau_j r_j} (g_{jj} - \sum_{i \neq j} d_i d_j^{-1} \mid g_{ij} \mid). \tag{17}$$

Since $r_j \geqq 0$, the right side of equation (17) is bounded from below by some positive constant ϵ uniformly in t and j. □

3.2 *Proof of Corollary 1*

By Corollary 3 of Ref. 5 there exists a unique $v \in E^{2p}$ such that

$$TF(v) + Gv = B_\infty \tag{18}$$

whenever G is such that all principal minors of $T^{-1}G$ are positive. In Reference 5 it is proved that $T^{-1}G$ will have this property if $T^{-1}G$ can be written as $A^{-1}B$ with both A and B stongly column-sum dominant.

Let $H = \text{diag}\,(d_1, d_2, \cdots, d_{2p})G$ be strongly column-sum dominant with all $d_j > 0$ and

$$\alpha_f^{(k)} < \frac{d_{2k-1}}{d_{2k}} < \frac{1}{\alpha_r^{(k)}}$$

for $k = 1, 2, \cdots, p$. Then $U \triangleq \text{diag}\,(d_1, d_2, \cdots, d_{2p})T$ is strongly column-sum dominant, and $T^{-1}G = U^{-1}H$, which proves that equation (18) possesses a unique solution v.

With v the solution of equation (18), let $u_\infty = C(v)$. Clearly if $B_\infty = \theta$, then $u_\infty = \theta$. Let u_b satisfy

$$\frac{du_b}{dt} + TF[C^{-1}(u_b)] + G[C^{-1}(u_b)] = B_\infty, \qquad t \geqq 0$$

with $u_b(0) = u_\infty$. Of course, $u_b(t) = u_\infty$ for all $t \geqq 0$. By Theorem 1, $[u(t) - u_\infty] \rightarrow \theta$ as $t \rightarrow \infty$, independent of u_0.

3.3 *Proof of Theorem 2*

Following the proofs of Theorem 1 and Corollary 1,

$$\frac{d}{dt}(u - u_\infty) + (TD_1 + GD_2)(u - u_\infty) = 0, \qquad t \geqq 0$$

in which

$$D_1 = \operatorname{diag}\left\{\frac{r_j(u_j\ , u_{\infty j})}{c_j + \tau_j r_j(u_j\ , u_{\infty j})}\right\}$$

and

$$D_2 = \operatorname{diag}\left\{\frac{1}{c_j + \tau_j r_j(u_j\ , u_{\infty j})}\right\}.$$

Therefore

$$\frac{d^+}{dt} \sum_j d_j \mid u_j(t) - u_{\infty j} \mid \leqq -\underline{K} \sum_j d_j \mid u_j(0) - u_{\infty j} \mid, \qquad t \geqq 0$$

in which

$$\underline{K} = \min_j \min \left\{\frac{1}{\tau_j}(t_{jj} - \sum_{i \neq j} d_i d_j^{-1} \mid t_{ij} \mid), \frac{1}{c_j}(g_{jj} - \sum_{i \neq j} d_i d_j^{-1} \mid g_{ij} \mid)\right\}.$$

But for $j = 1, 2, \cdots, 2p$

$$t_{jj} - \sum_{i \neq j} d_i d_j^{-1} \mid t_{ij} \mid = 1 - \tilde{d}_j d_j^{-1} \alpha_j \ . \ \square$$

3.4 *Proof of Theorem 3*

Since $TF[C^{-1}(z)] + GC^{-1}(z)$ depends continuously on $z \varepsilon E^{2p}$, u_∞ satisfies (see Ref. 6)

$$TF[C^{-1}(u_\infty)] + GC^{-1}(u_\infty) = B.$$

Therefore, following the proofs of Theorem 1 and Corollary 1,

$$\frac{d}{dt}(u - u_\infty) + (TD_1 + GD_2)(u - u_\infty) = 0, \qquad t \geqq 0$$

in which

$$D_1 = \operatorname{diag}\left\{\frac{r_j(u_j\ , u_{\infty j})}{c_j + \tau_j r_j(u_j\ , u_{\infty j})}\right\}$$

and

$$D_2 = \operatorname{diag}\left\{\frac{1}{c_j + \tau_j r_j(u_j\ , u_{\infty j})}\right\}.$$

For any $z \varepsilon E^{2p}$, let $\| z \|$ denote $\sum_i d_i \, |z_i|$. Then, for $t \geqq 0$

$$\left\| \frac{d}{dt}(u - u_\infty) \right\| = \| (TD_1 + GD_2)(u - u_\infty) \|$$

$$\leqq \max_i \left\{ (1 + \tilde{d}_j d_j^{-1} \alpha_j) \frac{r_j(u_j\,, u_{\infty j})}{c_j + \tau_j r_j(u_j\,, u_{\infty j})} + \sum_{i=1}^{2p} d_i d_j^{-1} \mid g_{ij} \mid \frac{1}{c_j + \tau_j r_j(u_j\,, u_{\infty j})} \right\} \| u(t) - u_\infty \|.$$

But, since $r_i(u_j\,, u_{\infty i}) \geqq 0$,

$$\bar{K} \geqq \max_i \left\{ (1 + \tilde{d}_j d_j^{-1} \alpha_j) \frac{r_j(u_j\,, u_{\infty j})}{c_j + \tau_j r_j(u_j\,, u_{\infty j})} + \sum_{i=1}^{2p} d_i d_j^{-1} \mid g_{ij} \mid \frac{1}{c_j + \tau_j r_j(u_j\,, u_{\infty j})} \right\}.$$

Thus

$$\left\| \frac{d}{dt}(u - u_\infty) \right\| \leqq \bar{K} \, \| u - u_\infty \|, \qquad t \geqq 0. \tag{19}$$

Clearly,

$$\left\| \frac{d}{dt}(u - u_\infty) \right\| = \lim_{\epsilon \to 0+} \frac{1}{\epsilon} \| u(t + \epsilon) - u_\infty - u(t) + u_\infty \|, \qquad t \geqq 0$$

Also, for $t \geqq 0$, the limit

$$\lim_{\epsilon \to 0+} \frac{1}{\epsilon} [\| u(t) - u_\infty \| - \| u(t + \epsilon) - u_\infty \|]$$

exists and is equal to $-\dfrac{d^+}{dt} \| u - u_\infty \|$ in which as before $\dfrac{d^+}{dt}$ denotes the right-hand derivative (see Appendix B). But, since for any $\epsilon > 0$ and $t \geqq 0$,

$$\| u(t) - u_\infty \| - \| u(t + \epsilon) - u_\infty \| \leqq \| u(t + \epsilon) - u_\infty - u(t) + u_\infty \|,$$

we have

$$-\frac{d^+}{dt} \| u - u_\infty \| \leqq \left\| \frac{d}{dt}(u - u_\infty) \right\|, \qquad t \geqq 0. \tag{20}$$

Therefore, using equations (19) and (20),

$$\frac{d^+}{dt} \| u - u_\infty \| \geqq -\bar{K} \, \| u - u_\infty \|, \qquad t \geqq 0$$

and, for $t \geqq 0$,

$$|| u - u_\infty || \geqq \exp(-\bar{K}t) \, || u(0) - u_\infty ||. \ \square$$

IV. A SIGNIFICANT EXTENSION

We can easily extend our results to cover an interesting class of networks containing diodes, capacitors (not necessarily linear), and (not necessarily linear) inductors, in addition to the elements of the Fig. 1 network.

Let each diode be represented by a model of the type shown in Fig. 7 in which

$$i_d = \frac{d}{dt} [c_d v_d + \tau_d f_d(v_d)] + f_d(v_d),$$

with c_d and τ_d positive constants. Assume that $f_d(\cdot)$ satisfies the conditions placed on $f_e(\cdot)$ and $f_c(\cdot)$ of the transistor model. Let there be q diodes and let v_{2p+k} and i_{2p+k} ($k = 1, 2, \cdots, q$) be the voltage and current associated with the kth diode.

Suppose that the kth capacitor (we assume that there are r capacitors) is governed by

$$\frac{d}{dt} [c_{2p+q+k}(v_{2p+q+k})] = i_{2p+q+k}$$

for $k = 1, 2, \cdots, r$, where $c_{2p+q+k}(\cdot)$ is a strictly-monotone-increasing continuously-differentiable mapping of E^1 *onto* itself such that $c_{2p+q+k}(0) = 0$ and the slope of $c_{2p+q+k}(\cdot)$ is uniformly bounded from above and from below by positive constants.

Finally, let there be s inductors which introduce constraints

$$\frac{d}{dt} [l_{2p+q+r+k}(i_{2p+q+r+k})] = v_{2p+q+r+k}$$

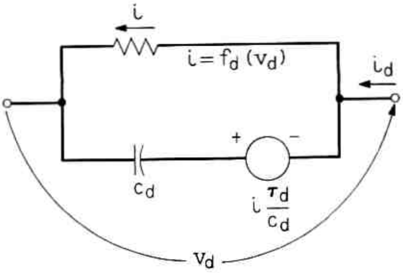


Fig. 7 — Diode model.

for $k = 1, 2, \cdots, s$, in which each $l_{2p+q+r+k}(\cdot)$ is a function of the same type as the $c_{2p+q+k}(\cdot)$.

Assume that the linear resistive portion of the network introduces the constraint

$$\tilde{\imath} = -H\tilde{v} + B, \qquad B \,\varepsilon\, \mathfrak{B}$$

in which $\tilde{\imath} = (i_1, i_2, \cdots, i_{2+p+q+r}, v_{2p+q+r+1}, \cdots, v_{2p+q+r+s})^{tr}$, $\tilde{v} = (v_1, v_2, \cdots, v_{2p+q+r}, i_{2p+q+r+1}, \cdots, i_{2p+q+r+s})^{tr}$, and H is a constant hybrid-parameter matrix of order $(2p + q + r + s)$. Then

$$\frac{d}{dt}[\tilde{C}(\tilde{v})] + \tilde{T}\tilde{F}(\tilde{v}) + H\tilde{v} = B$$

where

$$\begin{aligned}
[\tilde{C}(\tilde{v})]_j &= [C(v)]_j\,, & j &= 1, 2, \cdots, 2p\\
&= c_jv_j + \tau_jf_j(v_j), & j &= 2p+1, 2p+2, \cdots, 2p+q\\
&= c_j(v_j), & j &= 2p+q+1, \cdots, 2p+q+r\\
&= l_j(i_j), & j &= 2p+q+r+1, \cdots, 2p+q+r+s;
\end{aligned}$$

$\tilde{T}$ is the direct sum of matrices $T \oplus I_q \oplus 0_{r+s}$, in which I_q is the identity matrix of order q and 0_{r+s} is the zero matrix of order $(r + s)$, and

$$\begin{aligned}
[\tilde{F}(\tilde{v})]_j &= [F(v)]_j\,, & j &= 1, 2, \cdots, 2p\\
&= f_j(v_j), & j &= 2p+1, \cdots, 2p+q.
\end{aligned}$$

Under our assumptions $\tilde{C}(\cdot)^{-1}$ exists and, with $\tilde{u} = \tilde{C}(\tilde{v})$,

$$\frac{d\tilde{u}}{dt} + \tilde{T}\tilde{F}[\tilde{C}^{-1}(\tilde{u})] + H\tilde{C}^{-1}(\tilde{u}) = B. \tag{21}$$

Let $\tilde{\mathfrak{D}}$ denote the set of all real matrices M of order $(2p+q+r+s)$ such that there exist positive constants $d_1, d_2, \cdots, d_{2p+q+r+s}$ with the property that

$$\alpha_f^{(k)} < \frac{d_{2k-1}}{d_{2k}} < \frac{1}{\alpha_r^{(k)}}$$

for $k = 1, 2, \cdots, p$ (when $p \neq 0$) and $\operatorname{diag}(d_1, d_2, \cdots, d_{2p+q+r+s})M$ is strongly column-sum dominant.

With straightforward modifications of the arguments already presented, we can prove (*i*) that for each $\tilde{u}_0 \,\varepsilon\, E^{(2p+q+r+s)}$ equation (21) possesses a unique solution defined on $[0, \infty)$ such that $\tilde{u}(0) = \tilde{u}_0$, and (*ii*) the analogs of Theorems 1, 2, and 3 and Corollary 1. To be more specific, the analogs of Theorem 1, Corollary 1, and Theorem 2 are:

Theorem 1′: If $H \in \tilde{\mathfrak{D}}$, *and* $\tilde{u}_a$ *and* $\tilde{u}_b$ *are solutions of equation* (21) *with* $B = B_a$ *and* $B = B_b$, *respectively, for* $t \geqq 0$, *and if* $[B_a(t) - B_b(t)] \rightarrow \theta$ [*the zero vector of* $E^{(2p+q+r+s)}$] *as* $t \rightarrow \infty$ *with* $B_a \in \mathcal{B}$ *and* $B_b \in \mathcal{B}$, *then* $[\tilde{u}_a(t) - \tilde{u}_b(t)] \rightarrow \theta$ *as* $t \rightarrow \infty$.

Corollary 1′: Referring to equation (21), *if* $H \in \tilde{\mathfrak{D}}$, *and if there exists a constant vector* B_∞ *such that* $[B(t) - B_\infty] \rightarrow \theta$ *as* $t \rightarrow \infty$, *then there exists a constant vector* $\tilde{u}_\infty$ *such that* $[\tilde{u}(t) - \tilde{u}_\infty] \rightarrow \theta$ *as* $t \rightarrow \infty$, *and* $\tilde{u}_\infty$ *is independent of the initial condition* $\tilde{u}_0$. *In particular, if* $B_\infty = \theta$, *then* $\tilde{u}_\infty = \theta$.

Theorem 2′: If the hypotheses of Corollary 1′ are satisfied with $B(t) = B_\infty$ *for* $t \geqq 0$, *then with* $j_0 = (2p + q + r + s)$, *we have*

$$\sum_{j=1}^{j_0} d_j \mid \tilde{u}_j(t) - u_{\infty j} \mid \leqq \exp(-\tilde{\underline{K}}t) \sum_{j=1}^{j_0} d_j \mid \tilde{u}_j(0) - \tilde{u}_{\infty j} \mid, \qquad t \geqq 0$$

for every set of positive constants $d_1, d_2, \cdots, d_{2p+q+r+s}$ *such that* $0 < \tilde{\underline{K}} = \min \{\tilde{\underline{K}}_1, \tilde{\underline{K}}_2, \tilde{\underline{K}}_3\}$ *where*

$$\tilde{\underline{K}}_1 = \min_{1 \leqq j \leqq 2p} \min \left\{ \frac{1}{\tau_j}(1 - \tilde{d}_j d_j^{-1} \alpha_j), \frac{1}{c_j}(g_{jj} - \sum_{i \neq j} d_i d_j^{-1} \mid g_{ij} \mid) \right\}$$

$$\tilde{\underline{K}}_2 = \min_{2p+1 \leqq j \leqq 2p+q} \min \left\{ \frac{1}{\tau_j}, \frac{1}{c_j}(g_{jj} - \sum_{i \neq j} d_i d_j^{-1} \mid g_{ij} \mid) \right\}$$

$$\tilde{\underline{K}}_3 = \min_{2p+q+1 \leqq j \leqq 2p+q+r+s} \left\{ \frac{1}{s_j}(g_{ii} - \sum_{i \neq j} d_i d_j^{-1} \mid g_{ij} \mid) \right\}$$

in which $s_j = \sup c_j'(\cdot)$ *for* $j = 2p + q + 1, \cdots, 2p + q + r$; $s_j = \sup l_j'(\cdot)$ *for* $j = 2p + q + r + 1, \cdots, 2p + q + r + s$; $-\alpha_j$ *is the nonzero off-diagonal term in the jth column of* T; *and* $\tilde{d}_j = d_{j+1}$ *for j odd and* $\tilde{d}_j = d_{j-1}$ *for j even. Moreover there exists one such set of constants* $\{d_j\}$.

V. FINAL COMMENTS

The results presented here are quite encouraging in that they are concerned with the equations of reasonably realistic nonlinear network models, and provide some understanding of a precise nature in an area where there is a great need for many results of similar type.

VI. ACKNOWLEDGMENT

The writer is indebted to H. Shichman for supplying the data and computer results for the practical circuit of Fig. 6.

APPENDIX A

Proof that $r_j(u_{aj}, u_{bj})$ is continuous.

It is clear that $r_j(u_{aj}, u_{bj})$ is continuous at each point t such that $u_{aj}(t) \neq u_{bj}(t)$. Suppose now that t is such that $u_{aj}(t) = u_{bj}(t)$, and let $\epsilon > 0$ be given. Since u_{aj}, u_{bj}, g and f_j' are continuous, there exists $\delta_1 > 0$ such that

$$| f_j' \{g_j[u_{aj}(t + \eta)]\} - f_j' \{g_j[u_{aj}(t)]\} | \leqq \epsilon$$

for all $| \eta | \leqq \delta_1$. Then for $| \eta | \leqq \delta_1$ either $u_{aj}(t + \eta) = u_{bj}(t + \eta)$ in which case

$$| r_j[u_{aj}(t + \eta), u_{bj}(t + \eta)] - r_j[u_{aj}(t), u_{bj}(t)] | \leqq \epsilon$$

or $u_{bj}(t + \eta) \neq u_{bj}(t + \eta)$ and (using the mean-value theorem)

$$r_j[u_{aj}(t + \eta), u_{bj}(t + \eta)] = \frac{f_j\{g_j[u_{aj}(t + \eta)]\} - f_j\{g_j[u_{bj}(t + \eta)]\}}{g_j[u_{aj}(t + \eta)] - g_j[u_{bj}(t + \eta)]}$$

$$= f_j'(\xi)$$

in which

$$| \xi - g_j[u_{aj}(t)] | \leqq \max \{ | g_j[u_{aj}(t + \eta)] - g_j[u_{aj}(t)] |, | g_j[u_{bj}(t + \eta)] - g_j[u_{aj}(t)] | \}.$$

In the latter case, there exists $\delta_2 > 0$ such that $| f'(\xi) - f' \{g_j[u_{aj}(t)]\} | \leqq \epsilon$ for all $| \eta | \leqq \delta_2$. Thus for all $| \eta | \leqq \min \{\delta_1, \delta_2\}$, we have

$$| r_j[u_{aj}(t + \eta), u_{bj}(t + \eta)] - r[u_{aj}(t), u_{bj}(t)] | \leqq \epsilon.$$

APPENDIX B

Proof that the Right-Hand Derivative of $| x_j |$ exists and is equal to $s(x_j)(t)x_j'$.

If t is a point such that $x_j(t) \neq 0$, then it is clear that

$$\frac{d^+}{dt} | x_j | = s(x_j)(t)x_j'(t).$$

At t such that $x_j(t) = 0$ and $x_j'(t) \neq 0$,

$$s(x_j)(t)x_j' = \lim_{\epsilon \to 0} s(x_j)(t) \frac{x_j(t + \epsilon)}{\epsilon} = \frac{d^+}{dt} | x_j |.$$

Finally if $x_i(t) = 0$ and $x_i'(t) = 0$, then

$$0 = \lim_{\substack{\epsilon \to 0+ \\ t \leq \xi \leq t+\epsilon}} \mid x_i'(\xi) \mid = \lim_{\epsilon \to 0+} \frac{\mid x_i(t+\epsilon) \mid}{\epsilon} = \frac{d^+}{dt} \mid x_i \mid,$$

since x_i is continuously differentiable.

REFERENCES

1. Gummel, H. K., "A Charge-Control Transistor Model for Network Analysis Programs," Proc. IEEE, *56,* No. 4 (April 1968), p. 751.
2. Koehler, D., "The Charge-Control Concept in the Form of Equivalent Circuits, Representing a Link Between the Classic Large Signal Diode and Transistor Models," B.S.T.J., *46,* No. 3 (March 1967), pp. 523–576.
3. Nemytskii, V. V. and Stepanov, V. V., *Qualitative Theory of Differential Equations,* Princeton, New Jersey: Princeton University Press, 1960, pp. 9–13.
4. Rosenbrock, H. H., "A Method of Investigating Stability." Proc. 2nd International Federation Automatic Control Congress, Basel, Switzerland, 1963, pp. 590–594.
5. Sandberg, I. W. and Willson, A. N., "Some Theorems on Properties of DC Equations of Nonlinear Networks," B.S.T.J., this issue, pp. 1–34.
6. Bellman, R., *Stability Theory of Differential Equations,* New York: McGraw-Hill, 1953, p. 77.

Adaptive Equalization of Highly Dispersive Channels for Data Transmission

By ALLEN GERSHO

(Manuscript received May 29, 1968)

This paper analyzes an adaptive training algorithm for adjusting the tap weights of a tapped delay line filter to minimize mean-square intersymbol interference for synchronous data transmission. The significant feature of the adjustment procedure is that convergence is guaranteed for all channel response pulses, even for very severe amplitude and phase distortion.

The author examines convergence, rate of convergence, and the effect of noisy observations of the received pulses, and he shows that the noisy observations result in a random sequence of tap weight settings whose mean value converges to a suboptimal setting. The mean-square deviation of the tap weights from the suboptimal values is asymptotically bounded with a bound that can be made as small as desired by sufficiently reducing the speed of convergence.

The suboptimality arising here results from the use of isolated test pulses for the training signal. However, a training scheme using pseudorandom sequences or the actual data signal does not suffer from the suboptimality effect. Hence, although of possible utility in other pulse shaping applications, the technique presented here appears to be primarily of value in providing a conceptual framework for the closely related but more practical techniques to be examined in the sequel to this paper to be published shortly.

I. INTRODUCTION

A common approach to data transmission is to code the amplitudes of successive pulses in a periodic pulse train with a discrete set of possible amplitude levels. The coded pulse train is then linearly modulated, transmitted through the channel, demodulated, equalized, and synchronously sampled and quantized. As a result of dispersion of the pulse shape by the channel, the number of detectable amplitude levels has very often been limited by intersymbol interference rather than by additive noise.

In principle, if the channel is known precisely it is virtually always possible to design an equalizer that will make the intersymbol interference (at the sampling instants) arbitrarily small. However, in practice a channel is random in the sense of being one of an ensemble of possible channels. Consequently, a fixed equalizer designed on average channel characteristics may not adequately reduce intersymbol interference. An adaptive equalizer is then needed which can be "trained," with the guidance of a suitable training signal transmitted through the channel, to adjust its parameters to optimal values. If the channel is also time-varying, an adaptive equalizer operating in a tracking mode is needed which can update its parameter values by tracking the changing channel characteristics during the course of normal data transmission. In both cases the adaptation may be achieved by observing or estimating the error between actual and desired equalizer responses and using this error to estimate the direction in which the parameters should be changed to approach the optimal values.

A simple and effective technique for adaptive equalization was developed by Lucky using the tapped delay line filter structure for the equalizer.[1, 2] The main limitation of this technique is that convergence of the tap weight adjustment algorithm is assured only for relatively low dispersion channels. The convergence condition requires that the dispersed pulse shape have adequate quality so that, in the absence of noise, error-free binary data transmission would be possible without equalization. In other words the dispersed pulse must have an open binary "eye."

Using an approach to adaptation[3, 4] with virtually unrestricted convergence properties, Lucky and Rudin subsequently proposed and implemented an adaptive equalizer for minimizing the mean square error in frequency response of an analog channel.[5, 6] This approach was applied to synchronous data transmission by the author and independently by Lytle and by Niessen.[7–9] An implementation of the technique was described by Niessen and Drouilhet.[10] It has also been implemented for data communication at Bell Laboratories.

In this paper the approach is used for synchronous data transmission in a training mode where a sequence of isolated pulses is used as a test signal. The technique may be viewed equally as an adaptive design procedure for a sampled-data pulse shaping filter where the error criterion is to minimize the mean square error between actual and desired pulse shapes at the filter output. The important feature of the technique is that convergence is achieved for any channel pulse response whatever, thereby including highly dispersed pulses for which even binary data transmission would be impossible without equalization. Of particular interest are: (*i*) the analogous optimality condition to Lucky's zero forcing condition resulting with the change from a summed absolute error to a summed squared error criterion,[1] (*ii*) the manner in which noisy observations introduce randomness in the iterative corrections to the weights and the resulting stochastic convergence properties, (*iii*) the possibility of applying the technique where isolated pulses applied to a filter must be used to adaptively adjust the filter for optimum pulse shaping (unrelated to equalization), and (*iv*) the conceptual framework for the more practical adaptation techniques to be described in a sequel to this paper, planned for publication soon.

Perhaps the earliest application of the tapped delay line or "transversal" filter to pulse shaping for data transmission was made by W. P. Boothroyd and E. M. Creamer.[11] Tufts and George have shown that under a mean-square error criterion the optimal receiver structure includes a tapped-delay line filter with delay between taps equal to the symbol period.[12, 13] Aaron and Tufts have also shown that the same receiver structure is needed to minimize the average error probability for binary data transmission.[14]

The basic approach to adaptive adjustment of a set of weights where a mean-square error criterion is used with a gradient search procedure was considered by Widrow and Hoff who noticed that no derivative computation is needed.[3] Narendra and McBride proposed a self-optimizing Wiener filter using a continuous-time gradient algorithm and a filter structure whose transfer function is a weighted sum of fixed functions.[4] Koford and Groner used a mean-square error criterion and a gradient learning algorithm to find an optimum set of weights for pattern classifying.[15] Widrow described a general adaptive filtering problem with the tapped delay line filter.[16] Coll and George discussed the performance of George's optimum equalizer and indicated a possible adaptive adjustment technique.[17] Lucky and Rudin were the first to apply the mean square error criterion with the gradient search procedure to the field of adaptive equalization.[5, 6] This paper expands on a short presentation given at an international symposium on information theory.[18]

II. PERFORMANCE OBJECTIVES FOR EQUALIZATION

The objective of equalization, viewed as a pulse shaping problem, is to adjust the parameters of the equalizer to a setting which minimizes a suitable measure of the error between actual and desired pulse shapes. For the usual synchronous data transmission application, the desired pulse shape is one with the Nyquist property that the sample values y_k at the sampling instant kT are given by $y_k = \delta_{kr}$ where δ_{kr} is unity for $k = r$ and zero for all other integers k. The criterion used by Lucky[1] is peak distortion, D, given by

$$D = \sum_{k \neq r} | y_k | / | y_r | .$$

An alternate criterion of interest is the mean square distortion E, defined by

$$E = \sum_{k \neq r} y_k^2 / y_r^2 \ .$$

The physical interpretation of the peak distortion is that it is directly related to eye opening and determines the error probability for a worst case message pattern. The mean square distortion has a different interpretation. If the message pattern is such that the transmitted level for each time slot is statistically independent of the levels for other time slots, then the variance of the intersymbol interference in a given time slot is proportional to the mean square distortion. If the pulse shape has a large number of small sidelobes so that the intersymbol interference is normally distributed, then minimizing mean square distortion is equivalent to minimizing error rate.

Closely related to the mean square distortion is the mean square error

$$\mathcal{E} = \sum_{k} (y_k - d_k)^2 \tag{1}$$

where d_k is the desired pulse sample value at time instant kT. For the usual equalization problem where $d_k = \delta_{kr}$, the measure $\mathcal{E}$ has virtually the same interpretation as E; however, E is a normalized measure independent of pulse amplitude while $\mathcal{E}$ depends on both shape and amplitude. Optimization of the tapped delay line equalizer with respect to either criterion leads to equivalent results.

III. FORMULATION

Consider the transversal equalizer with N taps and tap spacing T equal to the symbol period. Let c_k be the weight at the kth tap for

$k = 0, 1, \cdots, N-1$ so that the input output relation of the transversal filter at the sample times is

$$y_n = \sum_{k=0}^{N-1} c_k x_{n-k} = \mathbf{c}'\mathbf{x}_n \tag{2}$$

where x_k and y_k denote the input and output pulse samples, respectively, at time instants kT, $\mathbf{c} = (c_0, c_1, \cdots, c_{N-1})$ is the tap weight vector, and $\mathbf{x}_n = (x_n, x_{n-1}, \cdots, x_{n-N+1})$ is the sample memory state of the delay line at the time instant nT; the vectors $\mathbf{c}$ and $\mathbf{x}_n$ are to be regarded as column matrices, and the prime denotes the transpose. We assume that the input sequence x_k has finite energy. Let $\epsilon_n = y_n - d_n$. Then from equation (1), using (2), the gradient of the error with respect to $\mathbf{c}$ may be written as

$$\nabla \mathcal{E} = 2 \sum_k \epsilon_k \mathbf{x}_k . \tag{3}$$

Therefore the optimality condition for minimum error $\nabla \mathcal{E} = 0$ is equivalent to the requirement that the (deterministic) corss-correlation between the input sequence x_k and output error sequence ϵ_k must have zeros for the N components with index values corresponding to the index values of the available tap weights. That is,

$$\varphi_{x\epsilon}(k) = \sum_n \epsilon_n x_{n-k} = 0 \qquad \text{for} \quad k = 0, 1, 2, \cdots, N-1.$$

This condition has an interesting similarity to Lucky's condition which states that the peak distortion, D, is minimized when the error sequence ϵ_n has zeros for the N components with index values corresponding to the index values of the available tap weights.[1] An important distinction is that Lucky's condition is generally not valid when the input pulse distortion D exceeds unity, while the mean square optimizing condition is valid for any input pulse with finite mean square distortion.

Using equation (2), the gradient (3) can be expressed explicitly as a function of the tap weight vector $\mathbf{c}$, namely:

$$\nabla \mathcal{E} = 2(\mathbf{Ac} - \mathbf{g}) \tag{4}$$

where

$$\mathbf{A} = \sum_n \mathbf{x}_n \mathbf{x}_n', \qquad \text{and} \qquad \mathbf{g} = \sum_n d_n \mathbf{x}_n .$$

Notice that $\mathbf{A}$ is symmetric and positive definite (see Appendix A). Setting equation (4) equal to zero yields the solution for the optimum tap weight vector $\mathbf{c}^*$,

$$\mathbf{c}^* = \mathbf{A}^{-1}\mathbf{g}.$$

Using equation (2), the error expression given by equation (1) may be expressed in the convenient form:

$$\mathcal{E}(\mathbf{c}) = \mathcal{E}(\mathbf{c}^*) + (\mathbf{c} - \mathbf{c}^*)'\mathbf{A}(\mathbf{c} - \mathbf{c}^*) \tag{5}$$

which shows explicitly the simple quadratic nature of the error surface and the unique optimality of the minimizing weight vector $\mathbf{c}^*$. It can be shown that the residual error $\mathcal{E}(\mathbf{c}^*)$ can be made as small as desired for all channels of practical interest by using a sufficiently large number, N, of taps.[19]

It is intuitively reasonable that successive corrections to the tap weight vector in the direction of steepest descent of the error surface should lead to the minimum error where $\mathbf{c} = \mathbf{c}^*$. This is the idea of the well-known[20] gradient algorithm:

$$\mathbf{c}_{i+1} = \mathbf{c}_i - \tfrac{1}{2}\alpha\nabla\mathcal{E}(\mathbf{c}_i) \tag{6}$$

where α is a suitably small positive proportionality constant, $\mathbf{c}_0$ is arbitrary, and $\mathbf{c}_i$ is the tap weight vector after the ith iteration.

The significant feature of the gradient algorithm for our quadratic error surface (5) is that the gradient can be conveniently evaluated without knowledge of the error surface itself. We have seen from equation (3) that the components of the gradient vector are values of the crosscorrelation between the input sequence and the output error sequence. This suggests the conceptually simple implementation where an isolated test pulse is transmitted through the channel and the requisite crosscorrelation values are formed by multiplying the delayed input pulse with the error pulse, sampling, and summing (or averaging). The tap weights are then incremented according to (6), the old crosscorrelation values "dumped" and a new iteration is begun with the transmission of a new test pulse.

The error pulse is formed by subtracting from the equalizer output pulse an "ideal" pulse whose sample values are the desired values d_k; the ideal pulse is locally generated at the appropriate time. The basic scheme is shown in Fig. 1. Naturally, the summation given by equation (3) cannot be performed over an infinite time interval. Suppose κT is a practical upper bound on the possible time duration of the input pulse, ξT is the time interval between successive test pulses with $\xi T > \kappa T$, ξ and κ as positive integers. Then if we include the effect of perturbing

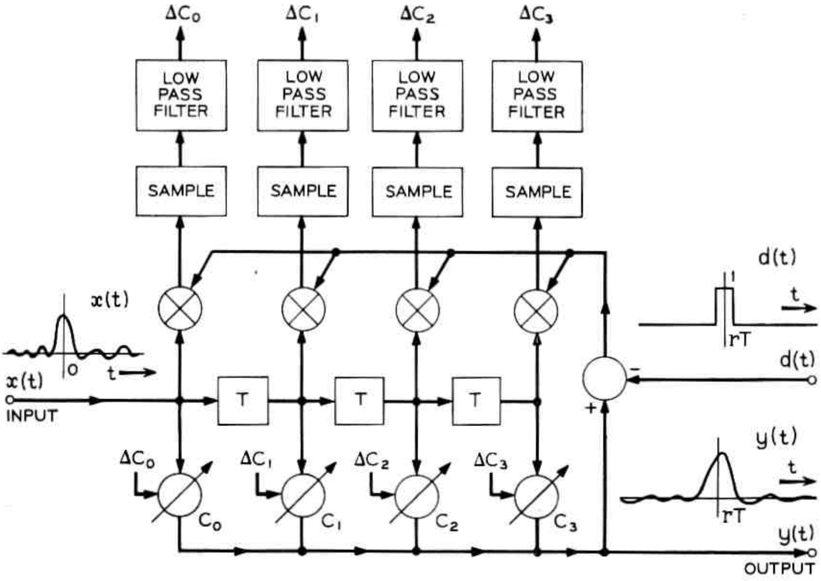


Fig. 1 — Four tap training mode adaptive equalizer.

receiver noise samples n_i and z_i at the equalizer input and output, respectively, the measured crosscorrelation vector $\hat{\varphi}_i$ after the ith iteration is given by:

$$\hat{\varphi}_i = \sum_{l=l_0+i\xi}^{l_0+i\xi+\kappa} (\mathbf{x}_{l-i\xi} + \mathbf{n}_l)(\epsilon_{l-i\xi} + z_l). \tag{7}$$

In the noiseless case the estimate $\hat{\varphi}_i$ reduces to one-half the deterministic gradient, that is, $\frac{1}{2}\nabla\mathcal{E}(\mathbf{c}_i)$ under the assumption that the pulse sequence x_l and desired sequence d_l are virtually zero outside of the interval $l_0 \leqq l \leqq l_0 + \kappa - N + 1$.

IV. CONVERGENCE PROPERTIES

In the presence of noise the tap weight corrections contain undesired random components consisting of products of input and output noise samples and products of pulse and noise sample. As a result, the random tap weights no longer converge to the optimal values but instead approach some neighborhood of a suboptimal setting and then fluctuate randomly about this setting. The error between the optimal and suboptimal settings is small for low noise levels and decreases with increasing signal-to-noise ratios. The size of the fluctuation neighborhood about the suboptimal setting is proportional to the noise level but can be made as small as desired by making the training time sufficiently long.

Assume the noise samples n_i have zero mean and finite variance σ^2. Define the vector $\mathbf{n}_k = (n_k, n_{k-1}, \cdots, n_{k-N+1})$ to be regarded as a column matrix. Then the output noise samples of the equalizer are:

$$z_k = \mathbf{c}'\mathbf{n}_k . \tag{8}$$

Define the matrix $\mathbf{B} = E(\mathbf{n}_k\mathbf{n}_k')$, where $E(\cdots)$ denotes the expected value. Notice that $\mathbf{B}$ is symmetric and positive semidefinite.

To formulate the iterative equations describing the tap weight behavior in the presence of noise, apply equations (2) and (8) to (7) to show how the gradient estimate depends on the tap weight vector:

$$\hat{\varphi}_i = \sum_l (\mathbf{x}_{l-i\xi} + \mathbf{n}_l)[(\mathbf{x}_{l-i\xi} + \mathbf{n}_l)'\mathbf{c}_i - d_{l-i\xi}].$$

Hence

$$\hat{\varphi}_i = \mathbf{H}_i\mathbf{c}_i - \mathbf{g} - \mathbf{v}_i , \tag{9}$$

where $\mathbf{H}_i$ is the random symmetric matrix

$$\mathbf{H}_k = \sum_l (\mathbf{x}_{l-i\xi} + \mathbf{n}_l)(\mathbf{x}_{l-i\xi} + \mathbf{n}_l)' \tag{10}$$

and

$$\mathbf{v}_i = \sum_l \mathbf{n}_l \, d_{l-i\xi} . \tag{11}$$

Let $\mathfrak{A} = E(\mathbf{H}_i)$, the expected value of $\mathbf{H}_i$. Then equation (10) yields

$$\mathfrak{A} = \mathbf{A} + \kappa\mathbf{B}. \tag{12}$$

which is positive definite since $\mathbf{A}$ is positive definite and $\mathbf{B}$ is positive semidefinite.

It is convenient to examine the random variation of the tap weight vector c_k about the suboptimal setting defined by

$$\tilde{\mathbf{c}} = \mathfrak{A}^{-1}\mathbf{g}, \tag{13}$$

and let $\mathbf{q}_i = \mathbf{c}_i - \tilde{\mathbf{c}}$. From equation (12) it is evident that the suboptimal setting $\tilde{\mathbf{c}}$ approaches the optimal setting $\mathbf{c}^*$ as the ratio of noise variance to input pulse sequence energy approaches zero. The iterative algorithm may be expressed in the form

$$\mathbf{q}_{i+1} = \mathbf{q}_i - \alpha\hat{\varphi}_i , \tag{14}$$

$$\hat{\varphi}_i = \mathbf{H}_i\mathbf{q}_i + \mathbf{h}_i \tag{15}$$

where

$$\mathbf{h}_i = \mathbf{H}_i\tilde{\mathbf{c}} - \mathbf{g} - \mathbf{v}_i \, . \tag{16}$$

Equations (14) and (15) constitute a system of first-order stochastic difference equations with a forcing function $\mathbf{h}_i$ which is statistically dependent on the stochastic state matrix $\mathbf{H}_i$. We assume that the perturbing noise samples in different iterations are uncorrelated, so that $\mathbf{H}_i$ and $\mathbf{h}_i$ are independent of $\mathbf{H}_j$ and $\mathbf{h}_j$ for $i \neq j$. Notice that the expected value of any function of $\mathbf{H}_i$ and $\mathbf{h}_i$ is independent of i. Under these conditions it is proved in Appendix C that for suitably small values of α the mean value of the solution vector $\mathbf{q}_i$ approaches zero as $i \rightarrow \infty$ and the sum of the variances of the components of $\mathbf{q}_i$ is bounded with a bound that approaches zero as α approaches zero. Consequently the mean value of the tap weight vector converges to the suboptimal setting $\tilde{\mathbf{c}}$ while the actual tap weights fluctuate randomly about the converging mean values with a variability that can be made arbitrarily small.

Notice from Appendix C that the norm of the mean solution vector $\langle \mathbf{q} \rangle_i$ is reduced at least by the factor ζ, the spectral norm[20] of $I - \alpha\mathfrak{A}$. Let ρ_1 and ρ_N denote the minimum and maximum eigenvalues, respectively, of $\mathfrak{A}$. Then

$$\zeta = \min \mid 1 - \alpha\rho_1 \mid , \quad \mid 1 - \alpha\rho_N \mid . \tag{17}$$

(For proof see p. 24 of Ref. 20.)

Then for $0 < \alpha < 2/(\rho_1 + \rho_N)$, we obtain $\zeta = 1 - \alpha\rho_1$. Consequently, while decreasing α offers a smaller bound on variability of the tap weight vector, increasing α assures a stronger bound on convergence rate. For the training mode it is likely that speed of adaptation will be relatively unimportant so that a very small value of α could be used to approach a tap weight setting that is very close to the suboptimal setting.

It is useful to obtain bounds on the eigenvalues of $\mathfrak{A}$ which can be determined without specific knowledge of the channel characteristics. If $x(t)$ denotes the channel pulse response and $n(t)$ the additive receiver noise so that the sampled values used earlier are given by $x_k = x(kT)$ and $n_k = n(kT)$, then the sampled spectrum $X^*(\omega)$ of x_k is

$$X^*(\omega) = \sum_k x_k e^{-i\omega kT} = \frac{1}{T} \sum_k X(\omega - 2\pi/T)$$

and the sampled spectral density $S^*(\omega)$ of n_k is

$$S^*(\omega) = \sum_k E(n_i n_{i+k}) e^{-i\omega kT} = \frac{1}{T} \sum_k S(\omega - 2\pi/T)$$

where $X(\omega)$ is the Fourier transform of $x(t)$ and $S(\omega)$ is the spectral density of $n(t)$. Let m and M denote the infimum and supremum, respectively, of $| X^*(\omega) |^2 + \kappa S^*(\omega)$ so that

$$m \leqq | X^*(\omega) |^2 + \kappa S^*(\omega) \leqq M. \tag{18}$$

In all cases of practical interest M will be finite; furthermore generally m will be greater than zero. It is shown in Appendix B that each eigenvalue, ρ_i, of $\mathbf{\mathfrak{a}}$ will be bounded according to

$$m \leqq \rho_i \leqq M. \tag{19}$$

To illustrate the use of this bound, notice from Appendix C that the condition for convergence of the mean tap weight vector to the suboptimal solution is that $\alpha < 2/\rho_N$. Thus a sufficient condition is that

$$\alpha < 2/M. \tag{20}$$

Furthermore, the mean tap setting converges exponentially with the convergence factor ζ, given by equation (17). Hence it can be inferred that the choice of α which provides the strongest bound (least value of ζ) is $\alpha = 2/(\rho_1 + \rho_N)$ yielding

$$\zeta = \frac{p - 1}{p + 1}$$

where $p = \rho_N/\rho_1$. Using the bounds given in (19) we obtain $p \leqq M/m$, and so

$$\zeta = \frac{M - m}{M + m}. \tag{21}$$

Therefore, for the best choice of α, convergence of the mean proceeds at least at a rate given by the geometric factor $(M - m)/(M + m)$. Thus useful information regarding the convergence speed can be determined without knowledge of the channel characteristics.

V. CONCLUSION

The degree of suboptimality of the tap weight setting reached by the training algorithm may or may not be consequential, depending on the application. In applications where multilevel pulse transmission with a large number of levels could be achieved with adequate equalization, the signal-to-noise ratio is necessarily very high and therefore the degree of suboptimality is not large. Even when the noise level is fairly substantial the suboptimal setting may still be adequate if the error surface given by $\mathcal{E}(\mathbf{c})$ is "shallow" in a large neighborhood of the minimum. Then a fairly large departure of $\tilde{\mathbf{c}}$ from $\mathbf{c}^*$ may correspond to a relatively small increase in mean-square error. Also, if training mode adaptation is used as a prelude to tracking mode adaptation, a fairly large degree of suboptimality may be a tolerable starting point for a tracking mode operation such as the one we plan to describe in a future paper.

When the noise level is substantial the criterion for optimality used here becomes inadequate because it does not consider the effect of the equalizer on the receiver noise. The price of reducing intersymbol interference may be a sizable increase in noise level at the equalizer output. In our future paper the error criterion is modified to include noise with the result that the problem of suboptimality does not arise.

The random fluctuation of the tap weights which prevents true convergence to the suboptimal setting can be eliminated by reducing the proportionality constant α in each iteration using a sequence of step sizes α_k with the properties

$$\sum \alpha_k = \infty \qquad \text{and} \qquad \sum \alpha_k^2 < \infty .$$

It may then be shown that the tap weight vector converges to the suboptimal solution with probability 1. The proof uses stochastic approximation theory and follows the lines taken by Tong and Liu who considered a training mode algorithm for low dispersion channels.[21] However, this modification complicates the implementation somewhat and cannot be applied to the tracking mode adaptation problem.

APPENDIX A

Proof that **A** *is Positive Definite*

The matrix **A** is defined by

$$\mathbf{A} = \sum_{k=-\infty}^{\infty} \mathbf{x}_k \mathbf{x}_k' . \tag{22}$$

Consequently

$$\mathbf{c}'\mathbf{A}\mathbf{c} = \sum_{-\infty}^{\infty} \mathbf{c}'\mathbf{x}_k\mathbf{x}_k'\mathbf{c} = \sum_{-\infty}^{\infty} y_k^2 .$$

But the sequence y_k is the convolution of the x_k sequence with the finite tap weight sequence c_k . Hence, using Parseval's equality,

$$\mathbf{c}'\mathbf{A}\mathbf{c} = \frac{1}{2\pi} \int_{-\pi/T}^{\pi/T} | X^*(\omega) |^2 \, | C(\omega) |^2 \, d\omega, \tag{23}$$

where

$$C(\omega) = \sum_{k=0}^{N-1} c_k e^{-jk\omega T}.$$

Equation (23) shows immediately that $\mathbf{c}'\mathbf{A}\mathbf{c}$ is nonnegative for all vectors $\mathbf{c}$. Also, $C(\omega)$ can have only isolated zeros and $| X^*(\omega) |$ is square integrable since the input pulse has finite mean square distortion. It may then be inferred that $\mathbf{c}'\mathbf{A}\mathbf{c} > 0$ unless $\mathbf{c} = 0$, which proves that $\mathbf{A}$ is positive definite.

APPENDIX B

Bounds on the Eigenvalues of $\mathfrak{A}$

Since $\mathbf{B} = E(\mathbf{n}_k\mathbf{n}_k')$ the quadratic form $\mathbf{c}'\mathbf{B}\mathbf{c}$ is the mean squared value of $\mathbf{y}_k^2$ of the response of the equalizer with weight vector $\mathbf{c}$ to the input noise n_t . Consequently

$$\mathbf{c}'\mathbf{B}\mathbf{c} = \frac{1}{2\pi} \int_{-\pi/T}^{\pi/T} S^*(\omega) \mid C(\omega) \mid^2 d\omega. \tag{24}$$

Combining equations (23) and (24) yields

$$\mathbf{c}'\mathfrak{A}\mathbf{c} = \frac{1}{2\pi} \int_{-\pi/T}^{\pi/T} \{\mid X^*(\omega) \mid^2 + \kappa S^*(\omega)\} \mid C(\omega) \mid^2 d\omega. \tag{25}$$

Applying to equation (25) the bounds m and M given by equation (18) yields

$$m\, \mathbf{c}'\mathbf{c} \leqq \mathbf{c}'\mathfrak{A}\mathbf{c} \leqq M\mathbf{c}'\mathbf{c}. \tag{26}$$

Let $\mathbf{c}$ be the eigenvector of $\mathfrak{A}$ corresponding to eigenvalue ρ. Then $\mathfrak{A}\mathbf{c} = \rho_i\mathbf{c}$ and equation (26) yields

$$m \leqq \rho_i \leqq M \tag{27}$$

which provides a convenient bound for the largest and smallest eigenvalues of $\mathfrak{A}$.

APPENDIX C

Convergence Proof

To examine the convergence properties of the tap weight adjustment algorithm, it is convenient to define the norm of a random vector $\mathbf{u}$ as

$$\| \mathbf{u} \| = [E(\mathbf{u}'\mathbf{u})]^{1/2}, \tag{28}$$

so that the squared norm of $\mathbf{u}$ is the sum of the second moments of the components of $\mathbf{u}$. For a deterministic vector the norm reduces to the usual Euclidian norm. The norm of a deterministic matrix will denote the usual spectral norm.[20]

Theorem: *Let* $\mathbf{H}_k$ *be a sequence of random symmetric* $N \times N$ *matrices and* $\mathbf{h}_k$ *a sequence of random N-tuple column vectors. Suppose* $\mathbf{H}_k$ *and* $\mathbf{h}_k$ *are stationary in k with* $\mathbf{H}_k$ *and* $\mathbf{h}_k$ *independent of* $\mathbf{H}_j$ *and* $\mathbf{h}_j$ *for* $k \neq j$. *Assume* $\mathbf{h}_k$ *has zero mean, and the elements of* $\mathbf{H}_k$ *and* $\mathbf{h}_k$ *have finite variance,* $E\mathbf{H}_k = \boldsymbol{\mathcal{A}}$, *independent of k with* $\mathcal{A}$ *positive definite. Define the random vector sequence* $\mathbf{q}_k$ *according to*:

$$\mathbf{q}_{k+1} = \mathbf{q}_k - \alpha \varphi_k \tag{29}$$

where

$$\varphi_k = \mathbf{H}_k \mathbf{q}_k + \mathbf{h}_k \tag{30}$$

for $k = 0, 1, 2, \cdots$ *and* $\mathbf{q}_0$ *is an arbitrary deterministic vector. Then for* α *positive and sufficiently small,*

$$\lim_{k \to \infty} \| E\, \mathbf{q}_k \| = 0 \tag{31a}$$

and

$$\limsup_{k \to \infty} \| \mathbf{q}_k \| \leqq V(\alpha) \tag{31b}$$

with $V(\alpha)$, *given in* (47), *satisfying*:

$$\lim_{\alpha \to 0} V(\alpha) = 0. \tag{32}$$

Proof: Combining equations (29) and (30) yields

$$\mathbf{q}_{k+1} = (\mathbf{I} - \alpha \mathbf{H}_k)\mathbf{q}_k - \alpha \mathbf{h}_k \, . \tag{33}$$

Noting that $\mathbf{q}_k$ is independent of $\mathbf{H}_k$, taking the expected value in equation (33), we find

$$E(\mathbf{q}_{k+1}) = (\mathbf{I} - \alpha \boldsymbol{\mathcal{A}}) E(\mathbf{q}_k). \tag{34}$$

It follows then that

$$\| E(\mathbf{q}_k) \| \leqq \zeta^k \| E\, \mathbf{q}_0 \| \tag{35}$$

where

$$\zeta = \| I - \alpha \boldsymbol{\mathcal{A}} \| \, . \tag{36}$$

Hence equation (31a) follows when $\zeta < 1$, or equivalently, for

$$0 < \alpha < 2/\rho_N \tag{37}$$

where ρ_N is the largest eigenvalue of $\mathcal{A}$.

To prove equation (31b), observe that

$$E(\mathbf{q}_{k+1}'\mathbf{q}_{k+1}) = E[\mathbf{q}_k'(I - \alpha H_k)^2\mathbf{q}_k] - E[2\alpha\mathbf{q}_k'(I - \alpha H_k)\mathbf{h}_k] + \alpha^2 \| \mathbf{h}_k \|^2 \tag{38}$$

from equation (33). Noting again that $\mathbf{q}_k$ is independent of H_k, we have

$$E[\mathbf{q}_k'(I - \alpha\mathbf{H}_k)^2\mathbf{q}_k] = E\{\mathbf{q}_k'E[(I - \alpha\mathbf{H}_k)^2]\mathbf{q}_k\} \leqq \mu \| \mathbf{q}_k \|^2, \tag{39}$$

where

$$\mu = \| E[(I - \alpha\mathbf{H}_k)^2] \| . \tag{40}$$

Also, using the Schwarz inequality,

$$E[-\mathbf{q}_k'(I - \alpha\mathbf{H}_k)\mathbf{h}_k] = \alpha\mathbf{q}_k'E(\mathbf{H}_k\mathbf{h}_k) \leqq \alpha \| \mathbf{q}_k' \| f$$

where $f = \| E(\mathbf{H}_k\mathbf{h}_k) \|$. Using equation (35) we obtain

$$-E[\mathbf{q}_k(I - \alpha\mathbf{H}_k)\mathbf{h}_k] \leqq \alpha\zeta^k \| E(\mathbf{q}_0) \| f. \tag{41}$$

The bounds (39) and (41) may be applied to equation (38), yielding

$$\| \mathbf{q}_{k+1} \|^2 \leqq \mu \| \mathbf{q}_k \|^2 + \alpha^2 f \| E \ \mathbf{q}_0 \| \zeta^k + \alpha^2 \| \mathbf{h}_k \|^2 . \tag{42}$$

If we now define the bounding sequence of positive numbers Q_k according to

$$Q_0 = \| E \ \mathbf{q}_0 \|^2$$

and

$$Q_{k+1} = \mu Q_k + \alpha^2 f \| E(\mathbf{q}_0) \| \zeta^k + \alpha^2 \|\mathbf{h}_k \|^2 , \tag{43}$$

then it follows from (42) that

$$\| \mathbf{q}_k \|^2 \leqq Q_k .$$

But the difference equation given by (43) has the asymptotic solution

$$\lim_{k\to\infty} Q_k = \frac{\alpha^2 \| \mathbf{h}_k \|^2}{1 - \mu} ,$$

for $\zeta < 1$ and $\mu < 1$. Then

$$\limsup_{k\to\infty} \| q_k \|^2 \leqq \frac{\alpha^2 \| \mathbf{h}_k \|^2}{1 - \mu}. \tag{44}$$

Notice that $|| h_k ||$ is independent of k by the hypothesis of stationarity.

Since

$$(I - \alpha \mathbf{H}_k)^2 = (I - \alpha \mathfrak{A})^2 + \alpha^2 E(\mathbf{G}_k^2).$$

where

$$\mathbf{G}_k = \mathbf{H}_k - \mathfrak{A}, \tag{45}$$

we find that

$$\begin{aligned} \mu &\leqq || \, I - \alpha \mathfrak{A} \, ||^2 + \alpha^2 \, || \, E(\mathbf{G}_k^2) \, || \\ \mu &\leqq \zeta^2 + \alpha^2 \gamma \end{aligned} \tag{46}$$

where $\gamma = || \, \mathbf{G}_k^2 \, ||$. Furthermore for $\alpha < 2/(\rho_1 + \rho_N)$, we have $\zeta = 1 - \alpha \rho_1$. Then, using (46), we see that

$$\frac{\alpha^2}{1 - \mu} \leqq \frac{\alpha^2}{2\alpha\rho_1 + \alpha^2(\rho_1^2 + \gamma)}.$$

We have therefore shown that for positive and sufficiently small α, equations (31b) and (32) are valid where

$$V(\alpha) = \frac{\alpha}{2\rho_1 + \alpha^2(\rho_1^2 + \gamma)}. \tag{47}$$

REFERENCES

1. Lucky, R. W., "Automatic Equalization for Digital Communication," B.S.T.J., *44,* No. 4 (April 1965), pp. 547–588.
2. Lucky, R. W., "Techniques for Adaptive Equalization of Digital Communication Systems," B.S.T.J., *45,* No. 2 (February 1966), pp. 255–286.
3. Widrow, B., and Hoff, M. E., Jr., "Adaptive Switching Circuits," IRE Wescon Conv. Record, Pt. 4 (August 1960), pp. 96–104.
4. Narenda, K. S. and McBride, L. E., "Multiparameter Self-Optimizing Systems Using Correlation Techniques," IEEE Trans. Automatic Control, *AC-9,* No. 1 (January 1964), pp. 31–38.
5. Lucky, R. W. and Rudin, H. R., Generalized Automatic Equalization for Communication Channels", Proc. IEEE (Letters), *54,* No. 3, pp. 439–40, March 1966.
6. Lucky, R. W. and Rudin, H. R. An Automatic Equalizer for General-Purpose Communication Channels, B.S.T.J., *46,* No. 9 (November 1967), pp. 2179–2208.
7. Gersho, A., unpublished work.
8. Lytle, D. W., unpublished work.
9. Niessen, C. W., "Automatic Channel Equalization Algorithm," Proc. IEEE, *55,* No. 5 (May 1967), p. 689.
10. Niessen, C. W., and Drouilhet, P. R., "Adaptive Equalizer for Pulse Transmission," Conf. Digest, IEEE Int. Conf. on Commun. (June 1967), p. 117.
11. Boothroyd, W. P. and Creamer, E. M., Jr., "A Time Division Multiplexing System," Trans. AIEE, *68,* Pt. I, (1949) pp. 92–97.
12. Tufts, D. W., "Nyquist's Problem—the Joint Optimization of Transmitter

and Receiver in Pulse Amplitude Modulation," Proc. IEEE, *53,* No. 3 (March 1965), pp. 248–259.

13. George, D. A., "Matched Filters for Interfering Signals," IEEE Trans. Inform. Theory *IT-11,* No. 1 (January 1965), pp. 153–154.
14. Aaron, M. R. and Tufts, D. W., "Intersymbol Interference and Error Probability," IEEE Trans. Inform. Theory *IT-12,* No. 1 (January 1966), pp. 26–34.
15. Koford, J. S., and Groner, G. F., "The Use of an Adaptive Threshold Element to Design a Linear Optimal Pattern Classifier," IEEE Trans. Inform. Theory, *IT-12,* No. 1 (January 1966), pp. 42–50.
16. Widrow, B., "Adaptive Filters I: Fundamentals," Stanford Electronics Lab., Stanford, Calif., Report SEL-60-126 (Technical Report 6764-6), December 1966.
17. Coll, D. C. and George, D. A., "A Receiver for Time-Dispersed Pulses, Conf. Record, 195, IEEE Annual Commun. Conv., pp. 753–757.
18. Gersho, A., "Automatic Equalization of Highly Dispersive Channels for Data Transmission," Int. Symp. on Inform. Theory, San Remo, Italy, September 1967.
19. Gersho, A. and Freeny, S. L., "Performance Capabilities of Transversal Equalizers for Digital Communication," Conf. Digest, IEEE Int. Conf. on Commun. 1967, p. 88.
20. Goldstein, A. A., *Constructive Real Analysis,* New York: Harper and Row, 1967, Chap. 1.
21. Tong, P. S. and Liu, B., "Automatic Time Domain Equalization in the Presence of Noise," Proc. Nat. Elec. Conf., *23,* (October 1967), pp. 262–266.
22. Varga, R. S., *Matrix Iterative Analysis,* New York: Prentice-Hall, 1962.

On the Probability of Error Using a Repeat-Request Strategy on the Additive White Gaussian Noise Channel

By A. D. WYNER

(Manuscript received July 5, 1968)

An upper bound on the error probability is obtained for digital communication (with average power P_o and no bandwidth constraint) in the presence of additive white gaussian noise (with one-sided spectral density N_o) with the use of a noiseless feedback link. A repeat-request strategy is used: the receiver decodes a signal only when it is relatively sure that one particular message was actually transmitted, otherwise it requests (via the feedback channel) a retransmission. We show that as the coding delay T becomes large, we can transmit at an effective rate $\bar{R} < C = P_o/N_o$, the channel capacity, with error probability P_e approximately exp $\{-T[(\sqrt{C} - \sqrt{\bar{R}})^2 + C - \bar{R}]\}$, which is a considerable improvement over the reliability attainable with a one-way channel. These results parallel those obtained earlier by Forney for the discrete memoryless channel.

I. INTRODUCTION

In a recent paper, Forney studied a repeat-request strategy for communication of digital information over a discrete memoryless channel when a fedback channel is available.[3] In this system the receiver decodes a received message only when it is relatively "sure" that one particular message was actually transmitted. If the receiver is not confident that one particular message was actually transmitted, then it requests (via the feedback channel) that the transmitter repeat the message. Forney showed that considerable improvement in the resulting error probability (over the best one-way scheme) was obtainable with a negligible degradation in the effective rate of transmission. In this paper we apply Forney's ideas to the additive white Gaussian noise channel (with no bandwidth constraint) and obtain analogous results. Furthermore, our coding scheme is constructive—the codes being orthogonal codes.

We will consider the following channel. The channel input signal is a real-valued function $s(t)$, defined on the interval $[0, T]$, which satisfies the "energy" constraint

$$\int_0^T s^2(t)\, dt = P_0 T. \tag{1}$$

The average signal "power" is therefore P_o. The channel output $r(t)$ is the sum of $s(t)$ and a sample $n(t)$ from a white Gaussian noise process with one-sided spectral density N_o (and with mean zero). By expanding $s(t)$, $r(t)$ and $n(t)$ on any orthonormal basis of $\mathfrak{L}_2[0, T]$, it is easy to show that an equivalent channel model is as follows.[8,9] (This equivalent channel model is the one we use in this paper.) The input signals are are (semi-infinite) vectors $\mathbf{x} = (x_1, x_2, \cdots)$ which satisfy

$$\sum_{k=1}^{\infty} x_k^2 = AT. \tag{2}$$

The channel output is a vector $\mathbf{y} = (y_1, y_2, \cdots)$, where

$$y_k = x_k + z_k, \qquad k = 1, 2, \cdots,$$

and the $z_k (k = 1, 2, \cdots)$ are independent Gaussian variates with zero mean and unit variance. The parameter A is equal to $2P_o/N_o$, and we assume that A is held fixed throughout the paper. We also assume that it takes T seconds for the channel to process $\mathbf{x}$, and that successive T-second transmissions are independent.

A *code* with parameters M and T is a set of M signals (called "code vectors" or "code words") $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots)$, $i = 1, 2, \cdots, M$, which satisfy equation (2), that is

$$\sum_{k=1}^{\infty} x_{ik}^2 = AT, \qquad i = 1, 2, \cdots, M. \tag{3}$$

We assume that each of the M code words is equally likely to be transmitted, so that the *transmission rate* is $R = 1/T \ln M$ nats (natural units) per second, and $M = e^{RT}$. It is the task of the receiver to examine the channel output $\mathbf{y}$ and to announce the code word, say $D(\mathbf{y})$, which it believes was actually transmitted. Let P_{ei} be the probability that $D(\mathbf{y}) \neq \mathbf{x}_i$ given that $\mathbf{x}_i$ is transmitted. The overall error probability is therefore

$$P_e = \frac{1}{M} \sum_{i=1}^{M} P_{ei}.$$

It is easy to show that for a given code, the "optimal" decoding rule D

(which minimizes P_e) selects for $D(\mathbf{y})$ that code word $\mathbf{x}_i$ which maximizes (with respect to i) the inner product

$$\langle \mathbf{x}_i , \mathbf{y} \rangle = \sum_{k=1}^{\infty} x_{ik} y_k$$

Define $P_e^*(M, T)$ as the smallest attainable error probability P_e for a code with parameters M and T. Set $M = [e^{RT}]$, and let $T \to \infty$ with the rate R held fixed. Then it is well known that if $R < A/2 = P_o/N_o \triangleq C$, the "channel capacity,"

$$P_e^*([e^{RT}], T) = \exp \{-E_0(R)T[1 + \epsilon_0(T)]\}, \tag{4}$$

where $E_0(R) > 0$, and $\epsilon_0(T) \to 0$ as $T \to \infty$.[1,8,9] Thus at rates $R < C$, the error probability tends to zero exponentially in T. Further, for rates $R > C$, $P_e^*([e^{RT}], T) \to 1$, so that the capacity C is the supremum of the rates for which "error-free" coding is possible.

Although this type of behavior of P_e^* is typical of a large class of channels, the present channel is unique in two ways. First the exponent $E_0(R)$ is known exactly, namely

$$E_0(R) = \begin{cases} C/2 - R, & 0 \leqq R \leqq C/4, \\ [C^{\frac{1}{2}} - R^{\frac{1}{2}}]^2, & C/4 \leqq R \leqq C. \end{cases} \tag{5}$$

Second, an explicit construction of codes which achieve error probability as in equation (4) is known. In fact, P_e as in equations (4) and (5) can be achieved when the code is any set of M orthogonal vectors. The simplest such code is that for which x_{ik} (the kth coordinate of $\mathbf{x}_i$) is given by

$$x_{ik} = \begin{cases} (AT)^{\frac{1}{2}}, & k = i, \\ 0, & k \neq i, \end{cases} \quad i = 1, 2, \cdots, M, \quad k = 1, 2, \cdots . \tag{6}$$

For this orthogonal code, the inner product of $\mathbf{y}$ and the ith code word is

$$\langle \mathbf{x}_i , \mathbf{y} \rangle = y_i (AT)^{\frac{1}{2}}, \qquad i = 1, 2, \cdots, M;$$

so that the optimal decoding rule is

$$D(\mathbf{y}) = \mathbf{x}_i \quad \text{if} \quad y_i > y_j \qquad \text{for all } j \neq i, \qquad 1 \leqq j \leqq M. \tag{7}$$

With probability one, (7) is satisfied for exactly one i. Notice that the coordinates $y_j (j > M)$ are irrelevant to the receiver. Further, from the symmetry of the orthogonal code (6), we can without loss of generality, assume that code word $\mathbf{x}_1$ is transmitted. Hence, the error probability is

$$P_e = P_{e1} = \Pr \bigcup_{j=2}^{M} \{y_1 \leqq y_j\}, \tag{8}$$

where the probability is computed with $\{y_i\}_1^M$ independent unit variance Gaussian random variables with $Ey_1 = (AT)^{\frac{1}{2}}$ and $Ey_j = 0$ $(2 \leqq j \leqq M)$.

Now suppose we can use a noiseless feedback link. As before, we transmit one of a set of $M = e^{RT}$ orthogonal signals $\{\mathbf{x}_i\}_1^M$, where $\mathbf{x}_i$ is given by (6). Instead of the decoding rule (7), let us use the rule

$$D(\mathbf{y}) = \mathbf{x}_i \text{ if } y_i > y_j + \Delta \quad \text{for all } j \neq i, \quad 1 \leqq j \leqq M, \tag{9}$$

where $\Delta > 0$ will be chosen later. If no y_i satisfies (9) then we request a retransmission via the feedback channel, and use (9) on the second received vector, and so on. The probability of error decreases as Δ increases. The price which we pay for this increased reliability is an increase in the length of time which it will take to complete the transmission of the M-ary message, and the consequential reduction in the effective rate of transmission. In fact, let E_R be the event that we ask for a retransmission, and let $P(E_R)$ be its probability. Then from the assumption that successive transmissions are independent, the expected number of T-second transmissions required to accept a message is

$$\begin{aligned}\sum_{j=1}^{\infty} j \Pr \{j \text{ transmissions are required}\} \\ = \sum_{j=1}^{\infty} j[1 - P(E_R)][P(E_R)]^{j-1} = [1 - P(E_R)] \sum_{j=1}^{\infty} jP(E_R)^{j-1} \\ = [1 - P(E_R)] \cdot \frac{1}{[1 - P(E_R)]^2} = \frac{1}{1 - P(E_R)}.\end{aligned}$$

Thus the average length of time required to transmit the M-ary message is $\bar{T} = T/(1 - P(E_R))$. If $P(E_R)$ is small, then $\bar{T}$ is not much greater than T.

Suppose that we use this repeat-request strategy repeatedly—that is, if the receiver does not call for a retransmission, then the transmitter sends a new M-ary message. For $k = 1, 2, \cdots$, let the random variable N_k be the number of M-ary messages which the receiver accepts (that is, it does not call for a retransmission) in kT seconds. Then we can write

$$N_k = \sum_{j=1}^{k} \xi_j ,$$

where the random variables $\xi_j = 1$ if the receiver accepts a message on the jth T-second interval, and $\xi_j = 0$ otherwise. Note that $Pr\{\xi_j = 0\} = P(E_R)$, and that the $\{\xi_j\}_1{}^k$, are independent (since we have assumed that successive T-second transmissions are independent). Thus

$$\begin{aligned} &(i)\quad E(N_k) = kE(\xi_j) = k(1 - P(E_R)) \\ &(ii)\quad N_k/k \to 1 - P(E_R), \qquad \text{as } k \to \infty, \\ &\qquad\qquad \text{with probability 1.} \end{aligned} \tag{10}$$

Statement (ii) follows from the strong law of large numbers (see Ref. 3, p. 190). Since each M-ary message contains $\ln M = RT$ nats, the effective rate of transmission $\bar{R}$, in the light of (10),

$$\begin{aligned} \bar{R} &= \frac{[E(N_k)]RT}{kT} \text{ nats/sec} \\ &= R[1 - P(E_R)] = R(T/\bar{T}). \end{aligned} \tag{11}$$

Let us turn our attention to the probability of error. Since we are using the orthogonal code of (6), we can, as above, without loss of generality, assume that code word $\mathbf{x}_1$ is transmitted. Using the decoding rule of equation (1.9) we make an error only when for some $j > 1$, $y_j > y_i + \Delta$ for *all* $i = 1, 2, \cdots, M$ and $i \neq j$. (In this case $D(\mathbf{y}) = \mathbf{x}_j$.) Thus the error probability is

$$P_e = \Pr \bigcup_{j=2}^{M} \bigcap_{i \neq j} \{y_j > y_i + \Delta\}. \tag{12}$$

As in (8), the probability in equation (12) is computed with $Ey_1 = (AT)^{\frac{1}{2}}$ and $Ey_j = 0$ $(2 \leqq j \leqq M)$.

Let us further define E_1 as the event that either an error occurs or a repeat-request occurs. If $\mathbf{x}_1$ is transmitted, E_1 has probability

$$\Pr(E_1) = \Pr \bigcup_{j=2}^{M} \{y_1 \leqq y_j + \Delta\}, \tag{13}$$

where as above, the probability in (13) is computed with $Ey_1 = (AT)^{\frac{1}{2}}$ and $Ey_j = 0$, $j > 1$. Clearly the probability of a repeat-request is

$$P(E_R) = P(E_1) - P_e \leqq P(E_1). \tag{14}$$

Consider the parameter Δ. In the interest of minimizing P_e, we want to make Δ large. However, in the interest of minimizing $P(E_R)$ and therefore making $\bar{R}$ as close to R as possible, we want to make Δ small. The approach which we will take is to choose Δ just small enough so that as the parameter $T \to \infty$ (R is held fixed), $P(E_1) \to 0$; so that by (14), $P(E_R) \to 0$. Thus the effective transmission rate $\bar{R} \approx R$. We will see that this results in a considerable improvement in P_e over that of equations (4) and (5). Roughly speaking, we will show that the resulting exponent is increased from that in equation (5) to approximately

$$E_F(R) = [C^{\frac{1}{2}} - R^{\frac{1}{2}}]^2 + C - R = 2C^{\frac{1}{2}}(C^{\frac{1}{2}} - R^{\frac{1}{2}}). \tag{15}$$

The exponents $E_0(R)$ and $E_F(R)$ are plotted in Fig. 1. Notice that the improvement is greatest in the neighborhood of capacity where (as $R \to C)E_F(R) \approx (C - R)$ and $E_0(R) \sim (C - R)^2/4C$.

II. SUMMARY AND DISCUSSION OF RESULTS

The main result is given as a corollary to the following two theorems which provide information on the trade-off between P_e and $P(E_1)$ as Δ is varied. The proofs are given in Section III.

Theorem 1: Let $\{y_i\}_1^M$, *be independent Gaussian random variables with unit variance and expectation*

$$\begin{aligned} Ey_1 &= (AT)^{\frac{1}{2}}, \\ Ey_j &= 0, \qquad 2 \leqq j \leqq M. \end{aligned} \tag{16}$$

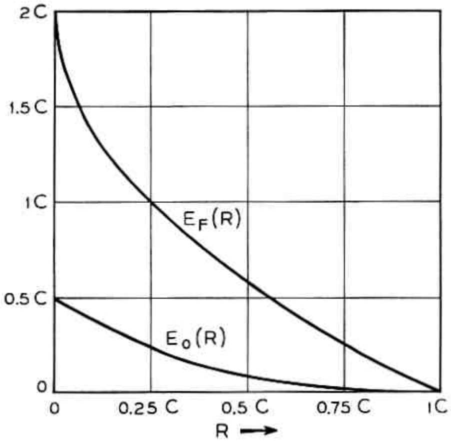


Fig. 1 — Exponents for white Gaussian noise channel: $E_0(R)$-one way exponent, $E_F(R)$-repeat-request exponent.

Let $M = e^{RT}$, where $0 < R < A/2 = C$, and let $\Delta = \delta(2T)^{\frac{1}{2}}$, where

$$C^{\frac{1}{2}} - (4R)^{\frac{1}{2}} \leqq \delta < C^{\frac{1}{2}} - R^{\frac{1}{2}}. \tag{17}$$

Then

$$P(E_1) = \Pr \bigcup_{j=2}^{M} \{y_1 \leqq y_j + \Delta\} \leqq 2 \exp \{-[C^{\frac{1}{2}} - R^{\frac{1}{2}} - \delta]^2 T\}. \tag{18}$$

Notice that $\delta = 0$ will satisfy (17) if $R \geqq C/4$. In this case $P(E_1) = P_e$ (see (8)), and (18) yields $E_0(R) \geqq [C^{\frac{1}{2}} - R^{\frac{1}{2}}]^2 (C/4 \leqq R \leqq C)$, a fact which is contained in (5). In fact, the proof of Theorem 1 closely parallels the derivation of P_e for orthogonal codes (for a one-way channel).

Theorem 2: Let $\{y_i\}_1^M$, be independent gaussian random variables with unit variance and expectation

$$Ey_1 = (AT)^{\frac{1}{2}},$$
$$Ey_j = 0, \qquad 2 \leqq j \leqq M. \tag{19}$$

Let $M = e^{RT}$, where $0 \leqq R < A/2 = C$, and let $\Delta = \delta(2T)^{\frac{1}{2}}$, where

$$\delta > C^{\frac{1}{2}} - (4R)^{\frac{1}{2}}. \tag{20}$$

With R and δ held fixed, and θ_1 , θ_2 arbitrary but satisfying

$$\theta_1 > 0, \tag{21a}$$

$$0 < \theta_2 < \begin{cases} R^{\frac{1}{2}} \\ \dfrac{\delta - [C^{\frac{1}{2}} - (4R)^{\frac{1}{2}}]}{2} \end{cases}, \tag{21b}$$

then for T sufficiently large,

$$P_e = \Pr \bigcup_{j=2}^{M} \bigcap_{i \neq j} \{y_j > y_i + \Delta\}$$
$$\leqq 2(1 + \theta_1) \exp \{-[(R^{\frac{1}{2}} + \delta - \theta_2)^2 + (C^{\frac{1}{2}} - R^{\frac{1}{2}} + \theta_2)^2 - R]T\}. \tag{22}$$

Again notice that $\delta = 0$ will satisfy (20) if $R > C/4$. In this case also, (22) yields $E_0(R) \geqq [C^{\frac{1}{2}} - R^{\frac{1}{2}}]^2$, when $R > C/4$ (since θ_2 can be made arbitrarily small).

Let us now use these theorems to find the value of $\Delta = \delta(2T)^{\frac{1}{2}}$ which gives the smallest upper bound on P_e without substantially changing the effective rate $\bar{R} \geqq R[1 - P(E_1)]$. Since P_e is a decreasing function of δ, we choose δ as large as possible with the proviso that $P(E_1) \rightarrow 0$.

From Theorem 1, this value of δ is

$$\delta = C^{\frac{1}{2}} - R^{\frac{1}{2}} - \gamma_1 , \tag{23}$$

where $\gamma_1 > 0$. If γ_1 is sufficiently small, this choice of δ satisfies (17) and (20). With δ so chosen, for any $\gamma_2 > 0$ we can find a T sufficiently large so that $\bar{R} \geqq R(1 - \gamma_2)$. Further, substitution of equation (23) into equation (22) yields an exponent

$$-[C^{\frac{1}{2}} - \gamma_1 - \theta_2)^2 + (C^{\frac{1}{2}} - R^{\frac{1}{2}} + \theta_2)^2 - R]T.$$

Finally, since γ_1, γ_2, θ_1 and θ_2 can be made arbitrarily small we have our main result:

Corollary: Let $\theta_1 > 0$, $\epsilon > 0$ be arbitrary. Let $\bar{R} < C$. Then for T sufficiently large, there is a repeat-request communication system using orthogonal codes with an effective rate of $\bar{R}$ and error probability

$$P_e \leqq 2(1 + \theta_1) \exp \{-[(C^{\frac{1}{2}} - \bar{R}^{\frac{1}{2}})^2 + C - \bar{R} - \epsilon]T\}.$$

Let us turn our attention to (4) and (5) which give the error probability for the one-way Gaussian channel. The fact that $E_0(R) \leqq (C^{\frac{1}{2}} - R^{\frac{1}{2}})^2$ can be demonstrated by a "sphere-packing" argument.[9] This argument states that $P_e^*(M, T) \geqq Q$, where Q is the probability of error which would result if it were possible to subdivide Euclidean M-space into M congruent cones (each with apex at the origin), one for each code word, and each code word were placed on the axis of its cone at a distance $(AT)^{\frac{1}{2}}$ from the origin. Setting the "sphere-packing exponent"

$$E_{SP}(R) = (C^{\frac{1}{2}} - R^{\frac{1}{2}})^2,$$

we have from the above corollary that for effective transmission rates $\bar{R} < C$ we can obtain an error exponent arbitrarily close to

$$E_F(\bar{R}) = E_{SP}(\bar{R}) + C - \bar{R}. \tag{24}$$

For discrete memoryless channels it is possible to find a lower bound to the optimal (one-way) error probability using an analogous sphere-packing argument.[7] Forney showed that using a repeat-request strategy similar to the one used here, one can obtain an error exponent arbitrarily close to that of equation (24) [with the appropriate $E_{SP}(\bar{R})$].[3] Forney also studied the so called (discrete) "very noisy channel," which is closely related to our Gaussian channel* and obtained results similar

* Our Gaussian channel may be thought of as a "very noisy channel" since the signal-to-noise ratio per coordinate is zero.

to our results. Thus, in the light of Forney's results, the above corollary is not surprising.

Let us also remark that Kramer has found a scheme for our white noise channel with a feedback link that attains an error exponent of $C - R$, which is less than that in equation (24).[4] In Kramer's scheme, the receiver observes the signal until it is sufficiently confident that one particular message was actually transmitted. It then informs the transmitter, via the feedback channel, to start the next M-ary transmission, thereby using the feedback channel only once per M-ary message. In the repeat-request scheme studied here, the number of uses of the feedback channel per M-ary transmission is an unbounded random variable. Thus the two schemes, while similar (in that the feedback channel is used only to convey a "decision"), are not directly comparable. On the other hand, there are schemes which use the feedback channel considerably more heavily (so called "information feedback") which in some cases attain somewhat better performance than the repeat-request strategy. (See for example Refs. 5, 6, and 10).

Finally, an important problem which has been completely ignored here is the requirement that the transmitter have a buffer in which it can store data which will accumulate at the transmitter at times when the receiver asks for retransmissions. If the buffer has finite capacity, it will occasionally overflow, introducing a further source of errors. Some quantitative results on this problem have been obtained by the author, and will be reported in a future paper.

III. PROOFS OF THEOREMS

We begin with some definitions. Let

$$g(\alpha) = \frac{1}{(2\pi)^{\frac{1}{2}}} \exp(-\alpha^2/2), \quad -\infty < \alpha < \infty,$$

be the standard Gaussian density, and let

$$\Phi(u) = \int_{-\infty}^{u} g(\alpha)\, d\alpha, \qquad -\infty < u < \infty,$$

be the cumulative error function, and let

$$\Phi_c(u) = \int_{u}^{\infty} g(\alpha)\, d\alpha = 1 - \Phi(u), \qquad -\infty < u < \infty,$$

be the complementary error function. Let $b = (AT)^{\frac{1}{2}} = (2CT)^{\frac{1}{2}}$ so that y_1 has density $g(\alpha - b)$ and y_j $(2 \leqq j \leqq M)$ has density $g(\alpha)$. We will use the following

Lemma 1: For $u \geqq 0, \Phi_c(u) \leqq \exp(-u^2/2)$; *and for* $u \leqq 0, \Phi_c(u) \leqq \exp(-u^2/2)$ (Wozencraft and Jacobs Ref. 8):

Proof: For $u \geqq 0$,

$$[\Phi_c(u)]^2 = \int_u^\infty \int_u^\infty g(\alpha)g(\beta)\, d\alpha\, d\beta \leqq \int_{\mathcal{R}} \int g(\alpha)g(\beta)\, d\alpha\, d\beta = \frac{\exp(-u^2)}{4},$$

where $\mathcal{R} = \{(\alpha, \beta): \alpha^2 + \beta^2 \geqq 2u^2, \alpha \geqq 0, \beta \geqq 0\}$. Taking square roots, we have

$$\Phi_c(u) \leqq \frac{\exp(-u^2/2)}{2} \leqq \exp(-u^2/2).$$

The rest of Lemma 1 follows on noting that $\Phi(u) = \Phi_c(-u)$.

Proof of Theorem 1: Let R $(0 < R < C)$ and δ satisfying (17) be given. Since y_1 has density $g(\alpha - b)$, and the $\{y_i\}_1^M$ are independent,

$$\begin{aligned} P(E_1) &= \Pr \bigcup_{j=2}^{M} \{y_j \geqq y_1 - \Delta\} \\ &= \int_{-\infty}^{\infty} d\alpha\, g(\alpha - b) \Pr \left\{ \begin{matrix} \text{at least one} \\ y_j \geqq y_1 - \Delta \end{matrix} \middle| \, y_1 = \alpha \right\} \qquad (25) \\ &= \int_{-\infty}^{\infty} d\alpha\, g(\alpha - b) \Pr \bigcup_{j=2}^{M} \{y_j \geqq \alpha - \Delta\}. \end{aligned}$$

Now since the y_j $(j > 1)$ have density $g(\alpha)$,

$$\Pr \bigcup_{j=2}^{M} \{y_j \geqq \alpha - \Delta\} \leqq \begin{cases} 1 \\ (M - 1) \Pr \{y_j \geqq \alpha - \Delta\} \leqq M\Phi_c(\alpha - \Delta). \end{cases} \qquad (26)$$

Letting a be a parameter to be specified later, we break the integral of equation (25) into two parts, $\alpha \leqq a$ and $\alpha \geqq a$. We then apply the first upper bound of (26) in the first part, and the second bound of (26) in the second part. Thus

$$P(E_1) \leqq \int_{-\infty}^{a} g(\alpha - b)\, d\alpha + M \int_a^\infty g(\alpha - b)\Phi_c(\alpha - \Delta)\, d\alpha.$$

If we assume that

$$a \geqq \Delta, \qquad (27)$$

we can use the bound of Lemma 1 on $\Phi_c(\alpha - \Delta)$ and obtain

$$P(E_1) \leqq \int_{-\infty}^{a} g(\alpha - b)\, d\alpha + M \int_{a}^{\infty} g(\alpha - b) \exp\,[-(\alpha - \Delta)^2/2]\, d\alpha$$
$$= P_1 + MP_2. \tag{28}$$

We now overbound P_1 and P_2. First,

$$P_1 = \int_{-\infty}^{a} g(\alpha - b)\, d\alpha = \int_{-\infty}^{a-b} g(\alpha)\, d\alpha = \Phi(a - b).$$

If we further assume that

$$a \leqq b, \tag{29}$$

we can use Lemma 1 and obtain

$$P_1 \leqq \exp\,[-(b - a)^2/2]. \tag{30}$$

Second,

$$P_2 = \int_a^\infty \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\,[-\tfrac{1}{2}(\alpha - b)^2] \exp\,[-\tfrac{1}{2}(\alpha - \Delta)^2]\, d\alpha$$
$$= \int_a^\infty \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left[-\left(\alpha - \frac{b + \Delta}{2}\right)^2\right] \exp\,[-(b - \Delta)^2/4]\, d\alpha$$
$$= \frac{\exp\,[-(b - \Delta)^2/4]}{\sqrt{2}} \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{\sqrt{2}[a-(b+\Delta)/2]}^{\infty} \exp\,(-v^2/2)\, dv$$
$$= \frac{\exp\,[-(b - \Delta)^2/4]}{\sqrt{2}}\, \Phi_c\left\{\sqrt{2}\left[a - \left(\frac{b + \Delta}{2}\right)\right]\right\}.$$

If we now make a third assumption that

$$a \geqq \frac{b + \Delta}{2}, \tag{31}$$

we can use Lemma 1 again (and $2^{-\frac{1}{2}} \leqq 1$) to bound P_2:

$$P_2 \leqq \exp\,[-(b - \Delta)^2/4] \exp\left\{-\left[a - \left(\frac{b + \Delta}{2}\right)\right]^2\right\}$$
$$= \exp\,[-(b - a)^2/2] \exp\,[-(a - \Delta)^2/2]. \tag{32}$$

Inserting the bounds on P_1 and P_2 into (28), we obtain

$$P(E_1) \leqq \exp\,[-(b - a)^2/2]\{1 + M \exp\,[-(a - \Delta)^2/2]\}, \tag{33a}$$

where from (27), (29), and (31),

$$\left.\begin{matrix} \Delta \\ \dfrac{b+\Delta}{2} \end{matrix}\right\} \leqq a \leqq b. \tag{33b}$$

It remains to choose the parameter a. A good choice will probably result when the upper bound of (28) is differentiated with respect to a and the result set equal to zero:

$$g(a - b) - Mg(a - b) \exp [-(a - \Delta)^2/2] = 0,$$

or

$$M \exp [-(a - \Delta)^2/2] = 1, \tag{34a}$$

or since $M = \exp (RT)$ and $\Delta = \delta(2T)^{\frac{1}{2}}$,

$$a = (R^{\frac{1}{2}} + \delta)(2T)^{\frac{1}{2}}. \tag{34b}$$

Let us now verify that when $0 < R < C$, constraints (33b) are satisfied for this choice of a. Since $R > 0$, $a \geqq \Delta$. Further, since $b = (2CT)^{\frac{1}{2}}$,

$$a - \left(\frac{b + \Delta}{2}\right) = \{\delta - [C^{\frac{1}{2}} - (4R)^{\frac{1}{2}}]\}\left[\frac{(2T)^{\frac{1}{2}}}{2}\right] \geqq 0,$$

since δ satisfies (17). Finally, from (17),

$$b - a = [C^{\frac{1}{2}} - (R^{\frac{1}{2}} + \delta)](2T)^{\frac{1}{2}} \geqq 0.$$

Thus constraints (33b) are, in fact, satisfied. Thus from (34) and (33a)

$$P(E_1) \leqq 2 \exp [-(C^{\frac{1}{2}} - R^{\frac{1}{2}} - \delta)^2 T],$$

which is Theorem 1.

Proof of Theorem 2: Let R $(0 \leqq R < C)$, $\delta > C^{\frac{1}{2}} - (4R)^{\frac{1}{2}}$, and θ_1, θ_2 satisfying equation (21) be given. Then

$$P_e = \Pr \bigcup_{j=2}^{M} \bigcap_{i \neq j} \{y_i < y_j - \Delta\} \leqq \sum_{j=2}^{M} \Pr \bigcap_{i \neq j} \{y_i < y_j - \Delta\},$$

or

$$P_e \leqq M \Pr \bigcap_{i \neq j} \{y_i < y_j - \Delta\}, \qquad j \geqq 2.$$

The last inequality follows from the symmetry of the distributions of the y_j $(j \geqq 2)$. Recalling that the density for y_j $(j \geqq 2)$ is $g(\alpha)$, and that the $\{y_i\}_1^M$ are independent,

$$= M \int_{-\infty}^{\infty} g(\alpha)\, d\alpha \Pr \left\{ \begin{matrix} \text{for all} \quad i \neq j \\ y_i < y_j - \Delta \end{matrix} \middle| \, y_j = \alpha \right\}$$

$$= M \int_{-\infty}^{\infty} g(\alpha)\, d\alpha \Pr \bigcap_{i \neq j} \{ y_i < \alpha - \Delta \}.$$

Again using the independence of the y_i and the fact that the density of y_1 is $g(\alpha - b)$ we have

$$\begin{aligned} \Pr \bigcap_{i \neq j} \{ y_i < \alpha - \Delta \} &= \left[\int_{-\infty}^{\alpha - \Delta} g(\alpha - b)\, d\alpha \right] \left[\int_{-\infty}^{\alpha - \Delta} g(\alpha)\, d\alpha \right]^{M-2} \\ &= \Phi(\alpha - \Delta - b) [\Phi(\alpha - \Delta)]^{M-2}. \end{aligned}$$

Substituting, we obtain

$$P_e \leqq M \int_{-\infty}^{\infty} g(\alpha) \Phi(\alpha - \Delta - b) [\Phi(\alpha - \Delta)]^{M-2}\, d\alpha. \tag{35}$$

Also note that

$$\begin{aligned} [\Phi(\alpha - \Delta)]^{M-2} &= [1 - \Phi_c(\alpha - \Delta)]^{M-2} \\ &\leqq \exp\,[-(M - 2)\Phi_c(\alpha - \Delta)]. \end{aligned} \tag{36}$$

As in the proof of Theorem 1, we break the integral in (35) into two parts $\alpha \leqq a$ and $\alpha \geqq a$, where a will be specified later. In the range $\alpha \leqq a$ we overbound $\Phi(\alpha - \Delta - b)$ by unity, and $[\Phi(\alpha - \Delta)]^{M-2}$ by (36). In the range $\alpha \geqq a$, we overbound $[\Phi(\alpha - \Delta)]^{M-2}$ by unity. Thus

$$\begin{aligned} P_e \leqq M \int_{-\infty}^{a} & g(\alpha) \exp\,[-(M - 2)\Phi_c(\alpha - \Delta)]\, d\alpha \\ & + M \int_{a}^{\infty} g(\alpha) \Phi(\alpha - \Delta - b)\, d\alpha = MP_1 + MP_2\,. \end{aligned} \tag{37}$$

We now overbound P_1 and P_2. First,

$$\begin{aligned} P_1 &= \int_{-\infty}^{a} g(\alpha) \exp\,[-(M - 2)\Phi_c(\alpha - \Delta)]\, d\alpha \\ &\leqq \exp\,[-(M - 2)\Phi_c(a - \Delta)] \int_{-\infty}^{a} g(\alpha)\, d\alpha \\ &\leqq \exp\,[-(M - 2)\Phi_c(a - \Delta)]. \end{aligned} \tag{38}$$

Second, if we assume that

$$a \leqq b + \Delta \tag{39}$$

we can write

$$P_2 = \int_a^\infty g(\alpha)\Phi(\alpha - \Delta - b)\,d\alpha$$

$$= \int_a^{\Delta+b} g(\alpha)\Phi(\alpha - \Delta - b)\,d\alpha + \int_{\Delta+b}^\infty g(\alpha)\Phi(\alpha - \Delta - b)\,d\alpha.$$

In the first integral, $\alpha - \Delta - b \leqq 0$, so that we may use Lemma 1 to bound $\Phi(\alpha - \Delta - b)$. In the second integral, we overbound $\Phi(\alpha - \Delta - b)$ by unity. Thus

$$P_2 \leqq \int_a^{\Delta+b} g(\alpha)\exp\left[-(\alpha - \Delta - b)^2/2\right] d\alpha + \int_{\Delta+b}^\infty g(\alpha)\,d\alpha$$

$$\leqq \int_a^\infty g(\alpha)\exp\left[-(\alpha - \Delta - b)^2/2\right] d\alpha + \Phi_c(\Delta + b).$$

Since from (20) and the fact that $R < C$,

$$\Delta + b = (\delta + C^{\frac{1}{2}})(2T)^{\frac{1}{2}} > 2(C^{\frac{1}{2}} - R^{\frac{1}{2}})(2T)^{\frac{1}{2}} > 0,$$

we can again use Lemma 1 to overbound $\Phi_c(\Delta + b)$. Using the definition of $g(\alpha)$, we have

$$P_2 \leqq \int_a^\infty \frac{1}{(2\pi)^{\frac{1}{2}}}\exp(-\alpha^2/2)\exp\left[-(\alpha - \Delta - b)^2/2\right] d\alpha + \exp\left[-(\Delta + b)^2/2\right]$$

$$= \exp\left[-(b + \Delta)^2/4\right]\frac{1}{(2\pi)^{\frac{1}{2}}}\int_a^\infty \exp\left[-\left(\alpha - \frac{b + \Delta}{2}\right)^2\right] d\alpha + \exp\left[-(\Delta + b)^2/2\right]$$

$$= \exp\left[-(b + \Delta)^2/4\right]\frac{2^{-1/2}}{(2\pi)^{\frac{1}{2}}}\int_{\sqrt{2}[a-(b+\Delta)/2]}^\infty \exp(-v^2/2)\,dv + \exp\left[-(\Delta + b)^2/2\right]$$

$$\leqq \exp\left[-(b + \Delta)^2/4\right]\Phi_c\left[\sqrt{2}\left(a - \frac{b + \Delta}{2}\right)\right] + \exp\left[-(\Delta + b)^2/2\right].$$

If we further assume that

$$a \geqq (b + \Delta)/2, \tag{40}$$

then we can again employ Lemma 1 to bound $\Phi_c[\sqrt{2}(a - (b + \Delta)/2)]$. Hence

$$\begin{aligned} P_2 &\leqq \exp\,[-(b + \Delta)^2/4]\,\exp\left\{-\left[a - \left(\frac{b + \Delta}{2}\right)\right]^2\right\} \\ &\qquad\qquad + \exp\,[-(\Delta + b)^2/2] \qquad (41) \\ &= \exp\,\{-\tfrac{1}{2}[a^2 + (a - \Delta - b)^2]\} + \exp\,[-(\Delta + b)^2/2]. \end{aligned}$$

The difference between the second and first exponents in (41) is

$$-\tfrac{1}{2}\{(\Delta + b)^2 - [a^2 + (a - \Delta - b)^2]\} = a[a - (\Delta + b)] \leqq 0,$$

by (39) and (40). Thus, the first term of (41) is not less than the second, and

$$P_2 \leqq 2\,\exp\,\{-\tfrac{1}{2}[a^2 + (a - \Delta - b)^2]\}. \qquad (42)$$

Inserting the bounds on P_1 (38) and P_2 (42) into (37), we obtain

$$\begin{aligned} P_e \leqq{}& M\,\exp\,[-(M - 2)\Phi_c(a - \Delta)] \\ &+ 2M\,\exp\,\{-\tfrac{1}{2}[a^2 + (a - \Delta - b)^2]\}, \qquad (43a) \end{aligned}$$

where from equations (39) and (40),

$$\frac{b + \Delta}{2} \leqq a \leqq b + \Delta. \qquad (43b)$$

It remains to choose the parameter a, and here we will simply state a good choice of a without giving a motivating argument. Let

$$a = (R^{\frac{1}{2}} + \delta - \theta_2)(2T)^{\frac{1}{2}} \qquad (44)$$

(where θ_2 is the arbitrary parameter which was selected at the beginning of the proof). We must verify that constraints (43b) are satisfied for this choice of a. First, since $R < C$ and $\theta_2 > 0$,

$$b + \Delta - a = (C^{\frac{1}{2}} - R^{\frac{1}{2}} + \theta_2)(2T)^{\frac{1}{2}} > 0.$$

Thus $a \leqq b + \Delta$. Second, from equation (21b),

$$a - \left(\frac{b + \Delta}{2}\right) = \tfrac{1}{2}\{\delta - [C^{\frac{1}{2}} - (4R)^{\frac{1}{2}}] - 2\theta_2\}(2T)^{\frac{1}{2}} \geqq 0,$$

so that $a \geqq (b + \Delta)/2$ and (43b) is satisfied.

Now consider the second term in (43a). Direct substitution of (44) shows that this term is

$$2\,M\,\exp\,\{-[(R^{\frac{1}{2}} + \delta - \theta_2)^2 + (C^{\frac{1}{2}} - R^{\frac{1}{2}} + \theta_2)^2]T\},$$

a single exponential decay in T (as $T \rightarrow \infty$). Finally consider the exponent of the first term of (43). Substituting (44), it is

$$-(M - 2)\Phi_c(a - \Delta) = -(\exp(RT) - 2)\Phi_c\{[R^{\frac{1}{2}} - \theta_2](2T)^{\frac{1}{2}}\}.$$

Making use of the asymptotic formula $\Phi_c(u) \approx (2\pi u)^{-\frac{1}{2}}e^{-u^2/2}$ as $u \rightarrow \infty$ (see p. 106 of Ref. 2), and letting $T \rightarrow \infty$ (and noting that from equation (21b), $R^{\frac{1}{2}} - \theta_2 > 0$), this exponent is asymptotic to

$$\frac{-1}{(2\pi)^{\frac{1}{2}}(R^{\frac{1}{2}} - \theta_2)(2T)^{\frac{1}{2}}} \exp(+KT),$$

where $K > 0$. Thus the first term of equation (43a) decays to zero as a double exponential in T, very much more rapidly than the second term of equation (43a). We can find a T sufficiently large so that the ratio of the first to second terms of equation (43a) $\leqq \theta_1$. With T so chosen

$$P_e \leqq (1 + \theta_1)2 \exp\{-[(R^{\frac{1}{2}} + \delta - \theta_2)^2 + (C^{\frac{1}{2}} - R^{\frac{1}{2}} + \theta_2)^2 - R]T\}$$

which is Theorem 2.

REFERENCES

1. Fano, R. M., *Transmission of Information,* Cambridge, Massachusetts: MIT Press, 1961.
2. Feller, W., *An Introduction to Probability Theory and Its Applications,* vol. I, New York: Wiley, 1950.
3. Forney, G. D., "Exponential Error Bounds for Erasure, List, and Decision Feedback Schemes," IEEE Trans. on Information Theory, *IT-14,* No. 2 (March 1968), pp. 206–220.
4. Kramer, A. J., "Use of Orthogonal Signaling in Sequential Decision Feedback," Information and Control, *10,* No. 5 (May 1967), pp. 509–521.
5. Kramer, A. J., "Analysis of Communication Schemes Using an Intermittent Feedback Link," Stanford University Center for Systems Research, Tech. Rept. No. 7050-11.
6. Schalkwijk, J. P. M. and Kailath, T., "A Coding Scheme for Additive Channels with Feedback, IEEE Trans. on Information Theory, *IT-12,* No. 2 (April 1966), pp. 172–188.
7. Shannon, C. E., Gallager, R. G., and Berlekamp, E. R., "Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels," Information and Control, *10,* No. 1, 5 (January and May 1967), pp. 65–103 and 522–552.
8. Wozencraft, J. and Jacobs, I., *Principles of Communication Engineering,* New York: Wiley, 1965.
9. Wyner, A. D., "On the Probability of Error for Communication in White Gaussian Noise," IEEE Trans. on Information Theory, *IT-13,* No. 1 (January 1967), pp. 86–90.
10. Wyner, A. D., "On the Schalkwijk-Kailath Coding Scheme with a Peak Energy Constraint," IEEE Trans. on Information Theory, *IT-14,* No. 1 (January 1968), pp. 129–134.

Second and Third Order Modulation Terms in the Distortion Produced when Noise Modulated FM Waves are Filtered

By S. O. RICE

(Manuscript received August 8, 1968)

This paper is concerned with the distortion produced in a frequency modulation wave when it passes through a filter. The phase or frequency modulation representing the signal is assumed to be a band of gaussian noise. The main result is an expression for the power spectrum $W_\theta(f)$ of the output phase angle $\theta(t)$. This expression holds for any filter, contains all of the distortion terms due to second and third order modulation, and is suited to computer evaluation. It is useful in many cases, but it has the shortcoming of not containing any modulation terms higher than the third order.

A second result is an approximation to $W_\theta(f)$, based on log $(1 + x) \approx x$ (that is, a "first-order" approximation), which is encountered in the derivation of the main result. Although it does not contain all of the second and third order modulation terms, it does contain higher order modulation terms which may give most of the distortion in some cases. The results given here are compared with those obtained earlier.

I. PREFACE

This work is a sequel to "Distortion and Crosstalk of Linearly Filtered Angle-Modulated Signals" by E. Bedrosian of the Rand Corporation and myself.[1] One of the principal results of that paper is an expression for the distortion produced when a frequency modulation wave, modulated by gaussian noise, passes through a filter assumed to be symmetrical about the carrier frequency.

The assumption of symmetry simplified the analysis, but led to zero second order modulation; and consequently the results do not apply to many cases of practical interest.

The main result of this paper is an expression for the distortion which contains all of the second and third order modulation terms produced by a general filter. It includes the earlier result as a special case.

As before, the input phase angle is assumed to be a gaussian noise.

Some time ago Mr. Bedrosian and I worked out, independently, the second order modulation terms. His analysis is somewhat different from mine and throws a different light on the problem. Since each approach is of interest in its own right and since we were unable to combine the two without losing useful results, we decided to publish our work separately. An early version of Bedrosian's analysis is given in a RAND memorandum.[2]

II. INTRODUCTION

When an angle-modulated wave (FM or PM) passes through a filter, the signal becomes distorted. For a multichannel system this distortion may produce crosstalk. In many practical cases the second and third order modulation terms give a good measure of the distortion. These terms have been studied by a number of investigators. In this paper we obtain some general expressions for them for the case in which the modulation is gaussian.

Our main results includes all of the second and third order modulation products. In this respect, it is more general than some of the earlier expressions for the distortion (see Medhurst,[3] Magnusson,[4] and Liou[5]). However, it does not give higher order modulation terms, some of which appear in earlier "first order" approximations. A first order approximation (similar to the earlier ones) occurs in our derivation of the main result. It is stated, along with the main result, in Section III.

As in Ref. 1, the complex form of the filter input is

$$s(t) = \exp\,[i\omega_0 t + i\varphi(t)] \tag{1}$$

where the carrier frequency is $\omega_0 = 2\pi f_0$ and the signal is carried by the real input phase angle $\varphi(t)$. Let the filter have the transfer function $G(f)$ and the impulse response $g(t)$:

$$\begin{aligned} g(t) &= \int_{-\infty}^{\infty} G(f) \exp\,(i\omega t)\, df, \\ G(f) &= \int_{-\infty}^{\infty} g(t) \exp\,(-i\omega t)\, dt, \qquad \omega = 2\pi f, \end{aligned} \tag{2}$$

where the response $g(t)$ may be complex and may be different from zero when $t < 0$.

The filter is regarded as a bandpass filter for which $G(f)$ is large only near $\pm f_0$. Let normalized functions be defined by

$$\Gamma(f) = \frac{G(f_0 + f)}{G(f_0)}, \qquad \gamma(t) = \frac{g(t) \exp\,(-i\omega_0 t)}{G(f_0)}. \tag{3}$$

From the definitions and the Fourier relations (2) it follows that

$$\begin{aligned} \gamma(t) &= \int_{-\infty}^{\infty} \Gamma(f) \exp(i\omega t)\, df, \qquad \Gamma(f) = \int_{-\infty}^{\infty} \gamma(t) \exp(-i\omega t)\, dt, \\ \int_{-\infty}^{\infty} \gamma(t)\, dt &= 1, \qquad \Gamma^*(-f) = \int_{-\infty}^{\infty} \gamma^*(t) \exp(-i\omega t)\, dt, \end{aligned} \tag{4}$$

where the asterisk denotes "conjugate complex."

The filter output corresponding to the input $s(t)$ is

$$\begin{aligned} s_0(t) &= \int_{-\infty}^{\infty} g(u)s(t-u)\, du \\ &= \int_{-\infty}^{\infty} g(u) \exp[i\omega_0(t-u) + i\varphi(t-u)]\, du \\ &= \left[G(f_0) \int_{-\infty}^{\infty} \gamma(u) \exp[i\varphi(t-u)]\, du\right] \exp(i\omega_0 t) \\ &= \exp(-\alpha_0 - i\beta_0)\{R(t) \exp[i\theta(t)]\} \exp(i\omega_0 t) \\ &= \exp(-\alpha_0 - i\beta_0)\{\exp[i\Theta(t)]\} \exp(i\omega_0 t). \end{aligned} \tag{5}$$

The definition of $\gamma(u)$ is used in going from the second to the third line. In going from the third to the fourth line, the attenuation and phase shift, α_0, β_0 at the carrier frequency have been introduced by writing $G(f_0)$ as $\exp(-\alpha_0 - i\beta_0)$. The complex phase angle $\Theta(t)$ is related to the envelope $R(t) \exp(-\alpha_0)$ and phase angle $-\beta_0 + \theta(t)$ of the output [α_0 and β_0 are constants which do not depend on $\varphi(t)$] by

$$\begin{aligned} \exp[i\Theta(t)] &= R(t) \exp[i\theta(t)], \qquad i\Theta(t) = \ln R(t) + i\theta(t), \\ \theta(t) &= \operatorname{Re} \Theta(t), \qquad \ln R(t) = -\operatorname{Im} \Theta(t). \end{aligned} \tag{6}$$

Comparing the third and fifth lines of equation (5) leads to

$$\Theta(t) = -i \ln\left[\int_{-\infty}^{\infty} \gamma(u) \exp[i\varphi(t-u)]\, du\right]. \tag{7}$$

The analysis may be simplified by introducing the "linear portion" $\Phi(t)$ of $\Theta(t)$. Working with the case in which the input $\varphi(t)$ is small gives

$$\begin{aligned} \Theta(t) &= -i \ln\left[1 + i \int_{-\infty}^{\infty} \gamma(u)\varphi(t-u)\, du + \cdots\right] \\ &\approx (-i)i \int_{-\infty}^{\infty} \gamma(u)\varphi(t-u)\, du, \end{aligned}$$

and this leads us to define the linear portion of $\Theta(t)$ as

$$\Phi(t) = \int_{-\infty}^{\infty} \gamma(u)\varphi(t - u)\, du. \tag{8}$$

The complex phase angle $\Theta(t)$ may be separated into its linear and nonlinear portions by adding and subtracting $i\Phi(t)$ in the exponent in equation (7):

$$\Theta(t) = \Phi(t) - i \ln \left[\int_{-\infty}^{\infty} \gamma(u) \exp \left[i\varphi(t - u) - i\Phi(t) \right] du \right]. \tag{9}$$

Adding and subtracting 1 in the integrand and using the fact that the integral of $\gamma(u)$ is unity gives the fundamental relation

$$\Theta(t) = \Phi(t) - i \ln \left[1 + K(t) \right] \tag{10}$$

where

$$K(t) = \int_{-\infty}^{\infty} du\, \gamma(u) \{ \exp \left[i\varphi(t - u) - i\Phi(t) \right] - 1 \}. \tag{11}$$

In most cases of practical interest, $K(t)$ tends to be small. When $| K(t) | < 1$, expansion of the logarithm gives the series

$$\Theta(t) = \Phi(t) - iK(t) + \frac{i}{2} K^2(t) - \frac{i}{3} K^3(t) + \cdots \tag{12}$$

upon which our analysis is based. When $\varphi(t)$ is gaussian, $| K(t) |$ will occasionally exceed unity. It appears that results obtained from the series of equation (12) represent the first few terms of an asymptotic series. This is further discussed in Appendix F.

If $\varphi(t)$ is small for all values of t, expansion of the exponential in the definition of $K(t)$ [equation (11)] shows that $K(t)$ is $0(\varphi^2)$. Our main expression for the distortion, given in Section III, neglects terms of order φ^8. For this accuracy, equation (12) can be written as

$$\Theta(t) = \Phi(t) - iK(t) + \frac{i}{2} K^2(t) + O(\varphi^6). \tag{13}$$

Since the variable portion $\theta(t)$ of the output phase angle is the real part of $\Theta(t)$, the dc portion, θ_{dc}, of $\theta(t)$ is the average value of Re $\Theta(t)$. When the input $\varphi(t)$ is a stationary gaussian process with zero mean,

$$\begin{aligned} \theta_{dc} &= \operatorname{Re} \langle \Theta(t) \rangle_{\text{av}} \\ &= \operatorname{Re} \left\langle -iK(t) + \frac{i}{2} K^2(t) \right\rangle_{\text{av}} + O(\varphi^6) \\ &= \operatorname{Im} \langle K(t) - 2^{-1} K^2(t) \rangle_{\text{av}} + O(\varphi^6) \end{aligned} \tag{14}$$

where $\langle\ \rangle_{\text{av}}$ denotes "ensemble average" and $\langle\Phi(t)\rangle_{\text{av}}$ is zero because $\Phi(t)$ depends linearly on $\varphi(t)$. Notice that from equation (5), the total output phase angle is $\hat{\theta}(t) = -\beta_0 + \theta(t)$ and that the dc part of $\hat{\theta}(t)$ is

$$\hat{\theta}_{dc} = -\beta_0 + \theta_{dc} .$$

The two-sided power spectra $W_{\hat{\theta}}(f)$, $W_\theta(f)$ of $\hat{\theta}(t)$, $\theta(t)$ contain the dc spikes $(-\beta_0 + \theta_{dc})^2\ \delta(f)$, $\theta_{dc}^2\ \delta(f)$, respectively. Furthermore,

$$W_{\hat{\theta}}(f) = (\beta_0^2 - 2\beta_0\theta_{dc})\ \delta(f) + W_\theta(f).$$

Here $\delta(f)$ is the unit impulse function.

In the following work it is convenient to ignore β_0, and we shall call $\theta(t)$ itself the output phase angle.

III. STATEMENT OF PRINCIPAL RESULTS

In all of the results stated here, the input phase angle $\varphi(t)$ is gaussian with zero mean. The two-sided power spectrum of $\varphi(t)$ is $W_\varphi(f)$. The two-sided power spectrum of the output phase angle $\theta(t)$ is $W_\theta(f)$.

3.1 *Second and Third Order Modulation Terms in $W_\theta(f)$*

The principal result given in this paper† is an expression for $W_\theta(f)$ which contains all of the second and third order modulation terms:

$$\begin{aligned} W_\theta(f) &= \theta_{dc}^2\ \delta(f) + \tfrac{1}{4}W_\varphi(f)\ |\ U(f) + U^*(-f)\ |^2 \\ &\quad + \frac{1}{8}\int_{-\infty}^{\infty} d\rho\ W_\varphi(\rho)W_\varphi(f-\rho)\ |\ T(\rho, f-\rho) - T^*(-\rho, -f+\rho)\ |^2 \\ &\quad + \frac{1}{24}\int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma\ W_\varphi(\rho)W_\varphi(\sigma)W_\varphi(f-\rho-\sigma) \\ &\quad\quad \cdot |\ S(\rho, \sigma, f-\rho-\sigma) + S^*(-\rho, -\sigma, -f+\rho+\sigma)\ |^2 \\ &\quad + O(\varphi^6 W_\varphi). \end{aligned} \qquad (15)$$

† Note added at press time: Equation (15) gives essentially the first few terms of a general expansion due to A. Mircea, Rev. Roum. Sci. Tech.—Electrotechn et Energ., 1967, *t. 12*, No. 3, pp. 359–371, and Proc. IEEE (Correspondence), October 1966, *54*, pp. 1463–1466. I regret the oversight of Mircea's excellent work. Use of his results would have substantially improved this article.

Here the dc part, θ_{dc}, of $\theta(t)$ is the imaginary part of

$$\begin{aligned} D_c = &-\frac{1}{2}\int_{-\infty}^{\infty} d\rho\, W_\varphi(\rho)S(\rho, -\rho) \\ &-\frac{1}{2}\int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma\, W_\varphi(\rho)W_\varphi(\sigma)S(-\sigma, -\rho)[\tfrac{3}{2}S(\sigma, \rho) - \Gamma(\sigma + \rho)] \\ &+ O(\varphi^6) \end{aligned} \tag{16}$$

and

$$\begin{aligned} T(\rho, f - \rho) = {} & S(\rho, f - \rho) \\ & + \int_{-\infty}^{\infty} d\sigma\, W_\varphi(\sigma)[2S(\sigma, \rho)S(-\sigma, f - \rho) - S(\sigma, f - \sigma) \\ & - \Gamma(\sigma)\Gamma(-\sigma)S(\rho, f - \rho) + S(\rho + \sigma, f - \rho - \sigma)] \end{aligned} \tag{17}$$

$$\begin{aligned} U(f) = {} & \Gamma(f) + \int_{-\infty}^{\infty} d\rho\, W_\varphi(\rho)\Gamma(\rho)S(-\rho, f) \\ & + \int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma\, W_\varphi(\rho)W_\varphi(\sigma)\{-\tfrac{1}{2}\Gamma(\rho + \sigma)S(-\rho - \sigma, f) + \Gamma(\sigma) \\ & \cdot[3S(-\sigma, \rho)S(-\rho, f) - S(\rho, f - \rho - \sigma) + S(\rho - \sigma, f - \rho)]\}. \end{aligned} \tag{18}$$

The $\Gamma(f)$ is the normalized filter transfer function defined by equation (3) and the functions S are discussed in Appendix B. They depend only on the filter. That is, they are independent of $W_\varphi(f)$, and are defined symbolically by

$$S(x_1, \cdots, x_n) = \prod_{k=1}^{n} [y^{x_k} - \Gamma(x_k)] \tag{19}$$

where the power y^z of y is to be replaced by $\Gamma(z)$ after multiplying out the product. The S's are symmetric functions of their arguments. For $n = 2$ and 3,

$$\begin{aligned} S(\rho, \sigma) &= y^{\rho+\sigma} - y^\rho\Gamma(\sigma) - y^\sigma\Gamma(\rho) + \Gamma(\rho)\Gamma(\sigma) \\ &= \Gamma(\rho + \sigma) - \Gamma(\rho)\Gamma(\sigma), \\ S(\rho, \sigma, \nu) &= \Gamma(\rho + \sigma + \nu) - \Gamma(\rho + \sigma)\Gamma(\nu) \\ &\quad - \Gamma(\rho + \nu)\Gamma(\sigma) - \Gamma(\sigma + \nu)\Gamma(\rho) + 2\Gamma(\rho)\Gamma(\sigma)\Gamma(\nu). \end{aligned} \tag{20}$$

The $S(\rho, \sigma, \nu)$ of Ref. 1 is the negative of the one used here.

In the "order of" symbol appearing in equation (15) the W_φ enters for dimensional reasons. This is in line with

$$\int_{-\infty}^{\infty} df\, W_\varphi(f) = \langle\varphi^2(t)\rangle_{\text{av}} \tag{21}$$

as can be seen by integrating both sides of equation (15) from $f = -\infty$ to $f = \infty$.

In some instances the expression for $W_\theta(f)$ is useful when rms $\varphi(t)$ is large but rms $d\varphi(t)/dt$ is small.

Bedrosian has computed curves showing the second order distortion for the case of quadratic phase shift, that is, for $\Gamma(f)$ of the form $\exp[iAf^2]$.[2]

As equation (15) stands, it is oriented towards phase modulation. For frequency modulation the time derivatives $\varphi'(t) = d\varphi(t)/dt$, $\theta'(t) = d\theta(t)/dt$ replace the phase angles $\varphi(t)$, $\theta(t)$ as the items of interest. Since the power spectrum $W_{\theta'}(f)$ of $\theta'(t)$ is equal to $(2\pi f)^2 W_\theta(f)$, multiplying (15) by $(2\pi f)^2$ converts it into an expression for $W_{\theta'}(f)$. On the right side of (15), the factor $W_\varphi(f)$ in the first line is replaced by $W_{\varphi'}(f)$ and the W_φ's appearing in the integrands may be transformed into $W_{\varphi'}$'s without introducing infinities. The last statement is seen to be true for the second order modulation integral when we write

$$W_\varphi(\rho)W_\varphi(f-\rho) = W_{\varphi'}(\rho)W_{\varphi'}(f-\rho)\ (2\pi)^{-4}\rho^{-2}(f-\rho)^{-2}$$

and observe that the product $\rho^{-1}(f-\rho)^{-1}T(\rho, f-\rho)$ remains finite even when ρ and $(f-\rho)$ approach zero. The third order term may be treated in a similar way.

In many applications $W_{\varphi'}(f)$ is proportional to D^2 where

$$D^2 = \left\langle\left[\frac{\varphi'(t)}{2\pi}\right]^2\right\rangle_{\text{av}}$$

and D is the rms frequency deviation in cycles per second. Then the second and third order modulation integrals in (15) are proportional to D^4 and D^6, respectively, as D tends to zero. This suggests that the remainder term, $0(\varphi^6 W_\varphi)$, is proportional to D^8. For this reason we shall sometimes refer to (15) as the "small deviation" approximation. When the FM signal to crosstalk ratio in dB is plotted as a function of log D, the behavior of the resulting curve as $D \to 0$ can be computed from (15). Indeed, if the second order modulation predominates, (15) furnishes an asymptote to the curve with a slope of 6 dB per octave. If, because of symmetry in the filter, the second order modulation term in (15) is zero, the third order term gives an asymptote with a slope of 12 dB per octave.

Some idea of how equation (15) begins to fail as D increases from 0 may be obtained by considering the case $\varphi(t) = A \sin \omega_a t$. Then the rms frequency deviation is $D = f_a A/2^{\frac{1}{2}}$ and the filter output is

$$s_0(t) = \sum_{-\infty}^{\infty} J_n(A) G(f_0 + nf_a) \exp [i(\omega_0 + n\omega_a)t]$$

where $J_n(A)$ is a Bessel function and $\omega_a = 2\pi fa$. Consider only the second harmonic. It is proportional to $J_2(A)$, and the approximation underlying (15) is roughly equivalent to replacing $J_2(A)$ by $A^2/8$, the leading term in its power series. The value of A which makes $A^2/8$ exceed $J_2(A)$ by 3 dB is $A \approx 2.0$ and the corresponding D is $1.4\ f_a$. If the baseband of a gaussian FM wave were flat and extended from 0 to B, the expected number of zeros per second would be $1.16\ B$. This is the same as the number of zeros of $A \sin \omega_a t$ with $f_a = 0.58\ B$. This representative value of f_a leads to the estimate that (15) will be in error by 3 dB when $D \approx (1.4)\ (0.58)B \approx 0.8\ B$. Comparison of (15) with experimental values indicates that the 3 dB error point typically occurs when D lies between $B/2$ and B.

3.2 *Power Spectrum of $a\theta(t) + b\ ln\ R(t)$.*

Equation (15) for $W_\theta(f)$ may be modified to give information regarding the fluctuation of the envelope $R(t)$. This information may be of interest, say, in determining the distortion produced by "AM to PM conversion."[5] More generally, suppose that one is interested in the power spectrum $W_x(f)$ of

$$x(t) = a\theta(t) + b \ln R(t) = \text{Re}\,[(a + ib)\Theta(t)] \tag{22}$$

where a and b are arbitrary real constants. Then $W_x(f)$ is given by an expression obtained from equation (15) upon replacing $U(f)$, $T(\rho, f-\rho)$ and $S(\rho, \sigma, f-\rho-\sigma)$ by $(a+ib)U(f)$, $(a+ib)T(\rho, f-\rho)$, and $(a+ib)S(\rho, \sigma, f-\rho-\sigma)$, respectively, so that $U^*\ (-f)$ is replaced by $(a-ib)U^*\ (-f)$, and so on. (See Appendix E.)

3.3 *Second and Third Order Modulation Terms for "Small and Slow" Frequency Deviations*

The expression (15) for $W_\theta(f)$ simplifies when

(*i*) $\Gamma(f)$ can be expanded as a power series

$$\Gamma(f) = \frac{G(f_0 + f)}{G(f_0)} = 1 + \sum_{n=1}^{\infty} \frac{\alpha_n f^n}{n!} \tag{23}$$

and

(*ii*) the effective spread of $W_\varphi(f)$ is so small that the Γ's used in equations (15) to (20) can be replaced by the first few terms of their power series expansion. Roughly, this means that the top baseband frequency is small compared with the filter bandwidth. The instantaneous frequency changes slowly in comparison with the envelope of the impulse response of the filter, and the quasistatic case is approached. With these assumptions the resulting simplified form of $W_\theta(f)$ is given by equation (126). A more complete form of the small and slow deviation approximation is given in equation (133) which brings out the asymptotic nature of the results.

The sum of the second and third order modulation terms given by the integrals in equation (126) [which are the simplified versions of the corresponding integrals in equation (15)] is

$$W_\theta^I(f) = 2^{-1}(\lambda_{2i} + 2^{-1}D^2\lambda_{4i})^2 \int_{-\infty}^{\infty} d\rho \; W_\varphi(\rho) W_\varphi(f - \rho)\rho^2(f - \rho)^2$$

$$+ 6^{-1}(\lambda_{3i})^2 \int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma \; W_\varphi(\rho) W_\varphi(\sigma) W_\varphi(f - \rho - \sigma)\rho^2\sigma^2(f - \rho - \sigma)^2. \tag{24}$$

Here $W_\theta^I(f)$ is the portion of $W_\theta(f)$ which gives the interchannel interference, that is, the noise a listener would hear in an idle channel in a multichannel frequency division multiplex angle-modulation system. D is the rms frequency deviation in cps,

$$D^2 = \left\langle \left[\frac{\varphi'(t)}{2\pi} \right]^2 \right\rangle_{\text{av}} \tag{25}$$

where $\langle\ \rangle_{\text{av}}$ denotes ensemble average. The quantities λ_{ni} are the imaginary parts of the semi-invariants λ_n defined by the expansion

$$\ln \Gamma(f) = \sum_{n=1}^{\infty} \frac{\lambda_n f^n}{n!}. \tag{26}$$

Equations (125) express the first five λ_n's in terms of the first five α_n's, the coefficients in the expansion of $\Gamma(f)$.

The corresponding approximation for the power spectrum $W_x^I(f)$ of $x = a\theta(t) + b \ln R(t)$ [see equation (22)] is obtained by replacing λ_{ni} in equation (24) by $(a\lambda_{ni} + b\lambda_{nr})$ where λ_{nr} denotes the real part of λ_n.

In the important FM case in which the baseband signal $\varphi'(t)$ has a flat power spectrum, the two-sided power spectrum of $\varphi'(t)$ may be taken to be

$$W_{\varphi'}(f) = \begin{cases} 0, & |f| > B \\ (2\pi D)^2/(2B), & |f| \leqq B \end{cases} \tag{27}$$

where the baseband extends from $f = 0$ to $f = B$ (that is, from $-B$ to $+B$ for two-sided spectra), and D is the rms frequency deviation. Then

$$W_{\varphi}(f) = W_{\varphi'}(f)/(2\pi f)^2 = \begin{cases} 0, & |f| > B \\ D^2/(2Bf^2), & |f| \leqq B. \end{cases} \tag{28}$$

When this expression is used, the integrals in equation (24) may be evaluated and it is found that, for $0 \leqq |f| \leqq B$,

$$W_{\theta}^{I}(f) = \frac{D^4}{8B^2}(2B - |f|)(\lambda_{2i} + 2^{-1}D^2\lambda_{4i})^2 + \frac{D^6}{48B^3}(3B^2 - f^2)(\lambda_{3i})^2,$$
$$0 \leqq |f| \leqq B. \tag{29}$$

The average signal power in an elementary frequency band extending from f to $f + \Delta f$ in the input base band is $W_{\varphi'}(f)\,\Delta f$ (radians per second)^2. The ratio of the interference power to the output power in the same elementary band is

$$\frac{W_{\theta'}^{I}(f)\,\Delta f}{W_{\theta'}(f)\,\Delta f} = \frac{W_{\theta}^{I}(f)}{W_{\theta}(f)}. \tag{30}$$

For the flat baseband FM case we may approximate $W_{\theta}(f)$ by $W_{\varphi}(f) = D^2/(2Bf^2)$ and use equation (29). This leads to the approximation

$$\frac{W_{\theta'}^{I}(f)\,\Delta f}{W_{\theta'}(f)\,\Delta f} = \frac{f^2D^2}{4}\left[\left(2 - \frac{|f|}{B}\right)(\lambda_{2i} + 2^{-1}D^2\lambda_{4i})^2 + \frac{D^2}{6}\left(3 - \frac{f^2}{B^2}\right)(\lambda_{3i})^2\right], \tag{31}$$

for the ratio of the interchannel interference power to the signal power in the elementary band $(f, f + \Delta f)$. This ratio has meaning only if $|f| \leqq B$.

Liou has given an approximation which is equivalent to equation (24) for $W_{\theta}^{I}(f)$ with several more terms included.[5] This approximation is discussed in Section XII.

The small and slow deviation approximation described above gives results which agree well with Monte Carlo computations made by C. L. Ruthroff.[6] For illustration we take the simplest of Ruthroff's cases, the one in which the transfer admittance of the filter is

$$G(f) = \frac{1}{1 + 2x + 2x^2 + x^3}, \qquad x = i(f - f_o)/f_c \, . \tag{32}$$

Here f_o is the carrier frequency and f_c is the filter semibandwidth.

Putting $f = f_o$ gives $G(f_o) = 1$. Putting $f = f' + f_o$ gives $x = if'/f_c$ and

$$\begin{aligned} \ln \Gamma(f') &= \ln \frac{G(f' + f_o)}{G(f_o)} \\ &= -\ln (1 + 2x + 2x^2 + x^3) \\ &= -2\frac{x}{1!} + 2\frac{x^3}{3!} - 48\frac{x^5}{5!} + \cdots \\ &= \sum_{n=1}^{\infty} \lambda_n \frac{f'^n}{n!}. \end{aligned} \tag{33}$$

The last line follows from the series of equation (26) defining the λ_n's as the coefficients in the expansion of $\ln \Gamma(f')$. In going from line 2 to line 3, the logarithm is expanded by setting $\alpha_1 = 2$, $\alpha_2 = 4$, $\alpha_3 = 6$, $\alpha_n = 0$ for $n > 3$ in

$$\ln\left(1 + \sum_1^{\infty} \alpha_n x^n/n!\right) = \sum_1^{\infty} \lambda_n x^n/n!$$

and by using the expressions (125) for the λ_n's.

Substituting $x = if'/f_c$ in line 3 of equation (33) and comparing the result with the last line gives

$$\lambda_1 = -2(if_c^{-1}), \qquad \lambda_2 = 0, \qquad \lambda_3 = 2(if_c^{-1})^3, \qquad \lambda_4 = 0.$$

Hence $\lambda_{2i} = 0$, $\lambda_{3i} = -2f_c^{-3}$, $\lambda_{4i} = 0$, and the approximation of equation (31) for the ratio of the interchannel interference power to the signal power leads to

$$-10 \log_{10} \frac{W_\theta^I(f)\,\Delta f}{W_\theta(f)\,\Delta f} \approx -10 \log_{10}\left[\frac{f^2 D^4}{6f_c^6}\left(3 - \frac{f^2}{B^2}\right)\right]. \tag{34}$$

In his Fig. 15 Ruthroff has plotted values of

$$-10 \log_{10} [W_\theta^I(f)\,\Delta f/W_\theta(f)\,\Delta f]$$

for several different values of D/B and f/B with $B = 7$ MHz and

$f_c = 119$ MHz.[6] The agreement with our equation (34) is good at $f/B = 1.0$. Our D/B is the same as Ruthroff's σ/W and our $f/B = 1.0$ corresponds to Ruthroff's slot 10. At $f/B = 0.4$, equation (34) gives values which are about 3 or 4 decibels less than the Monte Carlo values, but this may still be regarded as good agreement.

Similar agreement is found when the small and slow deviation approximation is applied to a number of the other cases examined by Ruthroff.

3.4 *A "First Order" Approximation*

Although equation (15) is useful in some FM distortion problems, in some cases it is of no help. One example concerns the distortion produced by an ideal filter centered on the carrier frequency and having a semibandwidth exceeding $3B$, where B is the baseband. That is, $W_\varphi(f)$ is 0 for $|f| > B$. In this case the distortion is produced by modulation terms of order higher than the third, and these are neglected in equation (15).

For such problems "first order" approximations can sometimes be used. The term "first order" refers to the approximation $\ln(1 + x) \approx x$ where x is of the nature of $K(t)$ in equation (10); it does not refer to the order of the modulation products in x. Different choices of x lead to different first order approximations. The first order approximation given by the first two terms in the series of equation (12) for $\Theta(t)$ is

$$\theta(t) \approx \mathrm{Re}\,\Phi(t) + \mathrm{Im}\,K(t). \tag{35}$$

The output phase angle $\theta(t)$ may be written as the sum

$$\theta(t) = \theta_{dc} + \theta_l(t) + \theta_{nl}(t)$$

where $\theta_l(t) = \mathrm{Re}\,\Phi(t)$ is the "linear portion" of $\theta(t)$, and $\theta_{nl}(t)$ given by

$$\begin{aligned} \theta_{nl}(t) &= \theta(t) - \theta_l(t) - \theta_{dc} \\ &= \theta(t) - \mathrm{Re}\,\Phi(t) - \theta_{dc} \end{aligned} \tag{36}$$

is the time varying part of the "nonlinear distortion" in $\theta(t)$. The first order approximation for $\theta_{nl}(t)$ corresponding to the first order approximation of equation (35) for $\theta(t)$ is

$$\theta_{nl}(t) \approx \mathrm{Im}\,K(t) - \theta_{dc} \approx \mathrm{Im}\,K(t) - \mathrm{Im}\,\langle K(t)\rangle_{\mathrm{av}} = y(t) \tag{37}$$

where

$$y(t) = \mathrm{Im}\,[K(t) - \langle K(t)\rangle_{\mathrm{av}}]. \tag{38}$$

The work of Section VI, which is part of the derivation of equation

(15), shows that the power spectrum of $y(t)$ is

$$W_y(f) = 2^{-1}\,\mathrm{Re}\,[P(f) + Q(f) + P^*(-f) + Q^*(-f)] \tag{39}$$

where

$$\begin{aligned}
P(f) &= -\frac{1}{2}\int_{-\infty}^{\infty} d\tau \exp(-i\omega\tau)\int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\, \gamma(u)\gamma(v) \\
&\qquad \cdot \exp[a(u) + a(v)]\{\exp[2c(u, v, \tau)] - 1\} \\
Q(f) &= \frac{1}{2}\int_{-\infty}^{\infty} d\tau \exp(-i\omega\tau)\int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\, \gamma^*(u)\gamma(v) \\
&\qquad \cdot \exp[a^*(u) + a(v)]\{\exp[2\hat{c}(u, v, \tau)] - 1\} \\
a(u) &= -\frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f)H_u(f)H_u(-f) \\
c(u, v, \tau) &= -\frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f)H_u(-f)H_v(f)\exp(i\omega\tau) \\
\hat{c}(u, v, \tau) &= \frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f)H_u^*(f)H_v(f)\exp(i\omega\tau) \\
H_u(f) &= \exp(-i\omega u) - \Gamma(f), \qquad \omega = 2\pi f.
\end{aligned} \tag{40}$$

The function $Q(f)$ is real when f is real, and $P(f)$ is an even function of f.

The first order approximation $W_y(f)$ for the power spectrum of $\theta_{nt}(t)$ contains some higher order modulation terms which are not contained in our main equation (15) for $W_\theta(f)$; conversely, equation (15) contains terms which are not in $W_y(f)$. In using the first order approximation of equation (35), which may be rewritten as

$$\theta(t) = \mathrm{Re}\,\Phi(t) + \mathrm{Im}\,K(t) + O(K^2),$$

one should guard against throwing away* [in the $O(K^2)$ terms] quantities which are of the same order as those being computed from $\mathrm{Re}\,\Phi(t) + \mathrm{Im}\,K(t)$ (the leading term). Although each case requires its own investigation, it is helpful to remember that $K(t)$ is $O(\varphi^2)$. Furthermore, when $\gamma(t)$ is real, $\mathrm{Im}\,K(t)$ is $O(\varphi^3)$. Also when $\Gamma(f) \equiv 1$, $K(t)$ becomes 0; and when

$$|\,\Gamma(f) - 1\,| < \epsilon \ll 1$$

for all real values of f (as in the case of small wave-guide echoes), it

* This type of error has been discussed by Enloe, Ruthroff, Gladwin, and Medhurst in Refs. 7 and 8.

may be conjectured that $K(t)$ itself is $O(\epsilon)$ irrespective of how large $\langle \varphi^2 \rangle_{av}$ may be.

The approximation $\theta_{nl}(t) \approx y(t)$, is not quite the same as earlier first order approximations.[3,4,9–12] It is a closer, but more complicated, approximation because $\varphi(t - u) - \Phi(t)$ is used in the integral equation (11) for $K(t)$ instead of $\varphi(t - u) - \varphi(t)$ as in the earlier approximations. Appendix D gives some results obtained when the present analysis is repeated with $\varphi(t - u) - \varphi(t)$ in place of $\varphi(t - u) - \Phi(t)$.

Equation (39) for $W_y(f)$ gives a first order approximation for the power spectrum of $\theta_{nl}(t) = \theta(t) - \operatorname{Re} \Phi(t) - \theta_{dc}$. The corresponding approximation for the power spectrum of

$$\ln R(t) + \operatorname{Im} \Phi(t) - [\ln R(t)]_{dc} \approx \operatorname{Re} [K(t) - \langle K(t) \rangle_{av}] = x(t)$$

is

$$W_x(f) = 2^{-1} \operatorname{Re} [-P(f) + Q(f) - P^*(-f) + Q^*(-f)].$$

3.5 *Simplification When Filter has Symmetry* $\Gamma(-f) = \Gamma^*(f)$

When the filter has the symmetry

$$\Gamma(-f) = \Gamma^*(f) \tag{41}$$

about the carrier frequency, the even order modulation terms disappear, $S^*(-x_1, \cdots, -x_n)$ becomes equal to $S(x_1, \cdots, x_n)$, and equation (15) becomes

$$\begin{aligned} W_\theta(f) = {} & W_\varphi(f) \mid U(f) \mid^2 \\ & + \frac{1}{6} \int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma \, W_\varphi(\rho) W_\varphi(\sigma) W_\varphi(f - \rho - \sigma) \\ & \cdot \mid S(\rho, \sigma, f - \rho - \sigma) \mid^2 + O(\varphi^6 W_\varphi). \end{aligned} \tag{42}$$

Here $U(f)$ is still given by equation (18) and $S(\rho, \sigma, \nu)$ by equation (20). This expression for $W_\theta(f)$ agrees with one of the main results of Ref. 1 when the double integral in equation (18) for $U(f)$ is assumed to be so small that it may be neglected.

When $\Gamma(-f)$ is equal to $\Gamma^*(f)$, the coefficients α_n in the power series of equation (23) for $\Gamma(f)$ are real when n is even, and imaginary when n is odd. The same is true for the λ_n's of equation (26). Hence λ_{2i}, λ_{4i} are zero and the second order modulation terms disappear from the small and slow deviation approximations equations (24), (29), and (31) for $W_\theta^I(f)$.

The relation $\Gamma(-f) = \Gamma^*(f)$ implies that $\gamma(u)$ is real and that $H_u(-f)$ is equal to $H_u^*(f)$. Then $a(u)$, $c(u, v, \tau)$, and $\hat{c}(u, v, \tau)$ are real and $\hat{c}(u, v, \tau)$

is equal to $-c(u, v, \tau)$. Both $P(f)$ and $Q(f)$ become even and real. Equation (39) for the power spectrum of $y(t)$, that is, the first order approximation for the power spectrum of the nonlinear distortion $\theta_{nl}(t)$, becomes

$$W_y(f) = P(f) + Q(f). \tag{43}$$

Here the triple integral for $P(f)$ is the same as that given in equation (40); and the triple integral for $Q(f)$ may be obtained from the integral for $P(f)$ by changing the sign of $2c(u, v, \tau)$. Hence equation (43) becomes

$$W_y(f) = -\int_{-\infty}^{\infty} d\tau \exp(-i\omega\tau) \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\, \gamma(u)\gamma(v) \cdot \exp[a(u) + a(v)] \sinh[2c(u, v, \tau)] \tag{44}$$

where $\omega = 2\pi f$, and $a(u)$, $c(u, v, \tau)$ are given by equation (40).

IV. INITIAL EXPRESSION FOR THE POWER SPECTRUM OF $\theta(t)$

When $\theta(t)$ is a stationary noise process its two-sided power spectrum $W_\theta(f)$ is the Fourier transform of its autocovariance:

$$W_\theta(f) = \int_{-\infty}^{\infty} \exp(-i\omega\tau)\langle\theta(t)\theta(t+\tau)\rangle_{\text{av}}\, d\tau, \qquad \omega = 2\pi f. \tag{45}$$

Denoting functions with arguments t, $t + \tau$ by subscripts 1, 2 and using

$$\theta(t) = \text{Re}\,\Theta(t) = 2^{-1}[\Theta(t) + \Theta^*(t)] \tag{46}$$

gives

$$\begin{aligned}\langle\theta(t)\theta(t+\tau)\rangle_{\text{av}} &= \langle\theta_1\theta_2\rangle_{\text{av}} \\ &= 2^{-1}\,\text{Re}\,[\langle\Theta_1\Theta_2\rangle_{\text{av}} + \langle\Theta_1^*\Theta_2\rangle_{\text{av}}].\end{aligned} \tag{47}$$

The procedure of Appendix A and equation (13) for the complex phase angle $\Theta(t)$ lead to

$$\begin{aligned}\langle\Theta_1\Theta_2\rangle_{\text{av}} &= A(\tau) + A(-\tau) + O(\varphi^8) \\ \langle\Theta_1^*\Theta_2\rangle_{\text{av}} &= B(\tau) + B^*(-\tau) + O(\varphi^8)\end{aligned} \tag{48}$$

where $A(\tau)$ and $B(\tau)$ are the ensemble averages

$$\begin{aligned}A(\tau) &= \langle\Phi_1(2^{-1}\Phi_2 - iK_2 + i2^{-1}K_2^2) - 2^{-1}K_1(K_2 - K_2^2)\rangle_{\text{av}} \\ B(\tau) &= \langle\Phi_1^*(2^{-1}\Phi_2 - iK_2 + i2^{-1}K_2^2) + 2^{-1}K_1^*(K_2 - K_2^2)\rangle_{\text{av}}\,.\end{aligned} \tag{49}$$

The remainder terms in equation (48) are $O(\varphi^8)$ instead of $O(\varphi^7)$ because the ensemble average of an odd order term is zero.

It also follows from Appendix A that equation (45) for the power spectrum of $\theta(t)$ goes into

$$W_\theta(f) = 2^{-1}\,\mathrm{Re}\,[P(f) + Q(f) + P^*(-f) + Q^*(-f)] + O(\varphi^6 W_\varphi) \quad (50)$$

$$\begin{aligned} P(f) &= \int_{-\infty}^{\infty} \exp(-i\omega\tau) A(\tau)\, d\tau \\ Q(f) &= \int_{-\infty}^{\infty} \exp(-i\omega\tau) B(\tau)\, d\tau. \end{aligned} \quad (51)$$

Although the functions $P(f)$, $Q(f)$ used here are not the same as those in Section III, they are of the same nature.

V. CALCULATION OF AVERAGES NEEDED FOR COVARIANCES

The equations of Section IV show that the value of $\langle \theta(t)\theta(t+\tau)\rangle_{\mathrm{av}}$ depends upon various ensemble averages of products of $\Phi(t)$ and $K(t)$. When the input phase angle $\varphi(t)$ is gaussian, these averages may be computed by using a result proved in Ref. 1.

Let L be a linear operator which operates on functions of t, and let

$$L \exp(i\omega t) = \exp(i\omega t)\, \ell(f), \qquad \omega = 2\pi f. \quad (52)$$

Let $\varphi(t)$ be a stationary gaussian process with two-sided power spectrum $W_\varphi(f)$. Then

$$\langle \exp[iL\varphi(t)]\rangle_{\mathrm{av}} = \exp\left[-\frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f)\ell(f)\ell(-f)\right]\cdot \quad (53)$$

Setting $xL\varphi(t)$ for $L\varphi(t)$ and comparing coefficients of x^2 in the power series expansions of the two sides of equation (53) shows that

$$\int_{-\infty}^{\infty} df\, W_\varphi(f)\ell(f)\ell(-f) = \langle [L\varphi(t)]^2\rangle_{\mathrm{av}} .$$

That $\langle \exp[iL\varphi(t)]\rangle_{\mathrm{av}}$ is equal to $\exp\{-2^{-1}\langle [L\varphi(t)]^2\rangle_{\mathrm{av}}\}$ follows from the fact that the real and imaginary parts of $L\varphi(t)$ are correlated gaussian processes.

In dealing with $K(t)$ it is convenient to introduce the function $J(v, \tau)$ defined by

$$J(v, \tau) = \exp[i\varphi(t + \tau - v) - i\Phi(t + \tau)]. \quad (54)$$

The dependence of $J(v, \tau)$ on t is ignored because the right side of equation (54) is a stationary random process, and $J(v, \tau)$ will be used only to calculate ensemble averages. The examples, which follow from the definition of equation (11) of $K(t)$,

$$\begin{aligned}\langle K_1\rangle_{\text{av}} &= \langle K(t)\rangle_{\text{av}} = \int_{-\infty}^{\infty} du\, \gamma(u)\langle J(u, o) - 1\rangle_{\text{av}} \\ \langle K_1K_2\rangle_{\text{av}} &= \langle K(t)K(t + \tau)\rangle_{\text{av}} \\ &= \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\, \gamma(u)\gamma(v)\langle[J(u, o) - 1][J(v, \tau) - 1]\rangle_{\text{av}} \\ \langle K_1^*K_2\rangle_{\text{av}} &= \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\, \gamma^*(u)\gamma(v)\langle[J^*(u, o) - 1][J(v, \tau) - 1]\rangle_{\text{av}}\end{aligned} \tag{55}$$

show that averages of the type $\langle J(u, o)\rangle_{\text{av}}$, $\langle J(u, o)J(v, \tau)\rangle_{\text{av}}$, and $\langle J^*(u, o)J(v, \tau)\rangle_{\text{av}}$ are needed.

To calculate $\langle J(u, o)\rangle_{\text{av}}$ from the general result, equation (53), let L be the operator which carries $\varphi(t)$ into $\varphi(t - u) - \Phi(t)$. Replacing $\Phi(t)$ by the integral which defines it gives

$$L\varphi(t) = \varphi(t - u) - \int_{-\infty}^{\infty} ds\, \gamma(s)\varphi(t - s).$$

The function $\ell(f)$ associated with L is obtained by setting $\exp(i\omega t)$ in place of $\varphi(t)$:

$$\begin{aligned}\exp(i\omega t)\ell(f) &= L[\exp(i\omega t)] \\ &= \exp[i\omega(t - u)] - \int_{-\infty}^{\infty} ds\, \gamma(s) \exp[i\omega(t - s)], \\ \ell(f) &= \exp(-i\omega u) - \Gamma(f) \equiv H_u(f).\end{aligned}$$

Then equation (53) gives

$$\begin{aligned}\langle\exp[iL\varphi(t)]\rangle_{\text{av}} &= \langle\exp[i\varphi(t - u) - i\Phi(t)]\rangle_{\text{av}} \\ &= \langle J(u, o)\rangle_{\text{av}} = \exp[a(u)]\end{aligned} \tag{56}$$

where

$$a(u) = -\frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f)H_u(f)H_u(-f). \tag{57}$$

The functions $a(u)$ and

$$H_u(f) = \exp(-i2\pi fu) - \Gamma(f) \tag{58}$$

play important parts in the analysis. The present $H_u(f)$ is the negative of the one used in Reference 1, a change made to simplify the analysis.

The calculation of $\langle J(u, o)J(v, \tau)\rangle_{\text{av}}$ proceeds in much the same way. Let

$$\begin{aligned} L\varphi(t) &= \varphi(t - u) - \Phi(t) + \varphi(t + \tau - v) - \Phi(t + \tau), \\ \ell(f) &= H_u(f) + \exp(i\omega\tau)H_v(f), \\ -\frac{1}{2}\int_{-\infty}^{\infty} & df\, W_\varphi(f)\ell(f)\ell(-f) \\ &= -\frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f)[H_u(f)H_u(-f) + H_v(f)H_v(-f) \\ &\qquad + 2\exp(i\omega\tau)H_u(-f)H_v(f)] \\ &= a(u) + a(v) + 2c(u, v, \tau) \end{aligned} \tag{59}$$

where $W_\varphi(f)$ is an even function of f and $c(u, v, \tau)$ is the integral

$$c(u, v, \tau) = -\frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f)H_u(-f)H_v(f)\exp(i2\pi f\tau). \tag{60}$$

Consequently,

$$\langle J(u, o)J(v, \tau)\rangle_{\text{av}} = \exp[a(u) + a(v) + 2c(u, v, \tau)]. \tag{61}$$

Similarly, to calculate $\langle J^*(u, o)J(v, \tau)\rangle_{\text{av}}$ let

$$\begin{aligned} L\varphi(t) &= -\varphi(t - u) + \Phi^*(t) + \varphi(t + \tau - v) - \Phi(t + \tau), \\ \ell(f) &= -\exp(-i\omega u) + \int_{-\infty}^{\infty} ds\, \gamma^*(s)\exp(-i\omega s) + \exp(i\omega\tau)H_v(f) \\ &= -H_u^*(-f) + \exp(i\omega\tau)H_v(f) \end{aligned} \tag{62}$$

where the Fourier transform of $\gamma^*(s)$ is $\Gamma^*(-f)$. The work of equations (59) and (60) goes through much as before with $-H_u^*(-f)$ in place of $H_u(f)$. The result is

$$\langle J^*(u, o)J(v, \tau)\rangle_{\text{av}} = \exp[a^*(u) + a(v) + 2\hat{c}(u, v, \tau)] \tag{63}$$

where

$$\hat{c}(u, v, \tau) = \frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f)H_u^*(f)H_v(f)\exp(i2\pi f\tau). \tag{64}$$

All of the averages needed are given in Tables I and II. Items 1, 3, and 6 in Table I have just been computed, and the others may be ob-

TABLE I—ENSEMBLE AVERAGES OF PRODUCTS OF $J(v, \tau)$'S

No.	Average	$l(f)$	Value
1	$\langle J(u, o)\rangle_{\text{av}}$	$H_u(f)$	$\exp[a(u)]$
2	$\langle J(v, \tau)\rangle_{\text{av}}$	$\exp(i\omega\tau)H_v(f)$	$\exp[a(v)]$
3	$\langle J(u, o)J(v, \tau)\rangle_{\text{av}}$	$H_u(f) + \exp(i\omega\tau)H_v(f)$	$\exp[a(u) + a(v) + 2c(u, v, \tau)]$
4	$\langle J(u, o)J(v, \tau)J(w, \tau)\rangle_{\text{av}}$	$H_u(f) + \exp(i\omega\tau)[H_v(f) + H_w(f)]$	$\exp[a(u) + a(v) + a(w) + 2c(u, v, \tau) + 2c(u, w, \tau) + 2c(w, v, o)]$
5	$\langle J^*(u, o)\rangle_{\text{av}}$	$-H_u^*(-f)$	$\exp[a^*(u)]$
6	$\langle J^*(u, o)J(v, \tau)\rangle_{\text{av}}$	$-H_u^*(-f) + \exp(i\omega\tau)H_v(f)$	$\exp[a^*(u) + a(v) + 2\hat{c}(u, v, \tau)]$
7	$\langle J^*(u, o)J(v, \tau)J(w, \tau)\rangle_{\text{av}}$	$-H_u^*(-f) + \exp(i\omega\tau)[H_v(f) + H_w(f)]$	$\exp[a^*(u) + a(v) + a(w) + 2\hat{c}(u, v, \tau) + 2\hat{c}(u, w, \tau) + 2c(w, v, o)]$

TABLE II—ENSEMBLE AVERAGES OF PRODUCTS CONTAINING $\varphi(t - u)$

No.	Average	$l(f)$	Value
1	$\langle \varphi(t - u)\varphi(t + \tau - v)\rangle_{\text{av}}$		$\int_{-\infty}^{\infty} df W_\varphi(f) \exp[i\omega(\tau - v + u)]$
2	$\langle \varphi(t - u)J(v, \tau)\rangle_{\text{av}}$	$x \exp(-i\omega u) + \exp(i\omega\tau)H_v(f)$	$i\int_{-\infty}^{\infty} df W_\varphi(f) \exp[i\omega(\tau + u)]H_v(f) \exp[a(v)]$
3	$\langle \varphi(t - u)J(v, \tau)J(w, \tau)\rangle_{\text{av}}$	$x \exp(-i\omega u) + \exp(i\omega\tau)[H_v(f) + H_w(f)]$	$i\int_{-\infty}^{\infty} df W_\varphi(f) \exp[i\omega(\tau + u)][H_v(f) + H_w(f)] \exp[a(v) + a(w) + 2c(w, v, o)]$

tained in a similar manner. The entries in the last column of Table I may be verified by expressing the a's and c's as ensemble averages [see equation (67)] and using $\langle \exp [iL\varphi(t)]\rangle_{\text{av}} = \exp \{-2^{-1}\langle [L\varphi(t)]^2\rangle_{\text{av}}\}$.

Table II gives averages of products in which one factor is $\varphi(t - u)$. The first average, $\langle \varphi(t - u)\varphi(t + \tau - v)\rangle_{\text{av}}$, is the Fourier transform of $W_\varphi(f)$. The second average, $\langle \varphi(t - u)J(v, \tau)\rangle_{\text{av}}$, is the coefficient of ix in the expansion of exp $\langle [iL\varphi(t)]\rangle$ where

$$\begin{aligned} L\varphi(t) &= x\varphi(t - u) + \varphi(t + \tau - v) - \Phi(t + \tau) \\ \ell(f) &= x \exp(-i\omega u) + \exp(i\omega\tau)H_v(f). \end{aligned} \tag{65}$$

The third average may be computed in a similar way.

The following list brings together the integrals $a(v)$, $c(u, v, \tau)$, $\cdots$ which appear in Tables I and II:

$$\begin{aligned} a(u) &= -\frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f)H_u(f)H_u(-f) \\ c(u, v, \tau) &= -\frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f)H_u(-f)H_v(f) \exp(i2\pi f\tau) \\ a^*(u) &= -\frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f)H_u^*(f)H_u^*(-f) \\ \hat{c}(u, v, \tau) &= \frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f)H_u^*(f)H_v(f) \exp(i2\pi f\tau) \end{aligned} \tag{66}$$

where $H_u(f) = \exp(-i2\pi fu) - \Gamma(f)$, and replacing $H_u(f)$ by $-H_u^*(-f)$ in $a(u)$, $c(u, v, \tau)$ gives $a^*(u)$, $\hat{c}(u, v, \tau)$. Also

$$\begin{aligned} a(u) &= c(u, u, o), \qquad c(w, v, o) = c(v, w, o), \\ c(u, v, -\tau) &= c(v, u, \tau), \qquad \hat{c}(u, v, -\tau) = \hat{c}^*(v, u, \tau), \\ c(u, v, \tau) &= -\tfrac{1}{2}\langle [\varphi(t - u) - \Phi(t)][\varphi(t + \tau - v) - \Phi(t + \tau)]\rangle_{\text{av}} \\ \hat{c}(u, v, \tau) &= \tfrac{1}{2}\langle [\varphi(t - u) - \Phi^*(t)][\varphi(t + \tau - v) - \Phi(t + \tau)]\rangle_{\text{av}}\,. \end{aligned} \tag{67}$$

When $\Gamma(-f)$ is equal to $\Gamma^*(f)$, both $\gamma(t)$ and $\Phi(t)$ are real; and it follows that $a(u)$, $c(u, v, \tau)$, $\hat{c}(u, v, \tau)$ are also real. Furthermore, $H_u(-f) = H_u^*(f)$ and $\hat{c}(u, v, \tau) = -c(u, v, \tau)$.

VI. THE POWER SPECTRUM OF $K(t)$

The dc portion of the complex random process $K(t)$ defined by the integral in equation (11) is the complex constant

$$\begin{aligned}\langle K(t)\rangle_{\text{av}} = \langle K_1\rangle_{\text{av}} &= \int_{-\infty}^{\infty} du\, \gamma(u)\langle J(u, o) - 1\rangle_{\text{av}} \\ &= \int_{-\infty}^{\infty} du\, \gamma(u)\{\exp[a(u)] - 1\}.\end{aligned} \tag{68}$$

This follows from equation (55) for $\langle K_1\rangle_{\text{av}}$ and the expression for $\langle J(u, o)\rangle_{\text{av}}$ given in Table I.

The power spectrum of $K(t)$ is the Fourier transform of $\langle K_1^* K_2\rangle_{\text{av}}$. The integral for $\langle K_1^* K_2\rangle_{\text{av}}$ given by equation (55) and the ensemble averages of $J^*(u, o)$, $J(v, \tau)$, and $J^*(u, o)J(v, \tau)$ given in Table I lead to

$$\begin{aligned}\langle K_1^* K_2\rangle_{\text{av}} = \langle K_1\rangle\langle K_1^*\rangle_{\text{av}} &+ \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\, \gamma^*(u)\gamma(v) \exp[a^*(u) + a(v)] \\ &\cdot \{\exp[2\hat{c}(u, v, \tau)] - 1\}.\end{aligned} \tag{69}$$

Integrals of the type appearing in equations (68) and (69) may be expressed as infinite series involving the S functions [which depend only on $\Gamma(f)$] described in Appendix B and the more complicated functions S_n described in Appendix C. Only the first few terms need be considered when most of the distortion arises from second and third order modulation.

The definition [equation (167)] of the complex constant S_0 and its series expansion [equation (171)] give

$$\begin{aligned}\langle K(t)\rangle_{\text{av}} &= S_0 - 1 \\ &= -\frac{1}{2}\int_{-\infty}^{\infty} d\rho\, W_{\varphi}(\rho)S(\rho, -\rho) \\ &\quad + \frac{1}{8}\int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma\, W_{\varphi}(\rho)W_{\varphi}(\sigma)S(\rho, \sigma, -\rho, -\sigma) + O(\varphi^6).\end{aligned} \tag{70}$$

Expanding $\exp[2\hat{c}(u, v, \tau)]$ in equation (69) in powers of $2\hat{c}(u, v, \tau)$, replacing each $\hat{c}(u, v, \tau)$ by its defining integral [equation (66)] with ρ in place of f, and integrating with respect to u and v with the help of

$$S_n(\rho_1, \cdots, \rho_n) = \int_{-\infty}^{\infty} dv\, \gamma(v) \exp[a(v)] \prod_{k=1}^{n} H_v(\rho_k) \tag{71}$$

leads to

$$\begin{aligned}\langle K_1^* K_2\rangle_{\text{av}} = |\langle K_1\rangle_{\text{av}}|^2 &+ \sum_{n=1}^{\infty} \frac{1}{n!}\int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_n \left[\prod_{k=1}^{n} W_{\varphi}(\rho_k)\right] \\ &\cdot \exp[i2\pi\tau(\rho_1 + \cdots + \rho_n)]S_n^*(\rho_1, \cdots, \rho_n)S_n(\rho_1, \cdots, \rho_n).\end{aligned} \tag{72}$$

When this expression for $\langle K_1^* K_2 \rangle_{\text{av}}$ is put in the integral

$$W_K(f) = \int_{-\infty}^{\infty} \exp(-i\omega\tau) \langle K_1^* K_2 \rangle_{\text{av}} \, d\tau, \qquad \omega = 2\pi f \tag{73}$$

for the power spectrum, $W_K(f)$, of $K(t)$ and the Fourier transform of unity denoted by $\delta(f)$, the result is

$$\begin{aligned} W_K(f) &= |\langle K(t) \rangle_{\text{av}}|^2 \, \delta(f) + \int_{-\infty}^{\infty} d\tau \exp(-i\omega\tau) \\ &\quad \cdot \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv \, \gamma^*(u)\gamma(v) \exp[a^*(u)+a(v)]\{\exp[2\hat{c}(u, v, \tau)]-1\} \\ &= |S_0 - 1|^2 \, \delta(f) + \sum_{n=1}^{\infty} \frac{1}{n!} \int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_n \\ &\quad \cdot \delta(f - \rho_1 - \cdots - \rho_n) \left[\prod_{k=1}^{n} W_\varphi(\rho_k) \right] |S_n(\rho_1, \cdots, \rho_n)|^2 \\ &= |S_0 - 1|^2 \, \delta(f) + W_\varphi(f) \, |S_1(f)|^2 \\ &\quad + \frac{1}{2} \int_{-\infty}^{\infty} d\rho \, W_\varphi(\rho) W_\varphi(f - \rho) \, |S_2(\rho, f - \rho)|^2 \\ &\quad + \frac{1}{6} \int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma \, W_\varphi(\rho) W_\varphi(\sigma) W_\varphi(f - \rho - \sigma) \\ &\quad \cdot |S_3(\rho, \sigma, f - \rho - \sigma)|^2 + \cdots . \end{aligned} \tag{74}$$

The leading terms in the series for S_0, S_1, S_2, S_3 in terms of unsubscripted S's are given by equations at the end of Appendix C. The inequality for S_n given in Appendix C may be used to show that the last series in equation (74) converges when $W_\varphi(f)$ remains finite for all values of f and $\langle [\varphi(t)]^2 \rangle_{\text{av}}$ is finite. The convergence of similar series which will be encountered later will be tacitly assumed.

VII. "FIRST ORDER" APPROXIMATION FOR POWER SPECTRUM OF $\theta(t)$

Before taking up the problem of computing $W_\theta(f)$ from $\theta(t) = \operatorname{Re} \Theta(t)$ and

$$\Theta(t) = \Phi(t) - iK(t) + \frac{i}{2} K^2(t) + O(\varphi^6),$$

which is the same as equation (13), we shall go through a similar, but simpler, calculation using the "first order" approximation

$$\Theta(t) = \Phi(t) - iK(t) + O(\varphi^4). \tag{75}$$

Neglecting $O(K^2)$ terms in equation (14) for θ_{dc}, and in the covariance expressions given in Section IV, shows that

$$\begin{aligned}\theta_{dc} &= \operatorname{Im}\langle K(t)\rangle_{\text{av}} + O(\varphi^4)\\ W_\theta(f) &= 2^{-1}\operatorname{Re}[P(f) + Q(f) + P^*(-f) + Q^*(-f)] + O(\varphi^4 W_\varphi)\end{aligned}\tag{76}$$

where $P(f)$, $Q(f)$ are the respective Fourier transforms of

$$\begin{aligned}A(\tau) &= \langle \Phi_1(2^{-1}\Phi_2 - iK_2) - 2^{-1}K_1K_2\rangle_{\text{av}}\\ B(\tau) &= \langle \Phi_1^*(2^{-1}\Phi_2 - iK_2) - 2^{-1}K_1^*K_2\rangle_{\text{av}}\end{aligned}\tag{77}$$

Expressions for $\langle K_1^*K_2\rangle_{\text{av}}$ have been obtained in the preceding section. Repeating the work with K_1 in place of K_1^* brings in $\gamma(u)$, $a(u)$, $c(u, v, \tau)$, $-H_u(-\rho)$, and $(-)^n S_n(-\rho_1, \cdots, -\rho_n)$ in place of $\gamma^*(u)$, $a^*(u)$, $\hat{c}(u, v, \tau)$, and $H_u^*(\rho)$, $S^*(\rho_1, \cdots, \rho_n)$. The result is

$$\begin{aligned}\langle K_1K_2\rangle_{\text{av}} &= [\langle K_1\rangle_{\text{av}}]^2 + \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\, \gamma(u)\gamma(v) \exp[a(u) + a(v)]\\ &\qquad \cdot \{\exp[2c(u, v, \tau)] - 1\}\\ &= [\langle K_1\rangle_{\text{av}}]^2 + \sum_{n=1}^{\infty} \frac{(-)^n}{n!} \int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_n \left[\prod_{k=1}^{n} W_\varphi(\rho_k)\right]\\ &\qquad \cdot \exp[i2\pi\tau(\rho_1 + \cdots + \rho_n)]\\ &\qquad \cdot S_n(-\rho_1, \cdots, -\rho_n) S_n(\rho_1, \cdots, \rho_n).\end{aligned}\tag{78}$$

The remaining portion of $A(\tau)$ in equation (77) is

$$\begin{aligned}&\langle \Phi_1(2^{-1}\Phi_2 - iK_2)\rangle_{\text{av}}\\ &\quad = \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\, \gamma(u)\gamma(v)\langle 2^{-1}\varphi(t-u)\varphi(t+\tau-v) - i\varphi(t-u)J(v, \tau)\rangle_{\text{av}}\\ &\quad = \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\, \gamma(u)\gamma(v)\\ &\qquad \cdot \int_{-\infty}^{\infty} df\, W_\varphi(f) \exp[i\omega(\tau + u)]\{2^{-1}\exp[-i\omega v] + H_v(f)\exp[a(v)]\}\end{aligned}\tag{79}$$

where $\omega = 2\pi f$ and Table II has been used in going from the first equation to the second. Integration with respect to u brings in $\Gamma(-f)$, and integration with respect to v brings in both $\Gamma(f)$ and the function $S_1(f)$ of Appendix C.

$$\langle \Phi_1(2^{-1}\Phi_2 - iK_2)\rangle_{\text{av}} = \int_{-\infty}^{\infty} df\, W_\varphi(f) \exp(i\omega\tau)\Gamma(-f)[2^{-1}\Gamma(f) + S_1(f)].\tag{80}$$

Replacing Φ_1, $\gamma(u)$ by Φ_1^*, $\gamma^*(u)$ causes $\Gamma^*(f)$ to appear in place of

$\Gamma(-f)$ and shows that the remaining portion of $B(\boldsymbol{\tau})$ in equation (77) is

$$\langle \Phi_1^*(2^{-1}\Phi_2 - iK_2)\rangle_{\text{av}} = \int_{-\infty}^{\infty} df\, W_\varphi(f) \exp(i\omega\tau)\Gamma^*(f)[2^{-1}\Gamma(f) + S_1(f)]. \tag{81}$$

The function $A(\tau)$ is the sum of $-2^{-1}\langle K_1K_2\rangle$, obtained from equation (78), and (80). Its Fourier transform is

$$\begin{aligned} P(f) = {} & W_\varphi(f)\Gamma(-f)[2^{-1}\Gamma(f) + S_1(f)] - 2^{-1}[\langle K_1\rangle_{\text{av}}]^2\,\delta(f) \\ & - \frac{1}{2}\int_{-\infty}^{\infty} d\tau \exp(-i2\pi f\tau)\int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\, \gamma(u)\gamma(v) \\ & \cdot \exp[a(u) + a(v)]\{\exp[2c(u, v, \tau)] - 1\} \end{aligned} \tag{82}$$

where the leading term follows immediately from equation (80) and the Fourier integral theorem.

The function $B(\tau)$ is the sum of $2^{-1}\langle K_1^*K_2\rangle_{\text{av}}$, obtained from equation (69), and (81). Its Fourier transform is

$$\begin{aligned} Q(f) = {} & W_\varphi(f)\Gamma^*(f)[2^{-1}\Gamma(f) + S_1(f)] + 2^{-1}\,|\,\langle K_1\rangle_{\text{av}}\,|^2\,\delta(f) \\ & + \frac{1}{2}\int_{-\infty}^{\infty} d\tau \exp(-i2\pi f\tau)\int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\, \gamma^*(u)\gamma(v) \\ & \cdot \exp[a^*(u) + a(v)]\{\exp[2\hat{c}(u, v, \tau)] - 1\}. \end{aligned} \tag{83}$$

A first order approximation for $W_\theta(f)$ may be obtained by combining equations (76), (82), and (83). Deleting the terms multiplied by $\delta(f)$ and $W_\varphi(f)$ gives the first order approximation to the power spectrum of the nonlinear distortion $\theta_{nl}(t)$. This approximation is stated by equations (38), (39), and (40) in the section describing the results. We now proceed to express the first order approximation for $W_\theta(f)$ as the series given by equation (90).

When equation (82) for $P(f)$ is added to equation (83) for $Q(f)$ and the triple integrals replaced by their series, namely, the Fourier transforms of the series appearing in equations (78) and (72), the result is

$$\begin{aligned} P(f) + Q(f) = {} & W_\varphi(f)[\Gamma(-f) + \Gamma^*(f)] \\ & \cdot[2^{-1}\Gamma(f) + S_1(f)] + 2^{-1}(|\,\langle K_1\rangle_{\text{av}}\,|^2 - [\langle K_1\rangle_{\text{av}}]^2)\,\delta(f) \\ & + \frac{1}{2}\sum_{n=1}^{\infty}\frac{1}{n!}\int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_n\, \delta(f - \rho_1 - \cdots - \rho_n)\left[\prod_{k=1}^{n} W_\varphi(\rho_k)\right] \\ & \cdot S_n(\rho_1, \cdots, \rho_n)[-(-)^n S_n(-\rho_1, \cdots, -\rho_n) + S_n^*(\rho_1, \cdots, \rho_n)]. \end{aligned} \tag{84}$$

Changing the signs of f and the variables of integration $\rho_1, \cdots, \rho_n$, and then taking the conjugate complex shows that the series in the expression for $P^*(-f) + Q^*(-f)$ differs from equation (84) only in that the S_n factors are replaced by

$$S_n^*(-\rho_1, \cdots, -\rho_n)[-(-)^n S_n^*(\rho_1, \cdots, \rho_n) + S_n(-\rho_1, \cdots, -\rho_n)]. \tag{85}$$

Taking $(-)^{n-1}$ out of the square brackets in equation (85) and then adding the series term in $P(f) + Q(f)$ to the series term in $P^*(-f) + Q^*(-f)$ gives

$$\frac{1}{2} \sum_{n=1}^{\infty} \frac{1}{n!} \int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_n \, \delta(f - \rho_1 - \cdots - \rho_n) \left[\prod_{k=1}^{n} W_\varphi(\rho_k) \right] \cdot | S_n(\rho_1, \cdots, \rho_n) - (-)^n S_n^*(-\rho_1, \cdots, -\rho_n) |^2 \tag{86}$$

for the series term in $P(f) + Q(f) + P^*(-f) + Q^*(-f)$.

The term for $n = 1$ in the series of equation (86) is

$$\tfrac{1}{2} W_\varphi(f) \, | S_1(f) + S_1^*(-f) |^2. \tag{87}$$

From the first line in equation (84), the sum of the other terms in $P(f) + Q(f) + P^*(-f) + Q^*(-f)$ containing the factor $W_\varphi(f)$ is

$$W_\varphi(f)[\Gamma(-f) + \Gamma^*(f)] \cdot [2^{-1}\Gamma(f) + 2^{-1}\Gamma^*(-f) + S_1(f) + S_1^*(-f)]. \tag{88}$$

The real part of the sum of equations (87) and (88) may be written as

$$\tfrac{1}{2} W_\varphi(f) \, | \Gamma(f) + \Gamma^*(-f) + S_1(f) + S_1^*(-f)|^2 \tag{89}$$

These results and equation (76) for $W_\theta(f)$ lead to

$$W_\theta(f) = \theta_{dc}^2 \, \delta(f) + 4^{-1} W_\varphi(f) \, | \Gamma(f) + S_1(f) + \Gamma^*(-f) + S_1^*(-f) |^2 + 4^{-1} \sum_{n=2}^{\infty} \frac{1}{n!} \int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_n \, \delta(f - \rho_1 - \cdots - \rho_n) \left[\prod_{k=1}^{n} W_\varphi(\rho_k) \right] \cdot | S_n(\rho_1, \cdots, \rho_n) - (-)^n S_n^*(-\rho_1, \cdots, -\rho_n) |^2 + O(\varphi^4 W_\varphi). \tag{90}$$

The remainder in equation (90) for $W_\theta(f)$ is $O(\varphi^4 W_\varphi)$ while the one in the main result, that is, equation (15), is $O(\varphi^6 W_\varphi)$. The result of neglecting all $O(\varphi^4 W_\varphi)$ terms in equation (90) agrees with the result obtained by neglecting the $O(\varphi^4 W_\varphi)$ terms in the main result. This may be verified with the help of

$$\begin{aligned} S_1(f) &= -\frac{1}{2}\int_{-\infty}^{\infty} d\rho\, W_\varphi(\rho) S(\rho, -\rho, f) + O(\varphi^4) \\ S_2(\rho, \nu) &= S(\rho, \nu) + O(\varphi^2) \end{aligned} \tag{91}$$

which follow from the equations at the end of Appendix C.

VIII. "SECOND ORDER" APPROXIMATION FOR THE POWER SPECTRUM OF $\theta(t)$

In Section VII the "first order" approximation to $W_\theta(f)$ is computed using the approximation

$$\Theta(t) = \Phi(t) - iK(t) + O(\varphi^4) \tag{92}$$

for the complex phase angle $\Theta(t)$. In this section the "second order" approximation to $W_\theta(f)$ will be computed using the approximation given by equation (13),

$$\Theta(t) = \Phi(t) - iK(t) + \frac{i}{2}K^2(t) + O(\varphi^6). \tag{93}$$

The equations needed are given in Section IV. Portions of the ensemble averages $A(\tau)$, $B(\tau)$ defined by equation (49) have already been obtained in Sections VI and VII. The remaining portions needed are

$$\langle i2^{-1}\Phi_1 K_2^2\rangle_{\text{av}}\,, \qquad \langle 2^{-1}K_1K_2^2\rangle_{\text{av}}\,, \tag{94}$$

for $A(\tau)$ and

$$\langle i2^{-1}\Phi_1^* K_2^2\rangle_{\text{av}}\,, \qquad \langle -2^{-1}K_1^*K_2^2\rangle_{\text{av}} \tag{95}$$

for $B(\tau)$.

From Table II,

$$\begin{aligned} \langle i2^{-1}\Phi_1 K_2^2\rangle_{\text{av}} &= i2^{-1}\int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv \int_{-\infty}^{\infty} dw\, \gamma(u)\gamma(v)\gamma(w) \\ &\qquad \cdot \langle \varphi(t-u)[J(v,\tau)J(w,\tau) - J(v,\tau) - J(w,\tau) + 1]\rangle_{\text{av}} \\ &= -2^{-1}\int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv \int_{-\infty}^{\infty} dw\, \gamma(u)\gamma(v)\gamma(w) \\ &\qquad \cdot \int_{-\infty}^{\infty} df\, W_\varphi(f) \exp\,[i\omega(\tau + u)][2H_v(f)][VWz - V], \\ &\qquad\qquad \omega = 2\pi f. \end{aligned} \tag{96}$$

In going from the first to the second equation, symmetry in v and w has been used to replace $H_v(f) + H_w(f)$ by $2H_v(f)$ and we have introduced part of the notation

$$
\begin{aligned}
U &= \exp\,[a(u)], & V &= \exp\,[a(v)], & W &= \exp\,[a(w)] \\
x &= \exp\,[2c(u, v, \tau)], & y &= \exp\,[2c(u, w, \tau)], & z &= \exp\,[2c(w, v, o)] \\
\hat{x} &= \exp\,[2\hat{c}(u, v, \tau)], & \hat{y} &= \exp\,[2\hat{c}(u, w, \tau)], & U^* &= \exp\,[a^*(u)].
\end{aligned}
\tag{97}
$$

As in equation (80), integration with respect to u brings in $\Gamma(-f)$, and integration with respect to v and w brings in the functions $S_{10}(f;)$, $S_1(f)$ of Appendix C:

$$\langle i2^{-1}\Phi_1 K_2^2\rangle_{\text{av}} = -\int_{-\infty}^{\infty} df\, \exp\,(i\omega\tau) W_\varphi(f)\Gamma(-f)[S_{10}(f;) - S_1(f)]. \tag{98}$$

The corresponding portion of $B(\tau)$, $\langle i2^{-1}\Phi_1^* K_2^2\rangle_{\text{av}}$, is equal to the expressions obtained when $\gamma(u)$ and $\Gamma(-f)$ are replaced by $\gamma^*(u)$ and $\Gamma^*(f)$ in the right sides of equations (96) and (98).

The last portion of $A(\tau)$ is

$$
\begin{aligned}
\langle 2^{-1}K_1K_2^2\rangle_{\text{av}} &= 2^{-1}\int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv \int_{-\infty}^{\infty} dw\, \gamma(u)\gamma(v)\gamma(w) \\
&\qquad \cdot\langle [J(u, o) - 1][J(v, \tau) - 1][J(w, \tau) - 1]\rangle_{\text{av}} \\
&= C_1 + D_1(\tau)
\end{aligned}
\tag{99}
$$

where C_1 is independent of τ and represents the value of equation (99) at $\tau = \infty$. With the help of Table I and the notation defined in equation (97), the ensemble average in the integrand may be written as

$$UVWxyz - UVx - UWy - VWz + U + V + W - 1. \tag{100}$$

The only variables in this expression which contain τ are x and y. When $\tau \to \infty$, $c(u, v, \tau)$ tends to 0 and x and y tend to 1. Therefore the portion of equation (100) which contributes to C_1 is

$$UVWz - UV - UW - VWz + U + V + W - 1$$

and the portion contributing to $D_1(\tau)$ is the remainder

$$UVWz(xy - 1) - UV(x - 1) - UW(y - 1). \tag{101}$$

The portion contributing to C_1 will be ignored since the Fourier transform of C_1, namely $C_1\,\delta(f)$, is part of $\theta_{dc}^2\,\delta(f)$, and θ_{dc} will be treated by itself.

When $xy - 1$ is written as $(x - 1)(y - 1) + (x - 1) + (y - 1)$

equation (101) becomes

$$UVWz(x-1)(y-1) + UV(Wz-1)(x-1) + UW(Vz-1)(y-1).$$

The symmetry in v and w allows the last summand to be replaced by the second and hence

$$D_1(\tau) = \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv \int_{-\infty}^{\infty} dw\, \gamma(u)\gamma(v)\gamma(w) \cdot [2^{-1}UVWz(x-1)(y-1) + UV(Wz-1)(x-1)]. \tag{102}$$

Expanding $(x-1)$, $(y-1)$ in powers of $c(u, v, \tau)$, $c(u, w, \tau)$, respectively, and integrating termwise in much the same way as in the passage from equation (69) to equation (72) leads to

$$\begin{aligned} D_1(\tau) = {} & \sum_{n=1}^{\infty} \frac{(-)^n}{n!} \int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_n \left[\prod_{k=1}^{n} W_{\varphi}(\rho_k) \right] \\ & \cdot \exp [i2\pi\tau(\rho_1 + \cdots + \rho_n)] \\ & \cdot S_n(-\rho_1, \cdots, -\rho_n)[S_{n0}(\rho_1, \cdots, \rho_n;) - S_n(\rho_1, \cdots, \rho_n)] \\ & + \frac{1}{2} \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \frac{(-)^{n+m}}{n!\, m!} \int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_n \\ & \cdot \int_{-\infty}^{\infty} d\sigma_1 \cdots \int_{-\infty}^{\infty} d\sigma_m \left[\prod_{k=1}^{n} W_{\varphi}(\rho_k) \right] \left[\prod_{\ell=1}^{m} W_{\varphi}(\sigma_k) \right] \\ & \cdot \exp [i2\pi\tau(\rho_1 + \cdots + \rho_n + \sigma_1 + \cdots + \sigma_m)] \\ & \cdot S_{n+m}(-\rho_1, \cdots, -\rho_n, -\sigma_1, \cdots, -\sigma_m) \\ & \cdot S_{nm}(\rho_1, \cdots, \rho_n; \sigma_1, \cdots, \sigma_m). \end{aligned} \tag{103}$$

When the last portion of $B(\tau)$ is written as

$$\langle -2^{-1}K_1^*K_2^2 \rangle_{\text{av}} = C_2 + D_2(\tau) \tag{104}$$

the work goes through much as for $C_1 + D_1(\tau)$. The functions $\gamma^*(u)$, $J^*(u, o)$, U^*, $\hat{x}$, and $\hat{y}$ replace $\gamma(u)$, $J(u, o)$, U, x, and y, respectively. The functions $H_u(-\rho_k)$, and $H_u(-\sigma_\ell)$ in $c(u, v, \tau)$, $c(u, w, \tau)$ are replaced by $-H_u^*(\rho_k)$, and $-H_u^*(\sigma_\ell)$. This carries x, y into $\hat{x}$, $\hat{y}$ and causes $S_n(\rho_1, \cdots, -\rho_n)$ to be replaced by $(-)^n S_n^*(\rho_1, \cdots, \rho_n)$. A similar replacement holds for S_{n+m}.

The resulting expression for $D_2(\tau)$ is obtained by changing the sign (because $-K_1^*$ replaces K_1) of the expression (103) for $D_1(\tau)$, and then replacing $S_n(-\rho_1, \cdots -\rho_n)$ and $S_{n+m}(-\rho_1, \cdots, -\sigma_m)$ by $(-)^n S_n^*(\rho_1, \cdots, \rho_n)$ and $(-)^{n+m} S_{n+m}^*(\rho_1, \cdots, \sigma_m)$, respectively.

Now that expressions for the portions (94) and (95) of $A(\tau)$ and $B(\tau)$ have been obtained (in effect), there remain two problems:

(*i*) taking their Fourier transforms to get their contributions to $P(f)$ and $Q(f)$, and

(*ii*) adding these to the first order approximation for $W_\theta(f)$ given by equation (90).

The Fourier transform of $\langle i2^{-1}\Phi_1 K_2^2\rangle_{\text{av}}$ follows from equation (98) and the Fourier integral theorem. The Fourier transform of $\langle i2^{-1}\Phi_1^* K_2^2\rangle_{\text{av}}$ may be obtained in much the same way and consequently the contribution of these terms to $P(f) + Q(f)$ is

$$\int_{-\infty}^{\infty} d\tau \exp(-i\omega\tau)[\langle i2^{-1}\Phi_1 K_2^2\rangle_{\text{av}} + \langle i2^{-1}\Phi_1^* K_2^2\rangle_{\text{av}}]$$

$$= -W_\varphi(f)[\Gamma(-f) + \Gamma^*(f)][S_{10}(f;) - S_1(f)]. \qquad (105)$$

Consequently their contribution to the right side of

$$W_\theta(f) = 2^{-1} \operatorname{Re}[P(f) + Q(f) + P^*(-f) + Q^*(-f)] + O(\varphi^6 W_\varphi) \qquad (106)$$

[from equation (50)] is

$$2^{-1}W_\varphi(f) \operatorname{Re}[\Gamma(-f, + \Gamma^*(f)][S_1(f) - S_{10}(f;) + S_1^*(-f) - S_{10}^*(-f;)]. \qquad (107)$$

The Fourier transform of the portion $D_1(\tau)$ of $\langle 2^{-1}K_1K_2^2\rangle_{\text{av}}$ is obtained by replacing $\exp[i2\pi\tau(\rho_1 + \cdots + \rho_n)]$ and $\exp[i2\pi\tau(\rho_1 + \cdots + \sigma_m)]$ in equation (103) for $D_1(\tau)$ by $\delta(f - \rho_1 - \cdots - \rho_n)$ and $\delta(f - \rho_1 - \cdots - \sigma_m)$, respectively. The Fourier transform of $D_2(\tau)$, from $\langle -2^{-1}K_1^*K_2^2\rangle_{\text{av}}$, can be obtained similarly. The sum of these two Fourier transforms gives the contribution of $D_1(\tau) + D_2(\tau)$ to $P(f) + Q(f)$. Changing the signs of f and the variables of integration $\rho_1, \cdots, \sigma_m$, and then taking the conjugate complex, gives the contribution of $D_1(\tau) + D_2(\tau)$ to $P^*(-f) + Q^*(-f)$. When the two contributions are added, it is found that the contribution of $D_1(\tau) + D_2(\tau)$ to $P(f) + Q(f) + P^*(-f) + Q^*(-f)$ is

$$\sum_{n=1}^{\infty} \frac{1}{n!} \int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_n \left[\prod_{k=1}^{n} W_\varphi(\rho_k)\right]$$

$$\cdot\, \delta(f - \rho_1 - \cdots - \rho_n)[S_n^*(\rho_1, \cdots) - (-)^n S_n(-\rho_1, \cdots)]$$

$$\cdot\, [S_n(\rho_1, \cdots) - (-)^n S_n^*(-\rho_1, \cdots)$$

$$- S_{n0}(\rho_1, \cdots) + (-)^n S_{n0}^*(-\rho_1, \cdots)]$$

$$
\begin{aligned}
&-\frac{1}{2}\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{1}{n!\,m!}\int_{-\infty}^{\infty}d\rho_1\cdots\int_{-\infty}^{\infty}d\sigma_m\\
&\cdot\left[\prod_{k=1}^{n}W_{\varphi}(\rho_k)\right]\left[\prod_{l=1}^{m}W_{\varphi}(\sigma_l)\right]\delta(f-\rho_1-\cdots-\sigma_m)\\
&\cdot[S^*_{n+m}(\rho_1\,,\,\cdots)-(-)^{n+m}S_{n+m}(-\rho_1\,,\,\cdots)]\\
&\cdot[S_{nm}(\rho_1\,,\,\cdots)-(-)^{n+m}S^*_{nm}(-\rho_1\,,\,\cdots)] \qquad (108)
\end{aligned}
$$

where the complete arguments of the S functions are shown in equation (103).

The desired expression corresponding to equation (106) for $W_\theta(f)$ is obtained by adding the significant terms in the first order approximation of equation (90) for $W_\theta(f)$ to the second order terms given by equations (107) and (108). The remainder term, $O(\varphi^4 W_\varphi)$, in equation (90) can be ignored because the significant terms are obtained from $\Phi(t) - iK(t)$ without approximation [compare equations (92) and (93) for $\Theta(t)$]. The result is

$$
\begin{aligned}
W_\theta(f) &= \theta_{dc}^2\,\delta(f) + 4^{-1}W_\varphi(f)\mid\Gamma(f)+S_1(f)+\Gamma^*(-f)+S_1^*(-f)\mid^2\\
&\quad+4^{-1}\sum_{n=2}^{\infty}\frac{1}{n!}\int_{-\infty}^{\infty}d\rho_1\cdots\int_{-\infty}^{\infty}d\rho_n\,\delta(f-\rho_1-\cdots-\rho_n)\left[\prod_{k=1}^{n}W_\varphi(\rho_k)\right]\\
&\quad\cdot\mid S_n(\rho_1\,,\,\cdots\,,\,\rho_n)-(-)^nS_n^*(-\rho_1\,\cdots\,-\rho_n)\mid^2\\
&\quad+\text{expression (107)}+2^{-1}\,\mathrm{Re}[\text{expression (108)}]+O(\varphi^6W_\varphi). \qquad (109)
\end{aligned}
$$

The next section is concerned with the elimination of all $O(\varphi^6 W_\varphi)$ terms from the significant portion of equation (109). When these terms are eliminated, the result is the "main result" stated in equation (15).

IX. ELIMINATION OF HIGHER ORDER MODULATION TERMS FROM $W_\theta(f)$

In this section all terms of $O(\varphi^6 W_\varphi)$ in equation (109) for $W_\theta(f)$ will be discarded, that is modulation terms of order higher than three will be discarded. Since the integral of $W_\varphi(f)$ is $O(\varphi^2)$, all terms in equation (109) containing the product of four or more W_φ's may be dropped immediately.

First consider the terms which explicitly contain the product of three $W\varphi$'s. This corresponds to $n = 3$ in the single series in expressions (108) and (109), and to the pairs of values $n = 1$, $m = 2$; $n = 2$, $n = 1$ in the double series. The contribution of the double series can be discarded because it is $O(\int\int W_\varphi^3\varphi^2) = O(\varphi^6 W_\varphi)$, the functions S_{21} and S_{12} being $O(\varphi^2)$ [from $S(f) = 0$, $S_1(f) = O(\varphi^2)$ and Appendix C]. The

$n = 3$ term in expression (108) can also be discarded because, from Appendix C, $S_3 - S_{30}$ is $O(\varphi^2)$. Using $S_3 - S = O(\varphi^2)$ in the $n = 3$ term in equation (109) shows that the contribution to $W_\theta(f)$ of the terms which explicitly contain the product of three W_φ's is

$$\frac{1}{24}\int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma\, W_\varphi(\rho)W_\varphi(\sigma)W_\varphi(\nu) \cdot | S(\rho, \sigma, \nu) + S^*(-\rho, -\sigma, -\nu) |^2 + O(\varphi^6 W_\varphi) \qquad (110)$$

where $\nu = f - \rho - \sigma$.

Next consider terms which explicitly contain the product of two W_φ's, namely the terms $n = 2$ in the single series and $n = m = 1$ in the double series. When we put $\rho_1 = \rho$, $\rho_2 = f - \rho = \nu$ in the single series terms, and $\rho_1 = \rho$, $\sigma_1 = f - \rho = \nu$ in the double series term, all of the integrands contain the factor

$$\beta^* = [S_2^*(\rho, \nu) - S_2(-\rho, -\nu)]$$

and the contribution of their sum to $W_\theta(f)$ can be written as

$$\frac{1}{8}\int_{-\infty}^{\infty} d\rho\, W_\varphi(\rho)W_\varphi(\nu) \text{ Re } [\beta^*(\beta + 2\gamma)]$$

where β is $O(1)$, and γ is not the earlier $\gamma(u)$. Here

$$\gamma = S_2(\rho, \nu) - S_{20}(\rho, \nu;) - S_{11}(\rho; \nu) - S_2^*(-\rho, -\nu) + S_{20}^*(-\rho, -\nu;) + S_{11}^*(-\rho; -\nu)$$

is $O(\varphi^2)$ since both $S_2 - S_{20}$ and S_{11} are $O(\varphi^2)$. Furthermore,

$$\text{Re } [\beta^*(\beta + 2\gamma)] = | \beta + \gamma |^2 - \gamma\gamma^* = | \beta + \gamma |^2 + O(\varphi^4),$$
$$\beta + \gamma = \hat{T}(\rho, \nu) - \hat{T}^*(-\rho, -\nu), \qquad (111)$$
$$\hat{T}(\rho, \nu) = 2S_2(\rho, \nu) - S_{20}(\rho, \nu;) - S_{11}(\rho, \nu).$$

The equations at the end of Appendix C may be used to show that $\hat{T}(\rho, \nu)$ is equal to $T(\rho, \nu) + O(\varphi^4)$ where

$$T(\rho, \nu) = S(\rho, \nu) + \int_{-\infty}^{\infty} d\sigma\, W_\varphi(\sigma)[-\tfrac{1}{2}S(\sigma, -\sigma, \rho, \nu) + \tfrac{1}{2}S(\sigma, -\sigma)S(\rho, \nu) + S(\sigma, \rho)S(-\sigma, \nu)] \qquad (112)$$

and consequently the contribution to $W_\theta(f)$ of the terms which explicitly contain the product of two W_φ's is

$$\frac{1}{8}\int_{-\infty}^{\infty} d\rho\, W_\varphi(\rho)W_\varphi(\nu)\mid T(\rho,\nu) - T^*(-\rho,-\nu)\mid^2 + O(\varphi^6 W_\varphi) \tag{113}$$

where $\nu = f - \rho$.

Now consider the terms in $W_\theta(f)$ which are multiplied by $W_\varphi(f)$. From equations (109), (107), and the term for $n = 1$ in the single series in equation (108), the sum of these terms is $W_\varphi(f)$ multiplied by

$$\begin{aligned}
&4^{-1}\mid \Gamma(f) + S_1(f) + \Gamma^*(-f) + S_1^*(-f)\mid^2 \\
&\quad + 2^{-1}\,\mathrm{Re}\,[\Gamma(-f) + \Gamma^*(f)][S_1(f) - S_{10}(f;) + S_1^*(-f) - S_{10}^*(-f;)] \\
&\quad + 2^{-1}\,\mathrm{Re}\,[S_1^*(f) + S_1(-f)][S_1(f) - S_{10}(f;) + S_1^*(-f) - S_{10}^*(-f;)] \\
&= 4^{-1}\mid \alpha + \beta\mid^2 + 2^{-1}\,\mathrm{Re}\,(\alpha^* + \beta^*)(\gamma) \\
&= 4^{-1}[(\alpha + \beta)(\alpha^* + \beta^*) + (\alpha^* + \beta^*)\gamma + (\alpha + \beta)\gamma^* + \gamma\gamma^* - \gamma\gamma^*] \\
&= 4^{-1}\mid \alpha + \beta + \gamma\mid^2 - 4^{-1}\gamma\gamma^* \\
&= 4^{-1}\mid \hat{U}(f) + \hat{U}^*(-f)\mid^2 - 4^{-1}\gamma\gamma^*.
\end{aligned} \tag{114}$$

Here, with β and γ different from those in equation (111),

$$\begin{aligned}
\alpha &= \Gamma(f) + \Gamma^*(-f) = O(1) \\
\beta &= S_1(f) + S_1^*(-f) = O(\varphi^2) \\
\gamma &= S_1(f) - S_{10}(f;) + S_1^*(-f) - S_{10}^*(-f;) = O(\varphi^4) \\
\hat{U}(f) &= \Gamma(f) + 2S_1(f) - S_{10}(f;).
\end{aligned} \tag{115}$$

The equations at the end of Appendix C may be used to show that $\hat{U}(f)$ is equal to $U(f) + O(\varphi^6)$ where

$$\begin{aligned}
U(f) = \Gamma(f) - \frac{1}{2}\int_{-\infty}^{\infty} d\rho\, W_\varphi(\rho)S(\rho, -\rho, f) + \int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma\, W_\varphi(\rho)W_\varphi(\sigma) \\
\cdot[\tfrac{1}{8}S(\rho, \sigma, -\rho, -\sigma, f) - \tfrac{1}{4}S(\rho, -\rho, f)S(\sigma, -\sigma) \\
- \tfrac{1}{2}S(\rho, f)S(\sigma, -\sigma, -\rho) - \tfrac{1}{2}S(\rho, \sigma, f)S(-\rho, -\sigma)].
\end{aligned} \tag{116}$$

Since $\gamma\gamma^*$ is $O(\varphi^8)$, the terms in $W_\theta(f)$ which are multiplied by $W_\varphi(f)$ can be written as

$$W_\varphi(f)\,|U(f) + U^*(-f)|^2 + O(\varphi^6 W_\varphi). \tag{117}$$

Finally consider the dc spike, $\theta_{dc}^2\,\delta(f)$, in $W_\theta(f)$. From equation (14), the dc component of $\theta(t)$ is

$$\theta_{dc} = \mathrm{Im}\,\langle K(t) - 2^{-1}K^2(t)\rangle_{\mathrm{av}} + O(\varphi^6). \tag{118}$$

The value of $\langle K(t)\rangle_{\mathrm{av}}$ is given by equation (70) and from equation (78) with $\tau = 0$,

$$\langle K^2(t)\rangle_{\mathrm{av}} = \langle K(t)\rangle_{\mathrm{av}}^2 - \int_{-\infty}^{\infty} d\rho\, W_\varphi(\rho)S_1(-\rho)S_1(\rho)$$

$$+\frac{1}{2}\int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma\, W_\varphi(\rho)W_\varphi(\sigma)S_2(-\rho,\, -\sigma)S_2(\rho,\, \sigma) + O(\varphi^6).$$

Since $S_1(\rho)$ is $O(\varphi^2)$, the single integral is $O(\varphi^6)$ and may be included in the remainder term. Squaring the leading term in equation (70) to get $\langle K(t)\rangle^2$, combining terms, and using $S_2(\rho,\, \sigma) = S(\rho,\, \sigma) + O(\varphi^2)$ leads to

$$\langle K(t) - 2^{-1}K^2(t)\rangle_{\mathrm{av}} = D_c$$

$$D = -\frac{1}{2}\int_{-\infty}^{\infty} d\rho\, W_\varphi(\rho)S(\rho,\, -\rho) + \int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma\, W_\varphi(\rho)W_\varphi(\sigma)$$

$$\cdot[\tfrac{1}{8}S(\rho,\, \sigma,\, -\rho,\, -\sigma) - \tfrac{1}{8}S(\rho,\, -\rho)S(\sigma,\, -\sigma) - \tfrac{1}{4}S(\rho,\, \sigma)S(-\rho,\, -\sigma)]$$

$$+ O(\varphi^6). \tag{119}$$

The imaginary part of D_c gives θ_{de}.

Addition of equations (110), (113), and (117) shows that

$$W_\theta(f) = \theta_{dc}^2\, \delta(f) + W_\varphi(f) \mid U(f) + U^*(-f) \mid^2$$

$$+\frac{1}{8}\int_{-\infty}^{\infty} d\rho\, W_\varphi(\rho)W_\varphi(f - \rho) \mid T(\rho,\, f - \rho) - T^*(-\rho,\, -f + \rho)\mid^2$$

$$+\frac{1}{24}\int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma\, W_\varphi(\rho)W_\varphi(\sigma)W_\varphi(\nu)$$

$$\cdot\mid S(\rho,\, \sigma,\, \nu) + S^*(-\rho,\, -\sigma,\, -\nu) \mid^2 + O(\varphi^6 W_\varphi) \tag{120}$$

where $\nu = f - \rho - \sigma$. This is the same as equation (15) in the statement of results. However, the expressions for D_c, $T(\rho,\, f - \rho)$, and $U(f)$ given in Section III are simpler than the ones given in this section. The method of obtaining the simpler expressions will be outlined in Section X.

X. SIMPLIFIED EXPRESSIONS FOR θ_{dc}, $U(f)$, AND $T(\rho,\, \nu)$

The expressions obtained for θ_{dc}, $U(f)$, and $T(\rho,\, \nu)$ in Section IX may be put in forms better suited to calculation by writing the higher order S functions in terms of S functions of two arguments,

$$S(\rho,\, \sigma) = \Gamma(\rho + \sigma) - \Gamma(\rho)\Gamma(\sigma). \tag{121}$$

These simplified forms are the ones stated in equations (16), (17), and (18).

Since no really satisfactory procedure of reduction was found, the expressions given here may not be the simplest. The procedure is illustrated for the double integral

$$I = \int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma \, W_\varphi(\rho)W_\varphi(\sigma)$$

$$\cdot [S(\rho, -\rho, \sigma, -\sigma) - S(\rho, -\rho)S(\sigma, -\sigma) - 2S(\rho, \sigma)S(-\rho, -\sigma)]$$

which appears in equation (119) for $\theta_{dc} = \text{Im } D_c$.

After some cancellation, the general equation (159) for $S(\rho, \sigma, \nu, \mu)$ shown in Appendix B gives

$$\begin{aligned} S(\rho, -\rho, \sigma, -\sigma) = 1 - (\rho)(-\rho) - (\sigma)(-\sigma) + (\rho + \sigma)(-\rho)(-\sigma) \\ + (\rho - \sigma)(-\rho)(\sigma) + (-\rho + \sigma)(\rho)(-\sigma) \\ + (-\rho - \sigma)(\rho)(\sigma) - 3(\rho)(-\rho)(\sigma)(-\sigma). \end{aligned}$$

Here $\Gamma(x)$ has been written as (x) and $(0) = \Gamma(0) = 1$ has been used. When this expression is multiplied by $W_\varphi(\rho)W_\varphi(\sigma)$ and integrated with respect to ρ and σ, changes in the variables of integration show that the value of I is unchanged by the substitution

$$\begin{aligned} S(\rho, -\rho, \sigma, -\sigma) \to 1 - 2(\rho)(-\rho) + 4(\rho + \sigma)(-\rho)(-\sigma) \\ - 3(\rho)(-\rho)(\sigma)(-\sigma). \end{aligned}$$

Here the arrow means "may be replaced in the double integral by".

Similarly,

$$\begin{aligned} -S(\rho, -\rho)S(\sigma, -\sigma) &= -[1 - (\rho)(-\rho)][1 - (\sigma)(-\sigma)] \\ &\to -1 + 2(\rho)(-\rho) - (\rho)(-\rho)(\sigma)(-\sigma) \end{aligned}$$

$$\begin{aligned} -2S(\rho, \sigma)S(-\rho, -\sigma) &= -2[(\rho + \sigma) - (\rho)(\sigma)][(-\rho - \sigma) - (-\rho)(-\sigma)] \\ &\to -2(\rho + \sigma)(-\rho - \sigma) + 4(\rho + \sigma)(-\rho)(-\sigma) - 2(\rho)(-\rho)(\sigma)(-\sigma). \end{aligned}$$

Addition shows that the quantity within the square brackets in the integrand of I may be replaced by

$$\begin{aligned} &8(\rho + \sigma)(-\rho)(-\sigma) - 2(\rho + \sigma)(-\rho - \sigma) - 6(\rho)(-\rho)(\sigma)(-\sigma) \\ &= 6(-\rho)(-\sigma)[(\rho + \sigma) - (\rho)(\sigma)] - 2(\rho + \sigma)[(-\rho - \sigma) - (-\rho)(-\sigma)] \\ &= 6(-\rho)(-\sigma)S(\rho, \sigma) - 2(\rho + \sigma)S(-\rho, -\sigma) \end{aligned}$$

$$= 6[(-\rho-\sigma)-(-\rho-\sigma)+(-\rho)(-\sigma)]S(\rho,\sigma)-2(\rho+\sigma)S(-\rho,-\sigma)$$

$$= 6(-\rho-\sigma)S(\rho,\sigma) - 6S(-\rho,-\sigma)S(\rho,\sigma) - 2(\rho+\sigma)S(-\rho,-\sigma)$$

$$\rightarrow 4(\rho+\sigma)S(-\rho,-\sigma) - 6S(-\rho,-\sigma)S(\rho,\sigma).$$

Hence

$$I = \int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma\, W_\varphi(\rho)W_\varphi(\sigma)S(-\rho,-\sigma)[4\Gamma(\rho+\sigma) - 6S(\rho,\sigma)]$$

which is the form of I used in equation (16) for D (except that ρ and σ are interchanged).

The simplification of equations (112) and (116) for $T(\rho, \nu)$ and $U(f)$, respectively, proceeds along the same lines. In dealing with $U(f)$, the symbolic substitution

$$S(\rho,\sigma,-\rho,-\sigma,f) \rightarrow [1 + (\rho)(-\rho) - 2y^\rho(-\rho)] \cdot[1 + (\sigma)(-\sigma) - 2y^\sigma(-\sigma)][y^f - (f)]$$

was found helpful.

In addition to the simplified forms for $T(\rho, \nu)$ and $U(f)$ given in equations (17) and (18), we also have

$$S(\rho,\sigma,\nu) = S(\rho+\sigma,\nu) - \Gamma(\rho)S(\sigma,\nu) - \Gamma(\sigma)S(\rho,\nu). \qquad (122)$$

XI. THE "SMALL AND SLOW" DEVIATION APPROXIMATION TO $W_\theta(f)$

This section and the following one are concerned with approximations to $W_\theta(f)$ which are obtained by replacing the Γ's used in equations (15) to (20) by the first few terms in their power series expansions. These expansions are assumed to exist and to converge rapidly over the range of frequencies for which $W_\varphi(f)$ is effectively different from zero. Roughly speaking, the top baseband frequency is assumed to be small compared with the filter bandwidth.

When the top baseband frequency is small, the modulating frequency, $\varphi'(t)$, changes slowly and we have the quasistatic case. The name "small and slow deviation approximation" is used because (15) holds only for "small" rms frequency deviations (D small), and here the requirement of "slowness" is added.

Two series which play important roles are

$$\Gamma(f) = \sum_{n=0}^{\infty} \alpha_n f^n/n!, \qquad \alpha_0 = 1 \qquad (123)$$

$$\ln \Gamma(f) = \sum_{n=0}^{\infty} \lambda_n f^n/n!, \qquad \lambda_0 = 0 \qquad (124)$$

The first one is the series assumed for $\Gamma(f)$. Substituting (123) in (124), expanding the logarithm, and equating coefficients of powers of f leads to expressions for the λ's in terms of the α's:

$$\begin{aligned}
\lambda_1 &= \alpha_1 ,\\
\lambda_2 &= \alpha_2 - \alpha_1^2 ,\\
\lambda_3 &= \alpha_3 - 3\alpha_2\alpha_1 + 2\alpha_1^3 ,\\
\lambda_4 &= \alpha_4 - 4\alpha_3\alpha_1 - 3\alpha_2^2 + 12\alpha_2\alpha_1^2 - 6\alpha_1^4 ,\\
\lambda_5 &= \alpha_5 - 5\alpha_4\alpha_1 - 10\alpha_3\alpha_2 + 20\alpha_3\alpha_1^2 + 30\alpha_2^2\alpha_1 - 60\alpha_2\alpha_1^3 + 24\alpha_1^5 .
\end{aligned} \tag{125}$$

and so on. When the α_n's are the moments of a probability distribution, the λ_n's are the associated "cumulants" or semi-invariants. In our problem the α_n's are proportional to the moments of the normalized response $\gamma(t)$, a relation which follows when the series (123) for $\Gamma(f)$ is compared with the one obtained by expanding $\exp(-i2\pi ft)$ in the Fourier integral (4) for $\Gamma(f)$.

The small and slow deviation approximation obtained from (15) and the first few terms of (123) is

$$\begin{aligned}
W_\theta(f) &\to \theta_{dc}^2\ \delta(f)\\
&+ W_\varphi(f)[1 + f^2\{(\lambda_{1i} + 2^{-1}D^2\lambda_{3i} + 8^{-1}D^4\lambda_{5i})^2 + (\alpha_{2r} + 2^{-1}D^2A_r)\}]\\
&+ 2^{-1}(\lambda_{2i} + 2^{-1}D^2\lambda_{4i})^2 \int_{-\infty}^{\infty} d\rho\, W_\varphi(\rho)W_\varphi(f - \rho)\rho^2(f - \rho)^2\\
&+ 6^{-1}(\lambda_{3i})^2 \int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma\, W_\varphi(\rho, W_\varphi(\sigma)W_\varphi(f - \rho - \sigma)\rho^2\sigma^2(f - \rho - \sigma)^2
\end{aligned} \tag{126}$$

Here the α_{ni}'s, λ_{ni}'s are the imaginary parts of the coefficients in the series (123), (124), $D^2 = \langle[\varphi'(t)/(2\pi)]^2\rangle$, D is the rms frequency deviation in Hz, and α_{2r}, A_r denote the real parts of α_2, A where

$$A = \alpha_4 - 2\alpha_3\alpha_1 - \alpha_2^2 + 2\alpha_2\alpha_1^2 .$$

The detailed derivation of equation (126) from equation (15) for $W_\theta(f)$ makes use of equation (162) which gives $S(x_1, x_2, \cdots x_n)$ for small values of the x's. The leading term in equation (162) gives

$$\begin{gathered}
S(\rho, \sigma) \to \rho\sigma\lambda_2\\
S(\rho, \sigma, \nu) \to \rho\sigma\nu\lambda_3\\
-S(\sigma, -\sigma, \rho, \nu) + S(\sigma, -\sigma)S(\rho, \nu) + 2S(\sigma, \rho)S(-\sigma, \nu) \to \sigma^2\rho\nu\lambda_4
\end{gathered} \tag{127}$$

where the left side of the last equation is proportional to the integrand in equation (112) for $T(\rho, \nu)$. There is a similar equation which shows that the integrand in the double integral in equation (116) for $U(f)$ tends to a quantity proportional to $\rho^2\sigma^2 f\lambda_5$. To deal with the single integral in $U(f)$ we use both terms in equation (162) to obtain

$$S(\rho, -\rho, f) \to -\rho^2 f\lambda_3 - 2^{-1}\rho^2 f^2 A. \tag{128}$$

Combining equations (127), (128), and the first three terms in the series for $\Gamma(f)$ gives

$$\begin{aligned} &S(\rho, \sigma, \nu) \to \rho\sigma\nu\lambda_3 \\ &\quad T(\rho, \nu) \to \rho\nu(\lambda_2 + 2^{-1}D^2\lambda_4) \\ &\qquad U(f) \to 1 + (\alpha_1 + 2^{-1}D^2\lambda_3 + 8^{-1}D^4\lambda_5)f + (2^{-1}\alpha_2 + 4^{-1}D^2A)f^2. \end{aligned}$$

Substitution in the small deviation approximation (15) for $W_\theta(f)$ then gives the small and slow deviation approximation shown in equation (126).

A form of the small and slow approximation which is more complete than (126) may be obtained by starting with the quasistatic form of equation (7) for $\Theta(t)$ instead of from the small deviation approximation (15) for $W_\theta(t)$. In the quasistatic case the instantaneous frequency $\Omega = \omega_0 + \varphi'(t)$ changes slowly and hence rms $\varphi''(t)$ is small. This leads us to replace $\varphi(t - u)$ in (7) by the equivalent expression $\varphi(t) - u\varphi'(t) + 2^{-1}u^2\varphi''(\xi)$ where ξ lies between $t - u$ and t. Let F denote the filter bandwidth and suppose that the impulse response $\gamma(u)$ is effectively 0 outside an interval of length $1/F$. Then, heuristically, the integral in (7) is given by

$$\begin{aligned} \int_{-\infty}^{\infty} \gamma(u) \exp\left[i\varphi(t - u)\right] du &= [1 + O(2^{-1}F^{-2} \text{ rms } \varphi'')] \\ &\quad \cdot \int_{-\infty}^{\infty} \gamma(u) \exp\left[i\varphi(t) - iu\varphi'(t)\right] du. \end{aligned}$$

The integral on the right is the desired quasistatic approximation. It is almost equal to the integral on the left when $2^{-1}F^{-2}$ rms $\varphi'' \ll 1$. However, for small rms frequency deviations, the contribution of $u\varphi'(t)$ may be less than the term $2^{-1}F^{-2}$ rms φ'' even though

(*i*) the latter may be $\ll 1$, and

(*ii*) despite the fact that when $\varphi(t)$ is band-limited with top frequency B we always have rms $\varphi'' \leqq (2\pi B)$ rms φ'. Therefore, in order to make the quasistatic approximation meaningful for small (as well as large) deviations, we impose the additional restriction $2^{-1}F^{-2}$ rms $\varphi''/(F^{-1}$ rms $\varphi') \ll 1$. Then, if

$$\text{rms } \varphi''/(2F^2) \ll 1, \qquad \text{rms } \varphi''/(2F \text{ rms } \varphi') \ll 1$$

the Fourier transform (4) gives the quasistatic approximations

$$\begin{aligned} \int_{-\infty}^{\infty} \gamma(u) \exp [i\varphi(t-u)]\, du &\approx \Gamma[\varphi'(t)/(2\pi)] \exp [i\varphi(t)] \\ \Theta(t) &\approx \varphi(t) - i \ln \Gamma[\varphi'(t)/(2\pi)] \end{aligned} \tag{129}$$

which are equivalent to the usual quasistatic approximation for the filter output, namely

$$s_0(t) \approx G(\Omega) \exp [i\omega_0 t + i\varphi(t)]. \tag{130}$$

D. T. Hess[13] has given a rigorous bound, roughly equivalent to rms $\varphi''/(2F^2) \ll 1$, for the error in (130).

For the flat FM baseband case discussed in Section 3.3 the above restrictions go into

$$10\, DB/F^2 \ll 1, \qquad 2B/F \ll 1$$

where D is the rms frequency deviation in Hz and B is the top baseband frequency in Hz. Notice that although the term "quasistatic" implies that rms φ'' tends to 0 in some sense or other, the requirements that the quasistatic approximations (129) and (130) hold differ from the requirement that the deviation ratio be large, a condition used in calculating the quasistatic approximation to the power spectrum of cos $[\omega_0 t + \varphi(t)]$[14] Thus, for the flat baseband case, the deviation ratio can be taken to be D/B, and this does not have to be large for (129) and (130) to hold.

To continue with the derivation of the more complete form of (126), we substitute the series (124) for ln $\Gamma(f)$ in (129) and take the real part. This gives

$$\begin{aligned} \theta(t) &\approx \varphi(t) + B(t) \\ B(t) &= \sum_{n=1}^{\infty} \lambda_{ni}[\varphi'(t)/(2\pi)]^n/n!. \end{aligned}$$

Since $B(t)$ depends only on $\varphi'(t)$, the power spectrum of $\varphi(t) + B(t)$ is $W_\varphi(f) + W_B(f)$ (Ref. 15). The covariance of $B(t)$ is

$$\langle B_1 B_2 \rangle = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \lambda_{ni}\lambda_{mi}\langle \varphi_1'^n \varphi_2'^m \rangle (2\pi)^{-n-m}/(n!\, m!) \tag{131}$$

where subscripts 1 and 2 denote arguments t and $t + \tau$, respectively.

From the characteristic function of the joint gaussian distribution function of φ_1' , φ_2' we have

$$\frac{\langle \varphi_1'^n \varphi_2'^m \rangle}{n!\, m!} = (2\pi i)^{-2} \int^{(0+)} \frac{du}{u} \int^{(0+)} \frac{dv}{v} \frac{\exp\left[-\psi_o(u^2 + v^2)2^{-1} - \psi_\tau uv\right]}{i^{n+m} u^n v^m}$$

$$= \sum_{k=0}^{\infty} \frac{[1 + (-)^{n-k}][1 + (-)^{m-k}]}{k!\, 4\Gamma[(n-k)2^{-1} + 1]\Gamma[(m-k)2^{-1} + 1]} \left(\frac{2\psi_\tau}{\psi_o}\right)^k \left(\frac{\psi_o}{2}\right)^{(n+m)/2} \tag{132}$$

where $\psi_\tau = \langle \varphi_1' \varphi_2' \rangle$, $\psi_0 = (2\pi D)^2$, and we have used, for integer l,

$$(2\pi i)^{-1} \int^{(0+)} u^{-\ell-1} \exp\left[-\psi_o u^2/2\right] du = \frac{(i^\ell + i^{-\ell})(\psi_o/2)^{\ell/2}}{2\Gamma(2^{-1}\ell + 1)}.$$

Actually, instead of ∞ the upper limit of summation for k in (132) is the smaller of n, m. Also the sum is 0 unless n, m are both even or both odd. When n is even, k runs over even integers; and when n is odd, k runs over odd integers. When (132) is substituted in (131), the Fourier transform of the resulting series for $\langle B_1 B_2 \rangle$ gives a series for $W_B(f)$ which leads to the more complete form of (126) we have been seeking, namely

$$W_\theta(f) \approx W_\varphi(f) + \sum_{k=0}^{\infty} \frac{1}{k!} \left[\sum_{n=0}^{\infty} \frac{\lambda_{2n+k,\,i} D^{2n}}{n!\, 2^n}\right]^2 \cdot (2\pi)^{-2k} \int_{-\infty}^{\infty} \psi_\tau^k \exp(-i2\pi f\tau)\, d\tau \tag{133}$$

The integral in (133) can be expressed as a $(k - 1)$-fold convolution of the power spectrum $W_{\varphi'}(f) = (2\pi f)^2 W_\varphi(f)$ of $\varphi'(t)$. This gives the first few terms of (126), except for the term $(\alpha_{2r} + 2^{-1}D^2 A_r)\, W_\varphi(f)$ which arises from terms neglected by (133).

Equation (133) is useful only when D is small because the summation with respect to n usually diverges. To illustrate this, consider the single pole filter for which $\Gamma(f)$ is $(1 + iff_c^{-1})^{-1}$ and λ_n is $(n - 1)!(-if_c^{-1})^n$. Equation (133), with k replaced by $2k + 1$, gives

$$W_\theta(f) \approx W_\varphi(f) + \sum_{k=0}^{\infty} \frac{1}{(2k+1)!} \left[\sum_{n=0}^{\infty} \frac{(-)^n (2n + 2k)!\, D^{2n}}{n!\, 2^n f_c^{2n+2k+1}}\right]^2 \cdot (2\pi)^{-4k-2} \int_{-\infty}^{\infty} \psi_\tau^{2k+1} \exp(-i2\pi f\tau)\, d\tau. \tag{134}$$

The quasistatic approximation for $W_\theta(f)$ obtained by starting with (130) is (see Ref. 13)

$$W_\theta(f) \approx W_\varphi(f) + \sum_{k=0}^{\infty} \frac{I_{2k}^2(2\pi D)^{-4k-2}}{(2k+1)!} \int_{-\infty}^{\infty} \psi_\tau^{2k+1} \exp(-i2\pi f\tau)\, d\tau,$$

$$I_{2k} = \int_0^\infty x^{2k} \exp\left[-\frac{x^2}{2} - \frac{x}{D} f_c\right] dx.$$

This series, which converges for all D, is of the same form as (134) in that it has I_{2k}/D^{2k+1} in place of the divergent sum with respect to n. When D becomes small, the two expressions for $W_\theta(f)$ approach equality in the sense that the sum with respect to n is the asymptotic expansion of I_{2k}/D^{2k+1}.

XII. LIOU'S APPROXIMATION FOR SECOND AND THIRD ORDER INTERCHANNEL MODULATION

It is instructive to relate our main result, equation (15) for $W_\theta(f)$, to an approximation for the interchannel modulation given by Liou.[5] Liou's approximation is equivalent to taking additional terms in the small frequency deviation approximation given in Section XI.

The interchannel modulation is represented by the portions of $W_\theta(f)$ in equation (15) which contains $T(\rho, f - \rho)$ and $S(\rho, \sigma, f - \rho - \sigma)$. Liou's approximation may be obtained by (*i*) approximating $T(\rho, f - \rho)$ by the leading term, namely $S(\rho, f - \rho)$, in equation (17) and (*ii*) expanding $S(\rho, f - \rho)$, $S(\rho, \sigma, f - \rho - \sigma)$ in powers of ρ, σ, and f out to and including degree 4. This leads to

$$\begin{aligned} T(\rho, f - \rho) \approx S(\rho, f - \rho) &= \Gamma(f) - \Gamma(\rho)\Gamma(f - \rho) \\ &= \rho(f - \rho)[\lambda_2 + f\ell_2 + f^2\ell_3 + \rho(f - \rho)\ell_4] + O(f^5) \\ S(\rho, \sigma, f - \rho - \sigma) &= \rho\sigma(f - \rho - \sigma)[\lambda_3 + f\ell_1] + O(f^5) \end{aligned} \tag{135}$$

where

$$\begin{aligned} \Gamma(f) &= \sum_{n=0}^{\infty} \alpha_n f^n/n!, \qquad \alpha_0 = 1 \\ \lambda_2 &= \alpha_2 - \alpha_1^2, \qquad \lambda_3 = \alpha_3 - 3\alpha_2\alpha_1 + 2\alpha_1^3 \\ \ell_1 &= (\alpha_4 - 2\alpha_3\alpha_1 - \alpha_2^2 + 2\alpha_2\alpha_1^2)/2 \\ \ell_2 &= (\alpha_3 - \alpha_2\alpha_1)/2 \\ \ell_3 &= (\alpha_4 - \alpha_3\alpha_1)/6 \\ \ell_4 &= 4^{-1}(\alpha_4 -- \alpha_2^2) - 3^{-1}(\alpha_4 - \alpha_3\alpha_1). \end{aligned} \tag{136}$$

Equation (162) of Appendix B gives the approximation for $S(\rho, \sigma,$

$f - \rho - \sigma)$ shown in equation (135). It does not give the higher order terms in $S(\rho, f - \rho)$ shown in equation (135). These must be calculated from the series for $\Gamma(f)$. Although the ℓ's and λ's used here are not precisely the same as those used by Liou, they are of the same general character.

When the expressions [equation (135)] for $T(\rho, f - \rho)$ and $S(\rho, \sigma, f - \rho - \sigma)$ are used in equation (15) for $W_\theta(f)$, the second and third order interchannel modulation terms are found to be

$$\begin{aligned}
&\frac{1}{2}\int_{-\infty}^{\infty} d\rho\, W_\varphi(\rho)W_\varphi(f - \rho)\rho^2(f - \rho)^2 \\
&\qquad \cdot\{[\lambda_{2i} + f^2\ell_{3i} + \rho(f - \rho)\ell_{4i}]^2 + f^2\ell_{2r}\} \\
&\qquad + \frac{1}{6}[\lambda_{3i}^2 + f^2\ell_{1r}^2]\int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma \\
&\qquad \cdot W_\varphi(\rho)W_\varphi(\sigma)W_\varphi(f - \rho - \sigma)\rho^2\sigma^2(f - \rho - \sigma)^2 \qquad (137)
\end{aligned}$$

where the second subscripts r and i denote "real part" and "imaginary part." The basic approximation used by Liou [his Eqs. (29) and (30)] may be put in this form by expressing his Fourier transforms as convolution integrals and combining terms.

APPENDIX A

Power Spectra of Real and Imaginary Parts of a Complex Random Process

Let $z(t)$ be a complex, stationary, ergodic, random process [for example the complex phase angle $\Theta(t)$] and let $x(t)$, $y(t)$ be its real and imaginary parts. We seek convenient expressions for the power spectra $W_x(t)$, $W_y(t)$ of $x(t)$ and $y(t)$ when $z(t)$ is the sum of several correlated complex random processes, say $a(t)$, $b(t)$, $c(t)$, $\cdots$. For illustration we take

$$z(t) = a(t) + b(t) + c(t) \qquad (138)$$

which corresponds to equation (13) for $\Theta(t)$ with $a(t)$, $b(t)$, and $c(t)$ in place of $\Phi(t)$, $-iK(t)$, and $i2^{-1}K^2(t)$, respectively.

Denoting functions with arguments t, $t + \tau$ by subscripts 1, 2 and using relations of the type

$$x_1 = \text{Re}\, z_1 = (z_1 + z_1^*)/2 \qquad (139)$$

leads to the following expression for the ensemble average $\langle x_1x_2\rangle_{\text{av}}$

$$\begin{aligned}\langle x_1x_2\rangle_{\text{av}} &= \langle (z_1 + z_1^*)(z_2 + z_2^*)\rangle_{\text{av}}/4 \\ &= \langle (z_1z_2 + z_1^*z_2^*) + (z_1^*z_2 + z_1z_2^*)\rangle_{\text{av}}/4 \\ &= 2^{-1}\ \text{Re}\ (\langle z_1z_2\rangle_{\text{av}} + \langle z_1^*z_2\rangle_{\text{av}}).\end{aligned} \tag{140}$$

It is convenient to write $\langle z_1z_2\rangle_{\text{av}}$ as

$$\langle z_1z_2\rangle_{\text{av}} = A(\tau) + A(-\tau) \tag{141}$$

where

$$A(\tau) = \langle \tfrac{1}{2}(a_1a_2 + b_1b_2 + c_1c_2) + a_1b_2 + a_1c_2 + b_1c_2\rangle_{\text{av}}. \tag{142}$$

This is suggested when the product z_1z_2 is multiplied out and terms of the type $a_1b_2 + b_1a_2$ are considered. Thus, if

$$\langle a_1b_2\rangle_{\text{av}} = \langle a(t)b(t + \tau)\rangle_{\text{av}} = f(\tau), \tag{143}$$

setting $t = t' - \tau$ and making use of stationarity leads to

$$\langle b_1a_2\rangle_{\text{av}} = \langle a(t + \tau)b(t)\rangle_{\text{av}} = \langle a(t')b(t' - \tau)\rangle_{\text{av}} = f(-\tau) \tag{144}$$

and hence to equation (141).

Similarly, replacing a_1, b_1, c_1 by a_1^*, b_1^*, c_1^* leads to writing the second ensemble average in equation (140) for $\langle x_1x_2\rangle_{\text{av}}$ as

$$\langle z_1^*z_2\rangle_{\text{av}} = B(\tau) + B^*(-\tau) \tag{145}$$

where

$$B(\tau) = \langle \tfrac{1}{2}(a_1^*a_2 + b_1^*b_2 + c_1^*c_2) + a_1^*b_2 + a_1^*c_2 + b_1^*c_2\rangle_{\text{av}}. \tag{146}$$

For terms of the type $a_1^*b_2 + b_1^*a_2$, the analogues of equation (143) and (144) are

$$\begin{aligned}\langle a_1^*b_2\rangle_{\text{av}} &= \langle a^*(t)b(t + \tau)\rangle_{\text{av}} = f(\tau), \\ \langle b_1^*a_2\rangle_{\text{av}} &= \langle a(t + \tau)b^*(t)\rangle_{\text{av}} = \langle a(t')b^*(t' - \tau)\rangle_{\text{av}} \\ &= \langle a^*(t')b(t' - \tau)\rangle_{\text{av}}^* = f^*(-\tau).\end{aligned} \tag{147}$$

Comparing equation (13) for $\Theta(t)$ with equation (138) for $z(t)$ suggests setting $a(t) = \Phi(t)$, $b(t) = -iK(t)$, and $c(t) = i2^{-1}K^2(t)$; this leads to equation (49) for $A(\tau)$, $B(\tau)$ given in Section IV.

Equation (140) for the autocovariance $\langle x_1x_2\rangle_{\text{av}}$ of $x(t)$ now takes the form

$$\langle x_1x_2\rangle_{\text{av}} = 2^{-1}\ \text{Re}\ [A(\tau) + A(-\tau) + B(\tau) + B^*(-\tau)], \tag{148}$$

and the power spectrum of $x(t)$ is

$$W_x(f) = \int_{-\infty}^{\infty} \exp(-i\omega\tau)\langle x_1 x_2\rangle_{\text{av}}\, d\tau, \qquad \omega = 2\pi f.$$

This may be written as

$$W_x(f) = 2^{-1}\,\text{Re}\,[P(f) + Q(f) + P^*(-f) + Q^*(-f)] \tag{149}$$

where

$$\begin{aligned} P(f) &= \int_{-\infty}^{\infty} \exp(-i\omega\tau)A(\tau)\, d\tau, \\ Q(f) &= \int_{-\infty}^{\infty} \exp(-i\omega\tau)B(\tau)\, d\tau. \end{aligned} \tag{150}$$

Equation (149) for $W_x(f)$ may be derived from

$$\begin{aligned} 2^{-1}\,\text{Re}\,[A(\tau) + A(-\tau)] &= 4^{-1}[A(\tau) + A^*(\tau) + A(-\tau) + A^*(-\tau)] \\ 2^{-1}\,\text{Re}\,[B(\tau) + B^*(-\tau)] &= 4^{-1}[B(\tau) + B^*(\tau) + B(-\tau) + B^*(-\tau)] \end{aligned} \tag{151}$$

and relations of the type

$$\begin{aligned} P^*(f) &= \int_{-\infty}^{\infty} \exp(i\omega\tau)A^*(\tau)\, d\tau = \int_{-\infty}^{\infty} \exp(-i\omega\tau)A^*(-\tau)\, d\tau \\ P(-f) &= \int_{-\infty}^{\infty} \exp(-i\omega\tau)A(-\tau)\, d\tau, \\ P^*(-f) &= \int_{-\infty}^{\infty} \exp(-i\omega\tau)A^*(\tau)\, d\tau. \end{aligned} \tag{152}$$

The power spectrum $W_y(f)$ of the imaginary part $y(t)$ of $z(t)$ may be computed in much the same way, starting with

$$\begin{aligned} \langle y_1 y_2\rangle_{\text{av}} &= \langle (z_1 - z_1^*)(z_2 - z_2^*)\rangle_{\text{av}}/(2i)^2 \\ &= 2^{-1}\,\text{Re}\,[-\langle z_1 z_2\rangle_{\text{av}} + \langle z_1^* z_2\rangle_{\text{av}}]. \end{aligned}$$

This differs from equation (140) for $\langle x_1 x_2\rangle_{\text{av}}$ only in the sign of $\langle z_1 z_2\rangle_{\text{av}}$. Therefore only the signs of $A(\tau)$ and $P(f)$ need be changed in the earlier work, and we get

$$W_y(f) = 2^{-1}\,\text{Re}\,[-P(f) + Q(f) - P^*(-f) + Q^*(-f)] \tag{153}$$

APPENDIX B

The Functions $S(\rho, \sigma)$, $S(\rho, \sigma, \nu)$, $\cdots$

The function

$$S(x_1 , x_2 , \cdots , x_n) = \int_{-\infty}^{\infty} du\, \gamma(u) H_u(x_1) H_u(x_2) \cdots H_u(x_n) \tag{154}$$

where

$$H_u(x) = \exp(-i2\pi xu) - \Gamma(x) \tag{155}$$

is a symmetrical function of the x's. It may be expressed as the sum of products of Γ's by replacing the H_u's by their definitions, multiplying out, and using the fact that $\Gamma(f)$ is the Fourier transform of $\gamma(t)$. This evaluation of the integral shows that $S(x_1, x_2, \cdots, x_n)$ is given symbolically by

$$S(x_1 , x_2 , \cdots , x_n) = \prod_{k=1}^{n} [y^{x_k} - \Gamma(x_k)] \tag{156}$$

where, after multiplying out, the various powers of y are replaced by Γ's according to the rule $y^z \rightarrow \Gamma(z)$.

For example,

$$\begin{aligned} S(\rho) &= 0, \\ S(\rho, \sigma) &= [y^\rho - \Gamma(\rho)][y^\sigma - \Gamma(\sigma)] \\ &= y^{\rho+\sigma} - y^\rho \Gamma(\sigma) - \Gamma(\rho) y^\sigma + \Gamma(\rho)\Gamma(\sigma) \\ &= \Gamma(\rho + \sigma) - \Gamma(\rho)\Gamma(\sigma) \end{aligned} \tag{157}$$

$$\begin{aligned} S(\rho, \sigma, \nu) = \Gamma(\rho + \sigma + \nu) &- \Gamma(\rho + \sigma)\Gamma(\nu) - \Gamma(\rho + \nu)\Gamma(\sigma) \\ &- \Gamma(\sigma + \nu)\Gamma(\rho) + 2\Gamma(\rho)\Gamma(\sigma)\Gamma(\nu). \end{aligned} \tag{158}$$

For four variables, writing (x) for $\Gamma(x)$,

$$\begin{aligned} S(\rho, \sigma, \nu, \mu) = (\rho + \sigma + \nu + \mu) &- (\rho + \sigma + \nu)(\mu) - (\rho + \sigma + \mu)(\nu) \\ &- (\rho + \nu + \mu)(\sigma) - (\sigma + \mu + \nu)(\rho) + (\rho + \sigma)(\nu)(\mu) \\ &+ (\rho + \nu)(\sigma)(\mu) + (\rho + \mu)(\sigma)(\nu) + (\sigma + \nu)(\rho)(\mu) \\ &+ (\sigma + \mu)(\rho)(\nu) + (\nu + \mu)(\rho)(\sigma) - 3(\rho)(\sigma)(\mu)(\nu). \end{aligned} \tag{159}$$

$S(x_1, x_2, \cdots, x_n)$ vanishes when one or more of its arguments are zero because $H_u(0)$ is zero. Of interest is the form taken by $S(x_1, \cdots, x_n)$ when the x's are small and $\Gamma(f)$ may be expanded as a power series in f. Let the power series be

$$\Gamma(f) = \sum_{n=0}^{\infty} \frac{f^n}{n!} \alpha_n , \qquad \alpha_0 = 1. \tag{160}$$

Since

$$\begin{aligned} \Gamma(f) &= \int_{-\infty}^{\infty} \gamma(u) \exp(\xi f)\, du, \qquad \xi = -i2\pi u \\ &= \sum_{n=0}^{\infty} \frac{f^n}{n!} \int_{-\infty}^{\infty} \gamma(u) \xi^n \, du \end{aligned}$$

it follows that

$$\alpha_n = \int_{-\infty}^{\infty} \gamma(u) \xi^n \, du. \tag{161}$$

When x is small, equation (155) for $H_u(x)$ gives

$$\begin{aligned} H_u(x) &= \exp(\xi x) - \Gamma(x) \\ &= x[(\xi - \alpha_1) + 2^{-1} x(\xi^2 - \alpha_2)] + O(x^3). \end{aligned}$$

Then

$$\prod_{k=1}^{n} H_u(x_n) = (x_1 x_2 \cdots x_n)\Big[(\xi - \alpha_1)^n + 2^{-1}(\xi - \alpha_1)^{n-1}(\xi^2 - \alpha_2) \sum_{k=1}^{n} x_k \Big] + O(x^{n+2})$$

and substitution in the integral [equation (154)] defining $S(x_1, \cdots, x_n)$ leads to

$$\begin{aligned} &S(x_1, \cdots, x_n) \\ &= (x_1 x_2 \cdots x_n) \int_{-\infty}^{\infty} du\, \gamma(u) \Big[\sum_{\ell=0}^{n} \binom{n}{\ell} \xi^{\ell} (-\alpha_1)^{n-\ell} \\ &\quad + 2^{-1} \sum_{1}^{n} x_k \sum_{\ell=0}^{n-1} \binom{n-1}{\ell} (\xi^{\ell+2} - \xi^{\ell}\alpha_2)(-\alpha_1)^{n-\ell-1} \Big] + O(x^{n+2}) \\ &= (x_1 x_2 \cdots x_n) \Big[\sum_{\ell=0}^{n} \binom{n}{\ell} \alpha_{\ell} (-\alpha_1)^{n-\ell} \\ &\quad + 2^{-1} \sum_{1}^{n} x_k \sum_{\ell=1}^{n-1} \binom{n-1}{\ell} (\alpha_{\ell+2} - \alpha_{\ell}\alpha_2)(-\alpha_1)^{n-\ell-1} \Big] + O(x^{n+2}). \end{aligned} \tag{162}$$

This is the approximation needed to examine the form taken by $W_\theta(f)$

when the bandwidth of $\varphi(t)$ becomes small, as it does in Section XI.

When $\Gamma(f)$ is such that $|\Gamma(f) - 1| < \epsilon \ll 1$ for all values of f, it may be shown that the symbolic form of equation (156) for $S(x_1, \cdots, x_n)$ becomes

$$S(x_1, \cdots, x_n) = \prod_{k=1}^{n} (y^{x_k} - 1) + O(\epsilon^2) \tag{163}$$

and the Γ's appear only linearly. Furthermore, $S(x_1, \cdots, x_n)$ is $O(\epsilon)$. For example

$$\begin{aligned} S(\rho, \sigma) &= y^{\rho+\sigma} - y^{\rho} - y^{\sigma} + 1 + O(\epsilon^2) \\ &= \Gamma(\rho + \sigma) - \Gamma(\rho) - \Gamma(\sigma) + 1 + O(\epsilon^2). \end{aligned}$$

This result is of interest in connection with the first order approximation discussed in Appendix D.

To establish equation (163) let

$$z_k = y^{x_k} - 1, \qquad \epsilon_k = 1 - \Gamma(x_k)$$

so that equation (156) for S becomes

$$\begin{aligned} S(x_1, \cdots, x_n) &= \prod_{k=1}^{n} (z_k + \epsilon_k) \\ &= \prod_{k=1}^{n} z_k + \prod_{k=1}^{n} \epsilon_k \prod_{\ell=1}^{n}{}' z_\ell + O(\epsilon^2). \end{aligned} \tag{164}$$

Here the factor z_k is omitted from $\prod'$. When the product

$$z_1 z_2 \cdots z_m = \prod_{k=1}^{m} (y^{x_k} - 1)$$

is multiplied out and the y^u's are replaced by $\Gamma(u)$'s, the result is the sum of $2^m \Gamma$'s $[(1 = \Gamma(0)]$. Half of the Γ's will have plus signs and the other half will have minus signs. Adding $+1$ for each $-\Gamma$ and -1 for each $+\Gamma$ shows that the entire sum is $O(\epsilon)$. Hence $z_1 z_2 \cdots z_m$ is $O(\epsilon)$, and when this is used to show that the $\prod'$ in equation (164) is $O(\epsilon)$, the result stated in equation (163) follows.

APPENDIX C

The Functions S_n and S_{nm}

The functions $S_n(x_1, x_2, \cdots, x_n)$ and $S_{nm}(x_1, \cdots, x_n; y_1, \cdots, y_m)$ are defined by the integrals, for $n \geqq 1$ and $m \geqq 1$,

$$S_n(x_1, \cdots, x_n) = \int_{-\infty}^{\infty} du\, \gamma(u) \exp[a(u)] \prod_{k=1}^{n} H_u(x_k), \tag{165}$$

$$\begin{aligned} &S_{nm}(x_1, \cdots, x_n; y_1, \cdots, y_m) \\ &\quad = \int_{-\infty}^{\infty} dv \int_{-\infty}^{\infty} dw\, \gamma(v)\gamma(w) \exp[a(v) + a(w) + 2c(w, v, o)] \\ &\qquad\qquad \cdot \left[\prod_{k=1}^{n} H_v(x_k)\right]\left[\prod_{l=1}^{m} H_w(y_l)\right] \end{aligned}$$

where $H_u(x)$, $a(u)$, $c(w, v, o)$ are given by

$$\begin{aligned} H_u(x) &= \exp(-i2\pi xu) - \Gamma(x) \\ a(u) &= -\frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f) H_u(-f) H_u(f) \\ c(w, v, o) &= -\frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f) H_w(-f) H_v(f) = c(v, w, o) \end{aligned} \tag{166}$$

[see equation (66)]. For $n = 0$, S_0 is defined as

$$S_0 = \int_{-\infty}^{\infty} du\, \gamma(u) \exp[a(u)] \tag{167}$$

and for the double subscripts,

$$\begin{aligned} &S_{n0}(x_1, \cdots, x_n;) \\ &\quad = \int_{-\infty}^{\infty} dv \int_{-\infty}^{\infty} dw\, \gamma(v)\gamma(w) \exp[a(v) + a(w) + 2c(w, v, o)] \prod_{k=1}^{n} H_v(x_k) \\ &S_{n0}(x_1, \cdots, x_n;) = S_{0n}(; x_1, \cdots, x_n) \\ &S_{00}(;) = \int_{-\infty}^{\infty} dv \int_{-\infty}^{\infty} dw\, \gamma(v)\gamma(w) \exp[a(v) + a(w) + 2c(w, v, o)]. \end{aligned} \tag{168}$$

The functions S_n and S_{nm} depend upon both $W_\varphi(f)$ and $\Gamma(f)$ [through $H_u(f)$]. This is in contrast with the function $S(x_1, \cdots, x_n)$, defined in Appendix B, which depends only on $\Gamma(f)$ and is independent of $W_\varphi(f)$.

The function S_n may be expressed as the sum of multiple integrals involving the functions S. Expanding $\exp[a(u)]$ in powers of $a(u)$ and replacing each $a(u)$ by its integral [equation (166)] with ρ in place of the variable of integration f leads to

$$S_0 = 1 + \sum_{j=1}^{\infty} \frac{(-\frac{1}{2})^j}{j!} \int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_j$$

$$\cdot\left[\prod_{k=1}^{j} W_{\varphi}(\rho_k)\right]S(\rho_1, \cdots, \rho_j, -\rho_1, \cdots, -\rho_j)$$

$$S_n(x_1, x_2, \cdots, x_n)$$
$$= S(x_1, x_2, \cdots, x_n) + \sum_{j=1}^{\infty} \frac{(-\frac{1}{2})^j}{j!} \int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_j$$
$$\cdot\left[\prod_{k=1}^{j} W_{\varphi}(\rho_k)\right]S(\rho_1, \cdots, \rho_j, -\rho_1, \cdots, -\rho_j, x_1, \cdots, x_n) \tag{169}$$

where $n \geqq 1$.

Similarly, expanding exp $[2c(w, v, o)]$ in the integrand of the integrals defining the S_{nm} functions leads to

$$S_{00}(;) = S_0^2 + \sum_{j=1}^{\infty} \frac{(-)^j}{j!} \int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_j$$
$$\cdot\left[\prod_{k=1}^{j} W_{\varphi}(\rho_k)\right]S_j(\rho_1, \cdots, \rho_j)S_j(-\rho_1, \cdots, -\rho_j)$$

$$S_{n0}(x_1, \cdots, x_n;)$$
$$= S_0S_n(x_1, \cdots, x_n) + \sum_{j=1}^{\infty} \frac{(-)^j}{j!} \int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_j \left[\prod_{k=1}^{j} W_{\varphi}(\rho_k)\right]$$
$$\cdot S_{j+n}(\rho_1, \cdots, \rho_j, x_1, \cdots, x_n)S_j(-\rho_1, \cdots, -\rho_j)$$

$$S_{nm}(x_1, \cdots, x_n; y_1, \cdots, y_m) = S_n(x_1, \cdots, x_n)S_m(y_1, \cdots, y_m)$$
$$+ \sum_{j=1}^{\infty} \frac{(-)^j}{j!} \int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_j \left[\prod_{k=1}^{j} W_{\varphi}(\rho_k)\right]$$
$$\cdot S_{j+n}(\rho_1, \cdots, \rho_j, x_1, \cdots, x_n)S_{j+m}(-\rho_1, \cdots, -\rho_j, y_1, \cdots, y_m) \tag{170}$$

when $n \geqq 1$ and $m \geqq 1$.

In order to obtain an inequality for $S_n(x_1, \cdots, x_n)$ assume that (*i*) the termwise integration in the derivation of the series in equation (169) is legitimate, and (*ii*) an $M > 1$ exists such that for all real values of f

$$M > | \Gamma(f) | = | G(f + f_0)/G(f_0) |.$$

Since $S(x_1, \cdots, x_n)$ may be expressed as the sum of 2^n terms, each of which is a product of not more than n of the Γ's,

$$| S(x_1, \cdots, x_n) | < 2^nM^n.$$

Furthermore, since the integral of $W_\varphi(f)$ from $-\infty$ to $+\infty$ is equal to $\langle\varphi^2(t)\rangle_{\text{av}}$, the terms of the series in equation (169) for $S_n(x_1, \cdots, x_n)$ are dominated by the terms in

$$\sum_{j=0}^{\infty} \frac{(\frac{1}{2})^j}{j!} \langle\varphi^2\rangle_{\text{av}}^j (2M)^{2j+n} = (2M)^n \exp [2M^2 \langle\varphi^2\rangle_{\text{av}}].$$

Therefore the series in equation (169) converges and

$$| S_n(x_1, \cdots, x_n) | < (2M)^n \exp [2M^2\langle\varphi^2\rangle].$$

This inequality may be used to show that the series in equation (170) for S_{nm} converges and that

$$| S_{nm}(x_1, \cdots, x_n ; y_1, \cdots, y_m) | < (2M)^{n+m} \exp [8M^2\langle\varphi^2\rangle_{\text{av}}].$$

The leading terms in the series required to handle the second and third order modulation are

$$\begin{aligned} S_0 = 1 &- \frac{1}{2}\int_{-\infty}^{\infty} d\rho\, W_\varphi(\rho) S(\rho, -\rho) \\ &+ \frac{1}{8}\int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma\, W_\varphi(\rho) W_\varphi(\sigma) S(\rho, \sigma, -\rho, -\sigma) + O(\varphi^6), \end{aligned} \tag{171}$$

$$\begin{aligned} S_1(f) = 0 &- \frac{1}{2}\int_{-\infty}^{\infty} d\rho\, W_\varphi(\rho) S(\rho, -\rho, f) \\ &+ \frac{1}{8}\int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma\, W_\varphi(\rho) W_\varphi(\sigma) S(\rho, \sigma, -\rho, -\sigma, f) + O(\varphi^6), \end{aligned} \tag{172}$$

$$S_2(\rho, \nu) = S(\rho, \nu) - \frac{1}{2}\int_{-\infty}^{\infty} d\sigma\, W_\varphi(\sigma) S(\sigma, -\sigma, \rho, \nu) + O(\varphi^4), \tag{173}$$

$$S_3(\rho, \sigma, \nu) = S(\rho, \sigma, \nu) + O(\varphi^2), \tag{174}$$

$$S_{00}(;) = S_0^2 + \frac{1}{2}\int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma\, W_\varphi(\rho) W_\varphi(\sigma) S(\rho, \sigma) S(-\rho, -\sigma) + O(\varphi^6), \tag{175}$$

$$\begin{aligned} S_{10}(f;) = S_1(f) &+ \int_{-\infty}^{\infty} d\rho \int_{-\infty}^{\infty} d\sigma\, W_\varphi(\rho) W_\varphi(\sigma) [4^{-1} S(\rho, -\rho, f) S(\sigma, -\sigma) \\ &+ 2^{-1} S(\rho, f) S(\sigma, -\sigma, -\rho) + 2^{-1} S(\rho, \sigma, f) S(-\rho, -\sigma)] + O(\varphi^6), \end{aligned} \tag{176}$$

$$S_{20}(\rho, \nu;) = S_2(\rho, \nu) - \frac{1}{2}\int_{-\infty}^{\infty} d\sigma\, W_\varphi(\sigma) S(\sigma, -\sigma) S(\rho, \nu) + O(\varphi^4).$$

$$S_{30}(\rho, \sigma, \nu;) = S_3(\rho, \sigma, \nu) + O(\varphi^2),$$

$$S_{11}(\rho; \nu) = -\int_{-\infty}^{\infty} d\sigma \, W_\varphi(\sigma) S(\sigma, \rho) S(-\sigma, \nu) + O(\varphi^4),$$

$$S_{21}(\rho, \sigma; \nu) = O(\varphi^2).$$

In obtaining the leading terms in $S_{00}(;)$ and $S_{10}(f;)$ the leading terms in S_0, $S_1(f)$ and $S_2(\rho, f)$ were used. $S_1(f)$ is $O(\varphi^2)$ in contrast with S_0, S_2, S_3, $\cdots$ which are $O(1)$.

APPENDIX D

Derivation of Earlier First Order Approximation by Present Procedure

The first order approximations which are given in Section VII are somewhat more complicated, as well as more accurate, than the ones which have appeared in the literature.[3, 4, 9–12] Here the relation between the earlier and present work will be brought out by applying the procedure of Section VII to obtain a first order approximation, which is the same as the given in Ref. 10 [the $\theta(t)$ of Ref. 10 is $\theta(t) - \varphi(t)$ in the present notation].

The derivation starts from the initial equations (5) and (7) for the filter output $s_0(t)$,

$$\begin{aligned} s_0(t) &= \{\exp[-\alpha_0 - i\beta_0 + i\Theta(t)]\} \exp(i\omega_0 t) \\ \Theta(t) &= -i \ln \left\{ \int_{-\infty}^{\infty} du \, \gamma(u) \exp[i\varphi(t-u)] \right\}. \end{aligned} \tag{177}$$

The difference between the output phase angle $\theta(t) = \operatorname{Re} \Theta(t)$ and the input phase angle $\varphi(t)$ is assumed to be small, and the filter delay is usually taken to be zero at the carrier frequency, that is, $\operatorname{Im} [d\Gamma(f)/df]$ is zero at $f = 0$.

Adding and subtracting $i\varphi(t)$ [instead of the linear portion $\Phi(t)$ of the output] in the exponent appearing in equation (177) for $\Theta(t)$ gives

$$\begin{aligned} \Theta(t) &= \varphi(t) - i \ln [1 + k(t)], \\ k(t) &= \int_{-\infty}^{\infty} du \, \gamma(u) \{\exp[i\varphi(t-u) - i\varphi(t)] - 1\}. \end{aligned} \tag{178}$$

The first order approximations to the complex phase angle $\Theta(t)$ and the output phase angle $\theta(t)$ are now

$$\Theta(t) \approx \varphi(t) - ik(t)$$

and

$$\theta(t) \approx \varphi(t) + \operatorname{Re}\,[-ik(t)], \tag{179}$$

respectively.

The analysis of the earlier sections goes through much as before with $k(t)$ in place of $K(t)$, and

$$h_u(f) = \exp\,(-i2\pi fu) - 1 \tag{180}$$

in place of

$$H_u(f) = \exp\,(-i2\pi fu) - \Gamma(f).$$

An illustration of how $h_u(f)$ enters the analysis is furnished by the computation of $\langle \exp\,[i\varphi(t-u) - i\varphi(t)]\rangle_{\text{av}}$. As suggested by equation (56) for $\varphi(t-u) - \Phi(t)$, let

$$\begin{gathered}
L\varphi(t) - \varphi(t-u) - \varphi(t) \\
\ell(f) = \exp\,(-i2\pi fu) - 1 = h_u(f) \\
-\frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f)\ell(f)\ell(-f) = -\frac{1}{2}\int_{-\infty}^{\infty} df\, W_\varphi(f)h_u(f)h_u(-f) = a'(u)
\end{gathered} \tag{181}$$

$$\langle \exp\,[i\varphi(t-u) - i\varphi(t)]\rangle_{\text{av}} = \exp\,[a'(u)].$$

Of most interest in practice is the power spectrum $W_\xi(f)$ of $\xi(t)$,

$$\begin{aligned}
\xi(t) &= \operatorname{Re}\,[-ik(t)] \\
\theta(t) &\approx \varphi(t) + \xi(t)
\end{aligned} \tag{182}$$

where $\xi(t)$ is an approximation to the distortion. The power spectrum $W_\xi(f)$ is the Fourier transform of the covariance $\langle \xi_1\xi_2\rangle$ where, as before, subscripts 1, 2 refer to times t, $t+\tau$, respectively. By putting $\xi(t)$, $-ik(t)$ for $\theta(t)$, $\Theta(t)$ in equation (47), or directly,

$$\langle \xi_1\xi_2\rangle_{\text{av}} = 2^{-1}\operatorname{Re}\,[-\langle k_1k_2\rangle_{\text{av}} + \langle k_1^*k_2\rangle_{\text{av}}]. \tag{183}$$

It may be shown that

$$\langle k_1\rangle_{\text{av}} = \int_{-\infty}^{\infty} du\, \gamma(u)\exp\,[a'(u)]$$

$$\langle k_1k_2\rangle_{\text{av}} = \langle k_1\rangle_{\text{av}}^2 + \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\, \gamma(u)\gamma(v)\exp\,[a'(u) + a'(v)] \cdot \{\exp\,[2c'(u, v, \tau)] - 1\}$$

$$\langle k_1^* k_2 \rangle_{\text{av}} = |\langle k_1 \rangle_{\text{av}}|^2 + \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\, \gamma^*(u)\gamma(v) \exp [a'(u) + a'(v)] \cdot \{\exp [-2c'(u, v, \tau)] - 1\} \tag{184}$$

where

$$\begin{aligned} c'(u, v, \tau) &= -\frac{1}{2} \int_{-\infty}^{\infty} df\, W_\varphi(f) h_u(-f) h_v(f) \exp (i2\pi f \tau) \\ a'(u) &= c'(u, u, o) \\ c'(u, v, -\tau) &= c'(v, u, \tau). \end{aligned} \tag{185}$$

Since $h(-f)$ is equal to $h^*(f)$, $a'(u)$ and $c'(u, v, \tau)$ are real. Furthermore,

$$\begin{aligned} c'(u, v, \tau) &= -\tfrac{1}{2}[R_\varphi(\tau + u - v) - R_\varphi(\tau - v) - R_\varphi(\tau + u) + R_\varphi(\tau)] \\ a'(u) &= -R_\varphi(o) + R_\varphi(u) \end{aligned} \tag{186}$$

where $R_\varphi(\tau)$ is the covariance $\langle \varphi(t)\varphi(t + \tau)\rangle_{\text{av}}$ of $\varphi(t)$.

The expression for $\langle \xi_1 \xi_2 \rangle_{\text{av}}$ obtained by combining equations (183) and (184) is similar to equation (8) of Ref. 10.

The power spectrum of $\xi(t)$ may be written as

$$W_\xi(f) = \xi_{dc}^2\, \delta(f) + 2^{-1} \operatorname{Re} [P(f) + Q(f) + P^*(-f) + Q^*(-f)] \tag{187}$$

in which ξ_{dc} is equal to $\operatorname{Im} \langle k_1 \rangle_{\text{av}}$ and

$$\begin{aligned} P(f) &= \int_{-\infty}^{\infty} d\tau \exp [-i2\pi f \tau][-\tfrac{1}{2}(\langle k_1 k_2 \rangle_{\text{av}} - \langle k_1 \rangle_{\text{av}}^2)], \\ Q(f) &= \int_{-\infty}^{\infty} d\tau \exp (-i2\pi f \tau)[\tfrac{1}{2}(\langle k_1^* k_2 \rangle_{\text{av}} - |\langle k_1 \rangle_{\text{av}}|^2)]. \end{aligned} \tag{188}$$

Addition gives

$$\begin{aligned} P(f) + Q(f) = \int_{-\infty}^{\infty} d\tau \exp (-i2\pi f \tau) \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv \exp [a'(u) + a'(v)] \\ \cdot (-\tfrac{1}{2}\gamma(u)\gamma(v)\{\exp [2c'(u, v, \tau)] - 1\} \\ + \tfrac{1}{2}\gamma^*(u)\gamma(v)\{\exp [-2c'(u, v, \tau)] - 1\}) \end{aligned} \tag{189}$$

which, when used in equation (187), leads to an expression for $W_\xi(f)$ which is similar to the main result given in Ref. 10 [equation (16) of Ref. 10].

The relation $c'(u, v, -\tau) = c'(v, u, \tau)$ may be used to show that $Q(f)$ is real and $P(f)$ is even. When $\gamma(u)$ is real, $\Gamma(-f) = \Gamma^*(f)$, and the expression for $W_\xi(f)$ may be simplified.

Expanding $\exp [\pm 2c'(u, v, \tau)]$ in powers of $c'(u, v, \tau)$ leads to

$$\begin{aligned} W_\xi(f) = \xi_{dc}^2\,\delta(f) + \frac{1}{4}\sum_{n=1}^{\infty}\frac{1}{n!}\int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_n\,\delta(f - \rho_1 - \cdots - \rho_n) \\ \cdot\left[\prod_{k=1}^{n} W_\varphi(\rho_k)\right] \mid S_n'(\rho_1, \cdots, \rho_n) - (-)^n S_n'^*(-\rho_1, \cdots, -\rho_n) \mid^2 \end{aligned} \tag{190}$$

where, as in Appendix C for the unprimed S's,

$$\begin{aligned} &S_n'(x_1, \cdots, x_n) \\ &\quad = \int_{-\infty}^{\infty} du\,\gamma(u)\exp[a'(u)]\prod_{k=1}^{n}[\exp(-i2\pi u x_k) - 1] \\ &\quad = S'(x_1, \cdots, x_n) + \sum_{j=1}^{\infty}\frac{(-\frac{1}{2})^j}{j!}\int_{-\infty}^{\infty} d\rho_1 \cdots \int_{-\infty}^{\infty} d\rho_j\left[\prod_{k=1}^{j} W_\varphi(\rho_k)\right] \\ &\qquad\qquad \cdot S'(\rho_1, \cdots, \rho_j, -\rho_1, \cdots, -\rho_j, x_1, \cdots, x_n). \end{aligned}$$

The series in equation (190) is analogous to the series in equation (90) for the more accurate first order approximation based on $\varphi(t - u) - \Phi(t)$. The function $S'(x_1, \cdots, x_n)$ is the analogue of $S(x_1, \cdots, x_n)$ discussed in Appendix B and is defined by

$$\begin{aligned} S'(x_1, \cdots, x_n) &= \int_{-\infty}^{\infty} du\,\gamma(u)\prod_{k=1}^{n} h_u(x_k) \\ &= \prod_{k=1}^{n}(y^{x_k} - 1). \end{aligned}$$

The second equation is symbolic in that y^z is to be replaced by $\Gamma(z)$ after expansion of the product. It is shown in Appendix B that when $\mid \Gamma(f) - 1 \mid < \epsilon \ll 1$ for all real values of f,

$$S(x_1, \cdots, x_n) = S'(x_1, \cdots, x_n) + O(\epsilon^2).$$

APPENDIX E

Power Spectrum of $a\theta(t) + b \ln R(t)$

Section III states that an expression for the power spectrum $W_x(f)$ of $x(t) = a\theta(t) + b \ln R(t)$ may be obtained from equation (15) for $W_\theta(f)$ by replacing $U(f)$, $T(\rho, f-\rho)$, and $S(\rho, \sigma, f-\rho-\sigma)$ by $(a+ib)U(f)$, $(a + ib)T(\rho, f - \rho)$, and $(a + ib)S(\rho, \sigma, f - \rho - \sigma)$, respectively. Here the steps leading to this result are outlined.

From equation (6) for the complex phase angle $\Theta(t)$ we have

$$i\Theta(t) = \ln R(t) + i\Theta(t),$$

and it follows that, for arbitrary real values of a and b,

$$a\theta + b \ln R(t) = \operatorname{Re}\,[(a + ib)\Theta(t)].$$

Consequently $W_x(f)$ may be obtained by replacing $\Theta(t)$ by $(a + ib)\Theta(t)$ in the analysis which led to equation (15) for $W_\theta(f)$.

The functions $A(\tau)$ and $B(\tau)$ appearing in equation (48) are replaced by $(a + ib)^2A(\tau)$ and $| a + ib |^2 B(\tau)$, and their respective Fourier transforms $P(f)$ and $Q(f)$ are replaced by $(a+ib)^2P(f)$ and $| a+ib |^2 Q(f)$. Each factor in $(a + ib)^2 = (a + ib)(a + ib)$ can be associated with factors in $P(f)$, and each factor in $(a + ib)(a + ib)^*$ with factors in $Q(f)$, in such a way that $U(f)$ becomes multiplied by $(a + ib)$ and $U^*(-f)$ by $(a + ib)^*$, and so on. This may be verified by repeating the analysis of Sections VIII and IX with the modified expressions.

APPENDIX F

Results Obtained from the Series for ln$[1 + K(t)]$ may be Asymptotic

For gaussian $\varphi(t)$ with average 0 and rms value σ, the following considerations suggest that results obtained from equations (10) and (12), namely

$$\begin{aligned}\Theta(t) &= \Phi(t) - i \ln [1 + K(t)] \\ &= \Phi(t) + i \sum_1^\infty n^{-1}[-K(t)]^n\end{aligned} \tag{191}$$

represent the first few terms of an asymptotic series when $\sigma \rightarrow 0$. Since $K(t)$ is difficult to handle, we replace it by $a[a^{i\varphi(t)} - 1]$ where a is somewhat like the integral of $\gamma(u)$ between $-\infty$ and ∞. The value of this integral is 1, and we regard a as being near 1. The series for $\ln [1 + K(t)]$ behaves somewhat like the series

$$\ln (1 + a\{\exp [i\varphi(t)] - 1\}) = -\sum_1^\infty \frac{(-a)^n}{n} \{\exp [i\varphi(t)] - 1\}^n \tag{192}$$

in which the mean square value of the modulus of the nth term is

$$\int_{-\infty}^{\infty} \frac{\exp [-\varphi^2/(2\sigma^2)]}{\sigma(2\pi)^{\frac{1}{2}}} n^{-2}[2a \sin \varphi/2]^{2n}\, d\varphi. \tag{193}$$

When $\sigma \ll 1$ and n is not too large, most of the contribution to the integral (193) arises from the region around $\varphi = 0$, and the integral is approximately

$$1 \cdot 3 \cdots (2n - 1)(a\sigma)^{2n}/n^2.$$

Consequently, the first few terms decrease rapidly when $\sigma \ll 1$. However, when n is very large most of the contribution arises from the regions around $\varphi = \pm\pi, \pm 3\pi, \cdots$, where $[\sin(\varphi/2)]^{2n}$ is a narrow pulse of height 1 and area $2(\pi/n)^{\frac{1}{2}}$. When $\sigma \ll 1$ only the regions around $\varphi = \pm\pi$ are important and the integral (193) is approximately

$$\sigma^{-1}2n^{3/2-5/2}(2a)^{2n}\exp\left[-\pi^2/(2\sigma^2)\right]$$

which tends to ∞ when a is near 1 and $n \to \infty$.

The fact that the rms values of the terms of the series in equation (192) decrease rapidly at first and then increase without limit suggests that results attained from the somewhat similar series in equation (191) may be asymptotic in nature as $\sigma \to 0$.

REFERENCES

1. Bedrosian, E., and Rice, S. O., "Distortion and Crosstalk of Linearly Filtered Angle-Modulated Signals," Proc. IEEE, *56,* No. 1 (January 1968), pp. 2–13.
2. Bedrosian, E., Transionospheric Propagation of FM Signals, RAND Memorandum RM-5379-NASA, August 1967.
3. Medhurst, R. G., "Explicit Form of FM Distortion Products with White-Noise Modulation," Proc. IEE, *107,* No. 11, part C (March 1960), pp. 120–126, and *107,* No. 12 (with J. H. Roberts) (September 1960), pp. 367–369.
4. Magnusson, R. I., "Intermodulation Noise in Linear FM Systems," Proc. IEE, *109,* No. 15, part C (March 1962), pp. 32–35.
5. Liou, M. L., "Noise in an FM System Due to an Imperfect Linear Transducer, B.S.T.J.," *45,* No. 9 (November 1966), pp. 1537–1561.
6. Ruthroff, C. L., "Exact Computation of FM Distortion in any Linear Network for Bandlimited Periodic Signals, B.S.T.J.," *47,* No. 6 (July-August 1968), pp. 1043–1063.
7. Enloe, L. H. and Ruthroff, C.L., "A Common Error in FM Distortion Theory," Proc. IEEE, *51,* No. 5 (May 1963), p. 846.
8. Gladwin, A. S., Medhurst, R. G., Enloe, L. H., and Ruthroff, C. L. "A Common Error in FM Distortion Theory, Proc." IEEE, *52,* No. 2 (February 1964), pp. 186–189.
9. Bennett, W. R., Curtis, H. E., and Rice, S. O., "Interchannel Interference in FM and PM Systems Under Noise Loading Conditions, B.S.T.J.," *34,* No. 3 (May 1955), pp. 601–636.
10. Rice, S. O, "Distortion in a Noise-Modulated FM Signal by Nonlinear Attenuation and Phase Shift," B.S.T.J., *36,* No. 4 (July 1967), pp. 879–890.
11. Bosse, G., "Die Verzerrung frequenzmodulierter Schwingungen beim Durchgang durch lineare Netzwerke," Frequenz 12, Sonderheft, (October 1958), pp. 6–13.
12. Medhurst, R. G., "Fundamental and Harmonic Distortion of Waves Freouency-Modulated with a Single Tone," Proc. IEE, *107,* No. 32, part B (March 1960), pp. 155–164. (Numerous references to earlier work are given.)
13. Hess, D. T., "Transmission of FM Signals through Linear Filters," Proc. Nat. Elec. Conf., *18* (1962), pp. 469–476.
14. Rowe, H. E., *Signals and Noise in Communication Systems,* Princeton, N. J.: D. Van Nostrand Co., Inc., 1965, p. 121.
15. Bedrosian, E. and Rice, S. O., "Reply to Comment by J. H. Roberts," scheduled to be published soon in the Proc. IEEE, Letters to the Editor.

Frequency Modulation of a Millimeter-wave IMPATT Diode Oscillator and Related Harmonic Generation Effects

By T. P. LEE and R. D. STANDLEY

(Manuscript received May 31, 1968)*

In this paper we report the performance of a continuous wave millimeter-wave IMPATT diode oscillator with a wide-band tunability. The diode is mounted in an iris wafer circuit; its oscillation frequency can be modulated either by a varactor diode or by direct modulation of the IMPATT diode bias current. The oscillator has been successfully used as a millimeter-wave frequency deviator in an experimental pulse code modulation millimeter-wave system.

We also report detailed measurements on subharmonic frequencies in IMPATT diode oscillators. Experimental results show that wide frequency tunability can be obtained with a circuit which provides an "idler" resonance at one-half the fundamental transit-time frequency. The results also show that by providing "idler" resonances at both the transit-time frequency and at one-half of the transit-time, the oscillation at $\frac{3}{2}$ the transit-time frequency is enhanced and yields a useful output power of 2 mW at 86 GHz.

I. INTRODUCTION

A continuous wave (CW) millimeter-wave silicon IMPATT diode oscillator was used satisfactorily as a local oscillator in an experimental pulse code modulation (PCM) millimeter-wave repeater system.[1, 2] That oscillator circuit used a radial-line resonant cavity whose resonant frequency was the primary factor determining the oscillation frequency. The oscillator was difficult to tune either mechanically or electronically.

While such characteristics are desirable for fixed frequency local oscillator applications, other applications which demand wideband

* The essential results of this paper were presented at the 26th annual conference on Electron Devicc Research, Boulder, Colorado, June 19–21, 1968.

performance and tunability are also of interest. For example, a millimeter-wave frequency deviator using an IMPATT diode would be advantageous over the L-band deviator and up-converter combination in the PCM repeater system. The IMPATT diode oscillator could deliver power at least one order of magnitude greater than the up-converter.

A previous paper showed that the circuit inductance of the radial-line cavity in shunt with the diode was much smaller than the diode inductance.[1] Therefore, the oscillation frequency could be tuned only a few hundred megahertz by varying the diode bias current.* In addition, the radial-line cavity was loosely coupled to the external waveguide circuit so that a tuning range of only 300 MHz was obtained when a varactor diode was used for frequency deviation. To improve the tunability, it is necessary to have (*i*) the IMPATT diode equivalent inductance dominate that contributed by other circuit elements, and (*ii*) the diode closely coupled to controllable external circuit parameters.

It is important to notice that the tunability can be further improved by providing more than one resonant circuit for the oscillator. The frequencies are harmonically related, and oscillation at each of these frequencies is reactively terminated except the one for the output. Take as an example the two-frequency case. The tunability of the oscillator near the transit-time frequency (say 60 GHz) can be improved if a lossless resonant circuit at half the output frequency (30 GHz) is incorporated in the oscillator circuit. The oscillation at 30 GHz is terminated by a cut-off waveguide and therefore is designated as the "idler," an analog to the idler in harmonic generators using varactor diodes. It is not necessary that the low frequency be used as the idler. Swan has shown that by providing an idler resonance at twice the transit-time frequency, improvements in both the power output and the tunability can be obtained.[3]

To achieve wide band tunability, IMPATT oscillators were designed using resonant-iris structures. It can be shown that the diode equivalent inductance is dominating in this circuit compared with that of the radial-line cavity structure. The loaded Q of the iris is about 10 which provides a wide bandwidth for oscillation. The oscillation frequency could be tuned over 9 GHz in the 50 to 60 GHz range by varying the bias current of the IMPATT diode, thus varying the diode

* The diode equivalent inductance is inversely proportional to the diode bias current.

equivalent inductance. It could also be tuned over 3 GHz by using a varactor diode closely coupled to the IMPATT diode.

Notice that by using the fundamental transit-time frequency as the idler, the power output at one-half the transit-time frequency is, in general, higher than the power output when the half frequency is used as the idler in our particular circuit. We refer to the frequency which is one-half the transit-time frequency as the subharmonic frequency throughout this paper.

In the sections which follow we describe the circuit structure of the oscillator, the performance as a tunable oscillator, and the results of frequency modulation of the oscillator in an experimental millimeter-wave PCM repeater system. Then measurements of various harmonic frequencies existing in the oscillator circuit, and the identification and effect of the subharmonic oscillation are described in detail.

II. DESCRIPTION OF THE OSCILLATOR

Figure 1a shows the resonant-iris structure. The iris is made in wafer form similar to the Sharpless wafer.[4] The range of the oscillation frequency and the Q of the resonant iris are determined by the size, thickness, and shape of the iris aperture. Refs. 5, 6, and 7 give details of iris characteristics. The wafers used were 0.100 inch thick,

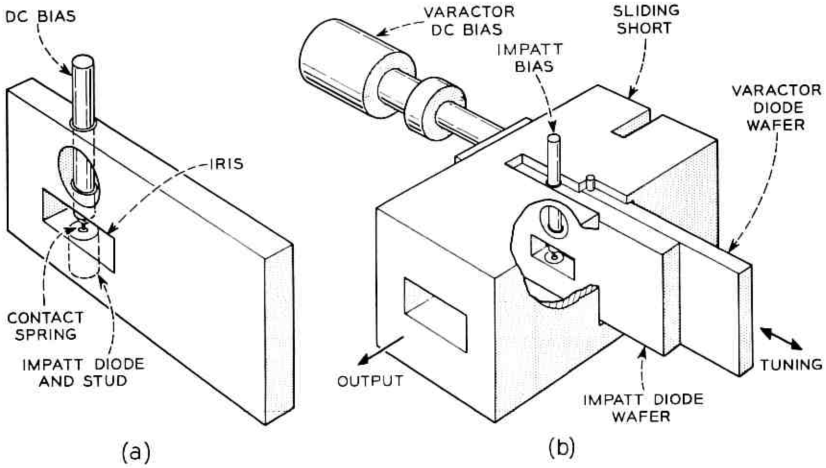


Fig. 1—Artist's view of the oscillator assembly. (a) IMPATT diode wafer. (b) Oscillator mount.

and had a rectangular aperture 0.010 or 0.030 inch high and 0.100 to 0.148 inch wide.

The IMPATT diodes had a 0.001 inch diameter mesa structure and nearly abrupt junctions as reported by Misawa.[8] Typical diode characteristics are given in Table I. The diodes were thermal-compression bonded onto gold-plated copper studs 0.063 inch in diameter. After bonding, the diodes were coated with a thin layer of silicone varnish which enchances the mechanical strength of the mesa structure. Electrical contact to the diode was made by a 0.032-inch nickel rod with a welded contact spring, as pictured in Fig. 1a. For the varactor-tuned oscillator, a varactor diode in a Sharpless wafer was mounted adjacent to the iris wafer as shown in Fig. 1b. The relative coupling between the IMPATT diode and the varactor diode (thus the tuning characteristic) was adjusted by varying the relative position of the two diodes in the plane of the cross section of the waveguide.[9] The complete assembly in a RG-98/U waveguide is shown in Fig. 1b.

Table I — Typical Diode Characteristics*

Epitaxial layer thickness†	2.1 μ
Epitaxial layer doping density‡	6×10^{16} cm^{-3}
Junction depth†	1.0 μ
Space charge layer width‡	0.75 μ
Breakdown voltage (at 1 μA)	19V
Capacitance at breakdown	0.19 pF
Junction area A	1.5×10^{-5} cm^2

* All diodes were from the LO 1114 series.
† Measured by staining and interference fringe.
‡ Obtained from voltage dependence of capacitance.

III. OSCILLATOR PERFORMANCE

Figure 2 shows the typical performance of an oscillator without the varactor. (The varactor wafer is replaced by a blank wafer.) The iris used had an aperture of 0.130 by 0.030 inch. The CW ouput power (above 1 mW) and the frequency are plotted as a function of the IMPATT diode current. Oscillation began at a diode current of about 50 mA compared with 100 mA for the cavity structure.[1] The output power was optimized at each current level by adjusting an E-H tuner in front of the diode and a sliding short in back of each

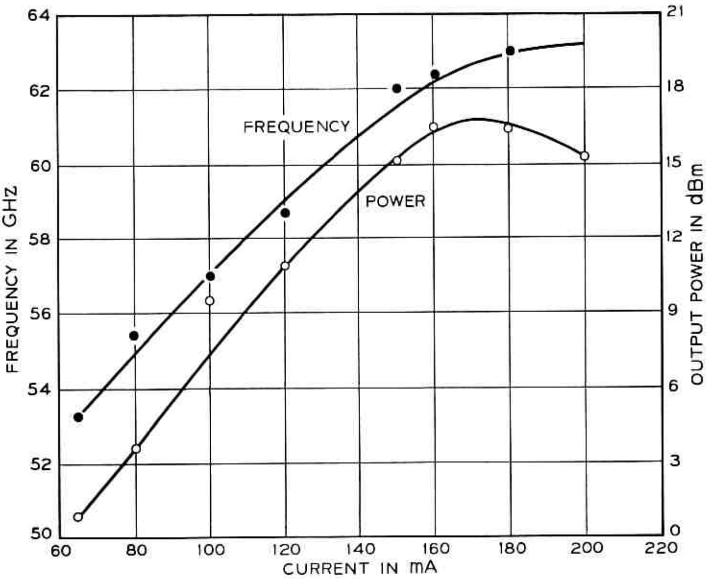


Fig. 2 — Maximum output power and frequency as a function of diode current.

diode. Note that the tunable band (about 10 GHz) is considerably wider than that previously reported for a radial line cavity structure. By varying the bias current with fixed mechanical tuning, the frequency could be tuned over 4 GHz with a 3 dB variation in the output power, and over 9 GHz with a minimum output power of 1 mW.

Table II summarizes the results of various circuits with different iris apertures. Although there are not enough data to draw a meaningful conclusion, it seems that the oscillation frequency was relatively independent of the iris width compared with iris height, which implies that the inductance of the center post (with contacting spring) partially controls the oscillation frequency. This statement is based on the assumption that diodes used in the test are of uniform characteristics since they were made from a single slice and batch processed. The dc characteristics of the diodes showed less than 5 percent variation in capacitance and in breakdown voltage.

IV. FREQUENCY MODULATION

As mentioned in Section I, one goal in this work was to design an IMPATT diode oscillator circuit which could be frequency modulated

TABLE II—RESULTS OF CIRCUITS WITH VARIOUS DIODES AND IRIS APERTURES

Diode No.	Width (in)	Height (in)	C_B‡ (pF)	Tunable¶ bandwidth (GHz)	P_{max} (dBm)
1	148	0.030	0.19	50 – 60	15
2	130	0.030	0.19	53 – 63	16.5
3	120	0.030	0.19	52 – 59	11.5
4	110	0.030	0.18	56 – 63.2	9.4
5	100	0.030	0.19	53 – 61	10.7
6	130	0.010	0.16	64 – 70	11.5
7	148	0.010	0.18	58 – 65	11.1

‡ C_B is the diode capacitance measured at breakdown ($V_B = -19$ volts).

¶ The lower limit on frequency is arbitrarily chosen as the frequency for which the power is 1 mW. The upper frequency limit occurs at a bias current of 160 mA which is well below the burn-out level for the LO 1114 series diodes.

and used as a frequency deviator in a PCM millimeter-wave repeater system.[2] Two approaches were taken. The first used a varactor diode to tune the circuit susceptance outside the diode; the second used the fact that the diode inductance (from avalanching) varies inversely with the bias current.

Both methods have advantages and disadvantages. The power output of the varactor-tuned oscillator remains almost constant over the frequency band so that amplitude modulation is negligible. However, to achieve ultimate performance the circuit is much more complex. The bias-current tuned oscillator has simpler circuitry, but inevitably its power output varies with bias current, which results in AM distortion. However, the AM distortion can be overcome by proper tuning of the circuit if the modulation index is small. (See Section 4.2)

4.1 *The Varactor-Tuned FM Deviator*

The varactor tuned oscillator is shown in Fig. 1b. The varactor diodes used were planar diffused GaAs diodes with a honeycomb structure.[10] The zero-bias capacitance was 0.04 – 0.05 pF and the breakdown voltage was 20 volts. The capacitance varied with voltage approximately as $C = C_0\ (1 + V/\varphi)^{-0.4}$, where C_0 is the junction capacitance at zero bias, V is the bias voltage, and φ is "built-in" voltage which is approximately 1 volt for GaAs. The diode was then mounted in a Sharpless wafer which in turn was inserted into the oscillator mount. The varactor diode was about 0.080 inch behind the IMPATT diode. The coupling between the varactor diode and the

IMPATT diode was tuned by sliding the varactor wafer relative to the IMPATT wafer. Additional tuning was provided by a sliding short behind the varactor diode and an E-H tuner in front of the IMPATT diode.

The self-resonant frequency of the varactor diode, as measured by the transmission resonance technique, was used to correlate with the tuning sensitivity of the oscillator. The self-resonant frequency can be varied by changing the length of the contacting wire in the varactor wafer.[11] It was found experimentally that the best tuning sensitivity was obtained when the varactor diode was self-resonant near the "idler" frequency.

A typical frequency tuning characteristic of the oscillator is shown in Fig. 3 where the frequency is plotted as a function of the dc bias voltage (reverse biased) on the varactor. A tunable band of 2.5 GHz was obtained. Notice that the linear region was 1.5 GHz with a sensitivity of 1 GHz per volt. Also notice that the power output varied ±1 dB over the tunable band.

Frequency modulation with sinusoidal drive on the varactor was achieved at both the 160 MHz and the 500 MHz modulation fre-

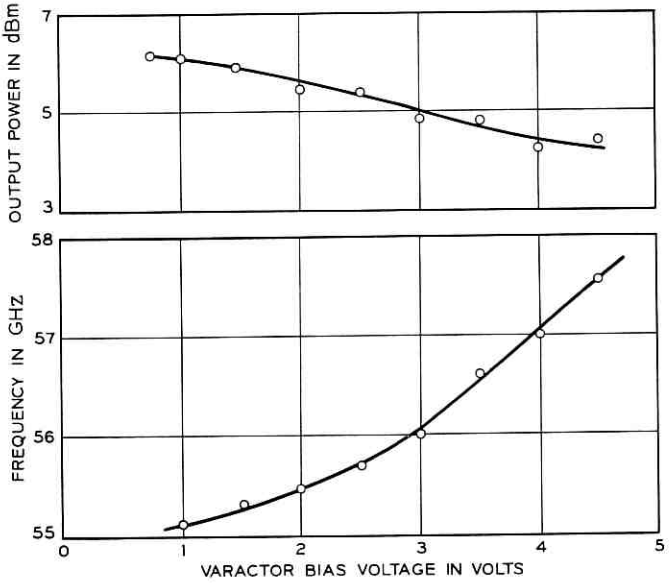


Fig. 3 — Performance of the varactor-tuned oscillator.

quency. Figure 4 shows a series of the FM spectrum at 160 MHz modulation as the modulating power is increased. For the particular diode tested, the minimum modulating power required at 160 MHz for complete carrier suppression was about 4 mW (to a 50 ohm load).

4.2 *The Bias-Current Tuned FM Deviator*

The second method of frequency modulation was achieved by direct modulation of the IMPATT diode bias current (the varactor diode wafer was replaced by a blank wafer). The modulating signal was directly applied to the bias through a coupling capacitor. The dc tuning sensitivity was between 100 to 200 MHz per milliampere at optimum tuning conditions; this resulted in a decrease of 1 dB in output power compared with the maximum power obtained without frequency deviation. For a sinusoidal drive at 160 MHz, complete carrier suppression was achieved at 0.25 mW drive power. Figure 5 shows the spectrum of the modulated signal at various drive levels. When it is tuned properly, no appreciable AM distortion was seen at low modulation index.

4.3 *System Performance with the IMPATT Diode Oscillator-Deviator*

The methods of IMPATT diode oscillator frequency deviation were tested in a two-level PCM millimeter-wave repeater system.[2] The error rate of the system with the IMPATT diode oscillator used as the deviator was measured as a function of the signal-to-noise ratio. Figure 6 is a simplified block-diagram of the test circuit. The IMPATT diode oscillator-deviator is driven by a random-word generator either through the bias of the varactor for varactor-tuned deviation or through the bias of the IMPATT diode for current-tuned deviation.

The random word generator produces 160 megabit pulses per second with a random distribution in polarity. These pulses modulate either the IMPATT diode current or the varactor diode bias to deviate the millimeter-wave output frequency according to the input pulse polarity. The IMPATT diode oscillator-deviator output is down converted to an IF signal in the 1.3 ± 0.25 GHz band. The IF signal is then amplified and injected through a 500 MHz bandwidth filter to a phase-locked oscillator. The function of this oscillator is to act as a limiter thus removing amplitude modulation from the frequency modulated signal. The output of the phase-locked oscillator is further amplified before being injected into the differential phase detector and timing recovery circuit. The latter gives two output voltages. The

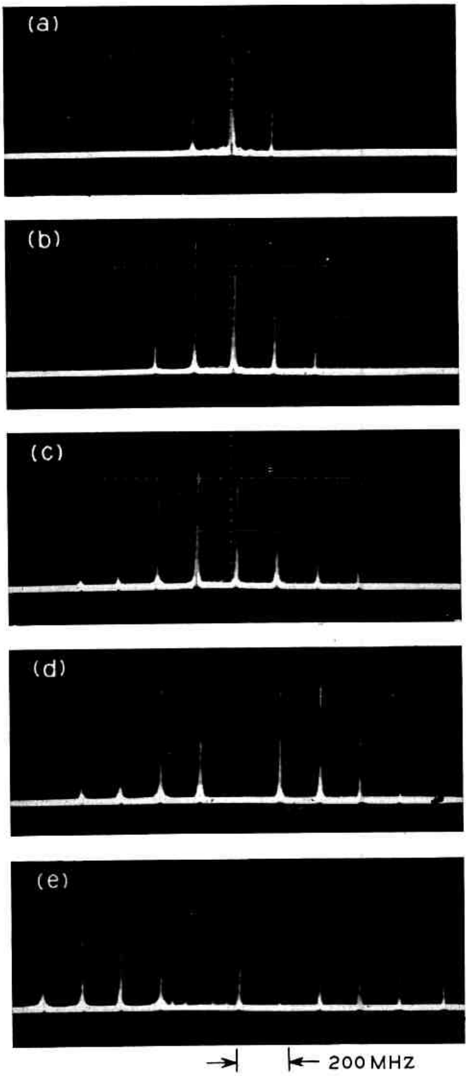


Fig. 4 — FM spectrum at 160 MHz modulation of varactor-tuned oscillator. Complete carrier suppression is shown in (d) with 6 dBm IF power. Others are (a) −9 dBm, (b) −4 dBm, (c) 1 dBm, and (e) 11 dBm (first sideband suppression).

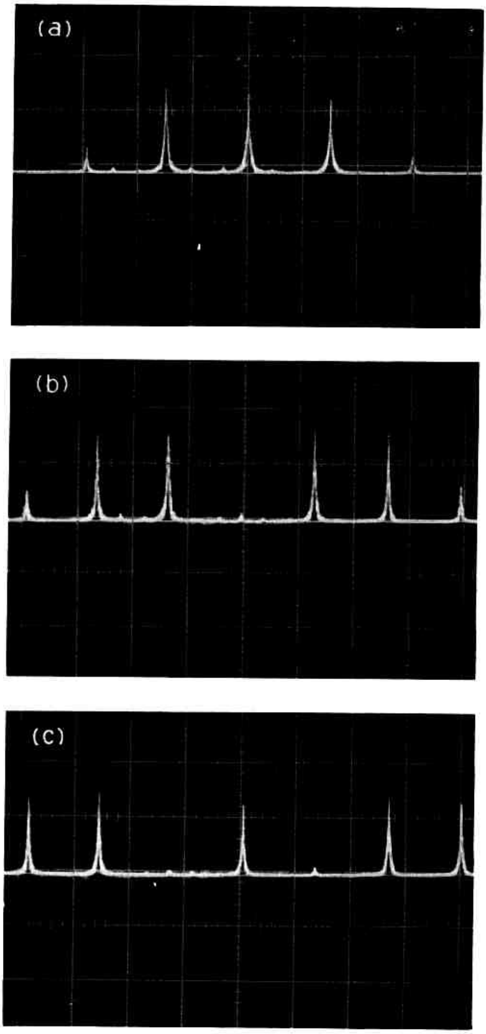


Fig. 5 — FM spectrum at 160 MHz modulation on the bias current. (a) IF power of −11 dBm for equal sidebands and carrier, $\Delta f/f = 1.435$, (b) IF power of −6 dBm for carrier suppression $\Delta f/f = 2.405$, and (c) IF power of −2 dBm for first sideband suppression, $\Delta f/f = 3.832$.

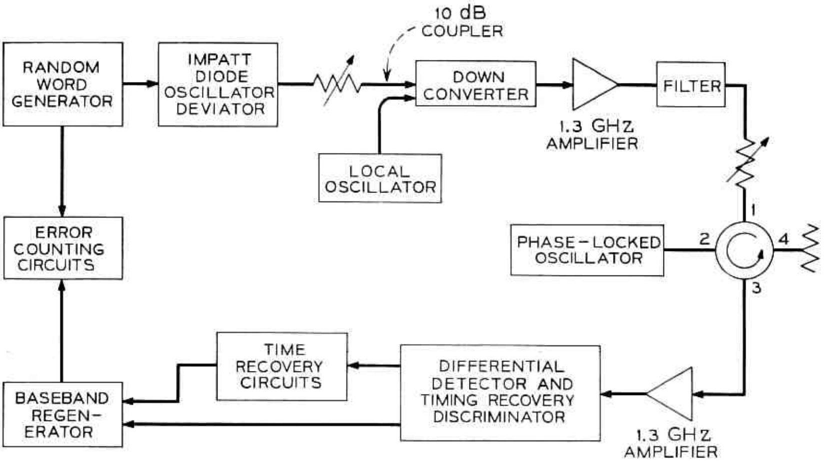


Fig. 6 — Circuit block diagram for the error rate measurement.

first output is ideally a replica of the baseband drive signal from the random word generator. The second output provides a timing signal for the regenerator, which samples the polarity of the first input at the time of arrival of the second input. The regenerator output is then compared with the random word generator output in an error counting circuit. (Reference 12 gives further details of circuit components.)

For the random-word modulation just described, the drive required at the 160 megabit rate was about 0.1 volt peak-to-peak (to a 75Ω load) for the varactor-tuned deviation scheme, and 0.03 volt peak-to-peak for the current-tuned deviation scheme at the optimum error-rate. Figure 7 compares typical eye-diagrams for the two deviation methods tested. For the latter case, the measured error-rate was 10^{-7} for a 16 dB signal-to-noise ratio, which was 2 dB worse than that obtained with a 1.3 GHz tunnel-diode deviator.[2] The major problem appeared to be the FM noise on the IMPATT oscillator output when it was optimally tuned for best deviation. The noise could have been reduced by increasing the circuit *Q* and consequently increasing the IF drive power.

Error rate improvement could have been obtained by matching the input impedance to the IMPATT bias circuit over the bandwidth of the baseband signal. A narrow band matching results in distortion of the baseband pulses.

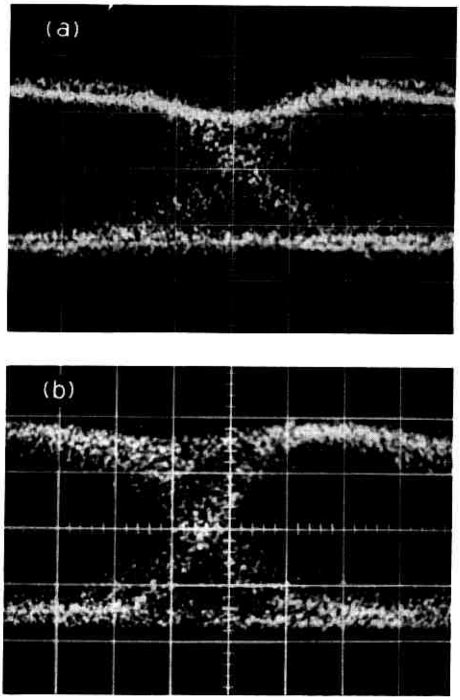


Fig. 7 — "Eye-diagrams" of the differentially-coherent-detected random word signal. (a) With the varactor-tuned FM deviator. (b) With the current-tuned FM deviator.

V. MEASUREMENTS OF SUBHARMONIC FREQUENCIES

As mentioned in Section I, the oscillator circuit under study provides at least two resonant circuits for the IMPATT diode, namely, an output circuit which is resonant at the transit-time frequency in M-band (50 to 75 GHz) and an "idler" circuit which is resonant at one-half the transit-time frequency in V-band (26.5 to 40 GHz). The two-frequency arrangement, with the frequencies harmonically related, has improved the tuning sensitivity of the oscillators, as we have shown in previous sections.

Although the existence of such a subharmonic oscillation can be deduced from the experiments described below, it is not surprising in view of the highly nonlinear nature of the impact ionization and avalanche process. Indeed, small-signal theories have predicted that the negative resistance exists over, at least, an octave in frequency.[13–15]

Parametric oscillations and mixing have been observed by DeLoach, Johnston, Evans, and Haddad.[16, 17] High-efficiency, low frequency, subtransit time oscillations have also been reported by Prager, Johnston, Scharfetter, and others,[18–19] and analyzed recently by Scharfetter and others.[20] In Scharfetter's analysis, the low-frequency oscillation is not necessarily harmonically related to the fundamental transit-time frequency. In recent work, Swan[3] showed that improvements in both output power and tuning bandwidth are obtained by providing an idler circuit which resonates at the second harmonic of the transit-time frequency. This result is similar to our findings except that in our case the idler is at one-half the transit-time frequency.

5.1 *Measurements*

Since the idler frequency is below the cutoff frequency of the RG-98/U waveguide, it is impossible to detect its existence directly at the output port of the oscillator as shown in Fig. 1. However, the existence of such a signal at the subharmonic frequency would result in mixing with the fundamental to produce an output at 3/2 the fundamental frequency. It was found that the next higher-order harmonic detected was always 3/2 of the oscillation frequency independent of bias current for all the diodes tested.

For direct detection of the below-cutoff subharmonic oscillation dielectric-filled waveguide tapers which had a cutoff frequency of 18 GHz were used as shown in Fig. 8a. However, a short section of air-filled RG-98/U waveguide with about 30 dB attenuation at 30 GHz remained between the IMPATT diode and the output waveguide taper. This arrangement retained the same subharmonic oscillation circuit conditions. Yet, if enough power exists for the signal at the subharmonic frequency to pass through the short section of air-filled RG-98/U waveguide, both the fundamental and the subharmonic frequencies should be present at the output port. A mixer and a spectrum analyzer were used to detect the subharmonic, while a wavemeter and a diode detector were used for the fundamental frequency. There was indeed an appreciable amount of power (estimated at 9 dBm at the diode) at the subharmonic of 27 GHz; the fundamental was exactly twice the subharmonic frequency, or 54 GHz, within measurement error.

When the short section of the RG-98/U waveguide preceding the diode wafer was also filled with dielectric the power output at the subharmonic was increased appreciably, but the frequency shifted

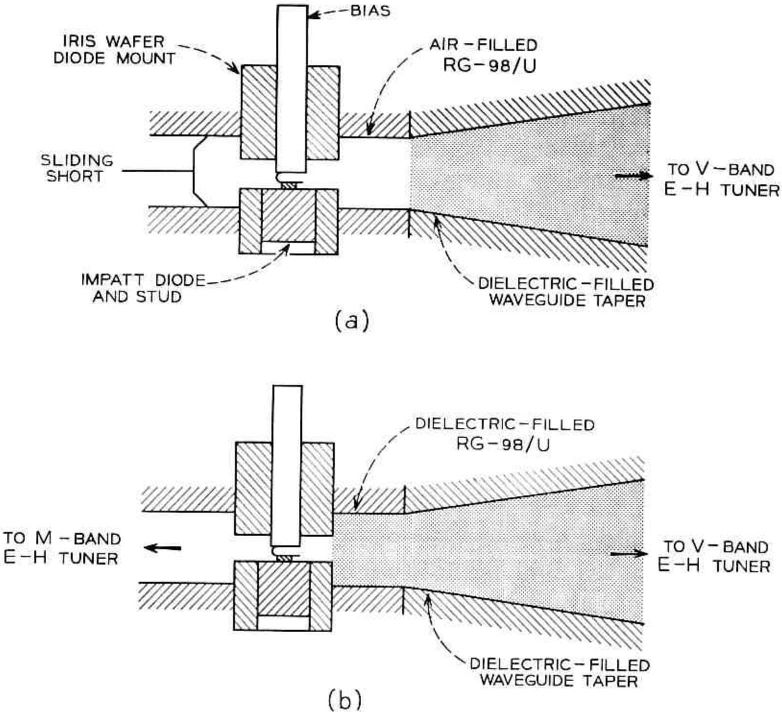


Fig. 8 — Waveguide arrangements for measuring the subharmonic signal.

slightly because of the change in loading impedance. The measurements of the subharmonic power on several diodes are summarized in Table III.

When the sliding short on one side of the oscillator mount was replaced by an E-H tuner followed by a wavemeter and a detector, as shown in Fig. 8b, both the fundamental and the subharmonic could be measured simultaneously. The results again confirmed that the subharmonic frequency was exactly one-half of the fundamental as shown in Fig. 9 for diode No. 3. This harmonic relation held true for any bias current.

The E-H tuner in the M-band circuit serves as an impedance matching device. By mismatching the oscillator to the load in the M-band circuit (thus reducing the 58 GHz power delivered to the load), the power delivered to the V-band load increases, and vice versa. Thus for efficient operation of IMPATT diodes, the subharmonic (and the harmonics)[3] should be reactively terminated. Likewise maximum sub-

TABLE III—MEASUREMENT OF POWER AT ONE-HALF THE TRANSIT-TIME FREQUENCY

Diode No.*	Bias current (mA)	Frequency (GHz)	Power output (dBm)
1	115	28.13	12.05
2	129	28.36	11.2
3	140	29.05	12.0
5	140	29.51	10.2
6	150	34.48	16.9
7	150	32.25	15.2
8	150	32.98	14.7
9	150	31.91	11.1

* The diode number is consistent with that in Table II.

harmonic power can be obtained when the fundamental and all harmonics are reactively terminated. Using this approach, by reactively terminating both the fundamental and the subharmonic, we obtain an output power at 3/2 the transit-time frequency of about 3.3 dBm at 86 GHz.

5.2 *Harmonic Phase-Locking*

The Fig. 8b circuit arrangement also was used for harmonic phase-locking. The locking-signal was injected in the V-band end through

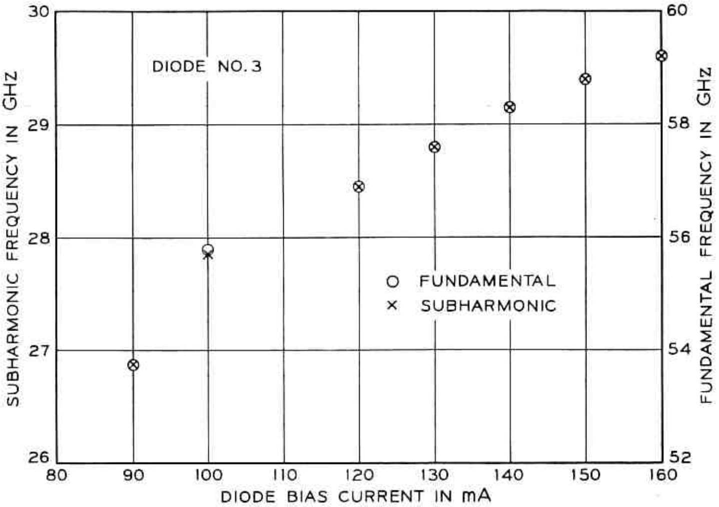


Fig. 9—Fundamental and subharmonic frequencies as a function of diode bias current.

a 10 dB directional coupler. Both the locked signals at the V-band and at the M-band were measured. The gain-bandwidth product $(2\Delta f/f_0)$ $(P_0/P_i)^{1/2}$ was about 0.05 to 0.07 for the V-band output.* The M-band output frequency was simultaneously phase-locked with the locked-bandwidth exactly twice that of the V-band. Since the power output in the M-band was about 1 mW compared with about 6 mW at V-band, the apparent gain bandwidth product was much less.

5.3 *V-Band Circuit with Cap Structure*

The experiments of Section 5.1 were conducted on IMPATT diodes in the iris circuit shown in Fig. 1. The same behavior was observed in a different circuit structure. A V-band oscillator was constructed using a resonant cap structure similar to the circuit described in Ref. 1. (see Fig. 10.) Caps were made in various diameters, and could be slid up and down the center bias rod, thus permitting the frequency of oscillation to be varied. A diode with very similar characteristics (breakdown voltage and capacitance) as the ones used before was selected from the same batch (LO 1114). The fundamental oscillation was in the range of 64 to 72 GHz with an output power of 10 dBm at 67 GHz for 150 mA bias. The subharmonic power was also detected.

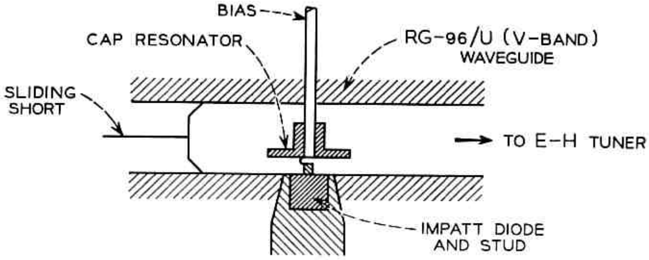


Fig. 10 — A V-band IMPATT oscillator using cap resonator structure.

To optimize the fundamental power, a V-band to M-band taper was used in addition to an E-H tuner so that the subharmonic oscillation was reactively terminated. Table IV shows the results. Notice that when the diode was biased at the same current level, the subharmonic frequency was close to ½ the fundamental frequency. The

* Here we define: f_0 = free running oscillation frequency
P_0 = free running oscillation output power
P_i = injected power at frequency $f_0 \pm \Delta f$
$2\Delta f$ = bandwidth over which the oscillator is phase-locked.

TABLE IV—OSCILLATION IN V-BAND CIRCUIT WITH CAP-RESONATOR STRUCTURE

Subharmonic			Fundamental (Subharmonic reactively terminated)		
Bias (mA)	Frequency (GHz)	Power output (dBm)	Bias (mA)	Frequency (GHz)	Power output (dBm)
80	32.75	7.5	80	64.95	4.0
100	32.87	9.5	105	65.7	5.7
150	33.4	14.0	150	67.0	10.5
160	36.0	14.0	160	72.0	11.1

slight difference resulted from the reactive termination of the subharmonic which required slight retuning for maximum output power at the fundamental frequency.

5.4 *Comparison of the Outputs*

Table V compares the fundamental and subharmonic power outputs in the two different circuits. For the iris-wafer circuit the diode was mounted in the RG-98/U waveguide; dielectric-filled waveguide and tapers were used when the subharmonic frequency power was measured. The dielectric-filled waveguide and tapers had an insertion loss of 1 dB in the frequency range of interest, which was taken into account for the power listed in the fourth column in Table V. For the resonant cap circuit, the diode was mounted in the V-band wave-

TABLE V—COMPARISON OF FUNDAMENTAL AND SUBHARMONIC POWER OUTPUTS

		Subharmonic		Fundamental	
Diode No.	Bias (mA)	Frequency (GHz)	Power (dBm)	Frequency (GHz)	Power (dBm)
			Iris circuit		
1	115	28.13	12.05	55.02	9.5
3	140	29.05	12.0	57.84	9.9
5	140	29.51	10.2	59.5	7.3
6	150	34.48	16.9	69.1	11.5
7	150	32.25	15.2	63.93	11.1
8	150	32.98	14.7	66.25	10.6
9	150	31.91	11.1	63.7	7.6
			Resonant Cap Circuit		
10	150	33.4	14	67.0	10.5
10	160	36.0	14	72.2	11.1

guide as previously described, so that no dielectric-loading was necessary.

In both circuits, the subharmonic power was maximized by reactively terminating the fundamental power, and vice versa. The subharmonic power is generally greater than that at the fundamental output by about 3 to 6 dB. This is in general agreement with Johnston, Schafetter, and others.[19–20] However, the current density used here is the same as for the diodes operated in the fundamental frequency mode, and the frequencies are harmonically related to each other.

VI. CONCLUSIONS

By using an idler resonance at the subharmonic frequency, one can design oscillators with 20 percent tuning range. Such tunable oscillators and frequency deviators were built and worked satisfactorily in an experimental millimeter-wave PCM repeater system.

Multiple frequency circuits for IMPATT diode oscillators offer an important means of generating millimeter waves with useful power. For example, 2 mW of power at 3/2 the transit-time frequency, or 86 GHz, has been obtained.

VII. ACKNOWLEDGMENT

The authors particularly wish to thank T. Misawa for supplying the IMPATT diodes and for helpful discussions. We would also like to thank F. A. Braun for mounting diodes into various circuits, and G. D. Mandeville for making the error-rate measurements in the PCM system.

REFERENCES

1. Lee, T. P., Standley, R. D., and Misawa, T., "A 50 GHz Silicon IMPATT Diode Oscillator and Amplifier," 1968 International Solid-State Circuit Conference Digest of Technical Papers, *IX* (February 1968), pp. 156–157.
2. Hubbard, W. M., and others, "A Solid-State Regenerative Repeater for Guided Millimeter-Wave Communication Systems," B.S.T.J., *46,* No. 9 (November 1967), pp. 1977–2018.
3. Swan, C. B., "IMPATT Oscillator Performance Improvement With Second Harmonic Tuning," 26th Annual Conference on Electron Device Research, Boulder, Colorado, June 19–21, 1968.
4. W. M. Sharpless, "Wafer-type Millimeter-Wave Rectifiers," B.S.T.J., *35,* No. 6 (November 1956), pp. 1385–1420.
5. Slater, J. C., *Microwave Transmission,* New York: McGraw-Hill Book Company, 1942.
6. Lewin, L., *Advanced Theory of Waveguides,* London: Iliffe & Sons, Ltd., 1951.
7. Tsung-Shan Chen, "Waveguide Resonant Iris Filters with Very Wide Passband and Stopbands," Int. J. Elec. *J(21),* No. 5, (1966), pp. 401–424.

8. Misawa, T., "CW Millimeter-Wave IMPATT Diodes with Nearly Abrupt Junctions," Proc. IEEE, *56,* No. 2 (February 1968), pp. 234–235.
9. Sharpless, W. M., unpublished work. (This tuning method was first used by W. M. Sharpless for an X-band tunnel diode oscillator.)
10. Burrus, C. A., "Planar Diffused Gallium Arsenide Millimeter Wave Varactor Diodes," Proc. IEEE, *55,* No. 6 (June 1967), pp. 1104–1105.
11. Lee, T. P. and Burrus, C. A., "A Millimeter-wave Quadrupler and an Up-converter using Planar Diffused GaAs Varactor Diodes," IEEE Transactions on Microwave Theory and Techniques, *MTT-16,* No. 5 (May 1968), pp. 287–295.
12. Hubbard, W. M. and Mandeville, G. D., "Experimental Verification of the Error-Rate Performance of Two Types of Regenerative Repeaters for Differentially Coherent Phase-shifted-keyed Signals" B.S.T.J., *46,* No. 6 (July 1967), pp. 1173–1202.
13. Read, W. T., Jr., "A Proposed High-Frequency Negative Resistance Diode," B.S.T.J., *37,* No. 2 (March 1958), pp. 401–446.
14. Misawa, T., "Negative Resistance in p-n Junctions under Avalanche Breakdown Conditions," Part I and Part II, IEEE Transactions on Electron Devices, *ED-13,* No. 1 (January 1966), pp. 137–151.
15. Gummel, H. K. and Scharfetter, D. L., "Avalanche Region of IMPATT Diodes," B.S.T.J., *45,* No. 10 (December 1966), pp. 1797–1827.
16. DeLoach, B. C. and Johnston, R. L., "Avalanche Transit-time Microwave Oscillators and Amplifiers," IEEE Transactions on Electron Devices, *ED-13,* No. 1 (January 1966), pp. 181–186.
17. Evans, W. J. and Haddad, G. I., "Frequency Conversion in Read Diodes," Informal Conference on Active Microwave Effects in Bulk Semiconductors, New York, February 2–3, 1967; Also Evans, W. J., "Non-linear and Frequency Conversion characteristics of IMPATT Diodes," Tech. Rep. No. 104, Electron Physics Laboratory, The University of Michigan, Ann Arbor, Michigan, February 1968.
18. Prager, H. J., Chang, K. K. N., and Weisbrod, S., "High Power, High Efficiency Silicon Avalanche Diodes at Ultrahigh Frequencies," Proc. IEEE, *55,* No. 4 (April 1967), pp. 586–587.
19. Johnston, R.L. and Scharfetter, D. L.,"Low Frequency High Efficiency Oscillations in Ge IMPATT Diodes," 26th Annual Conference on Electron Device Research, Boulder, Colorado, June 19–21, 1968.
20. Scharfetter, D. L., Bartelink, D. J., and Johnson, R. L., "Computer Simulation of Low Frequency Oscillation in Ge IMPATT Diodes," 26th Annual Conference on Electron Device Research, Boulder, Colorado, June 19–21, 1968.

Minimal Synthesis of Two-Variable Reactance Matrices*

By T. N. RAO

(Manuscript received July 3, 1968)

A simple algebraic method stemming from ideas in minimal state-variable realization theory is developed for the synthesis of two-variable reactance matrices. The method rests mainly on the factorization of a one variable polynomial matrix which is para-Hermitian and positive semidefinite on the imaginary axis, and always yields a realization minimal in both variables.

I. INTRODUCTION

Two-variable reactance functions and matrices, originally introduced to represent the characteristics of lumped passive networks with variable elements,[1, 2] have become more important because of their application to the synthesis of lumped-distributed networks. Ansell first showed the two-variable reactance property of networks composed of lossless transmission lines and lumped reactances.[3] The two-variable theory has also been applied to the synthesis of networks consisting of lumped resistors capacitors and uniformly distributed RC lines,[4, 5] which are of importance in microelectronic structures.[6, 7] Besides the various applications, the two-variable reactance theory is of theoretical interest in itself since it can be shown that passive RLC synthesis is a special case of two-variable reactance synthesis.[2]

Koga[8] demonstrated that every $n \times n$ two-variable reactance matrix $W(p, s)$ can be realized as the impedance seen at the first n ports of a lossless $(n+qr)$-port network in the p-plane terminated at its last qr ports with unit inductors in the s-plane; q is the rank of $W(p, s)$, and r is the highest degree of s in the least common denominator of

* This work is based on Chapter III of the author's dissertation, "Synthesis of Lumped-Distributed RC Networks" submitted in partial fulfilment of the requirements for the Ph.D. degree at Stanford University, May 1967.

the elements of w. The method is quite complicated and rests heavily on the theory of algebraic functions and the structure of para unitary matrices. Also it does not guarantee the use of a minimum number of elements. Youla[9] solved the problem of synthesizing a lossless two-variable scattering matrix by adapting an earlier method for synthesizing one-variable scattering matrices.[10] The method could be adapted to the direct synthesis of an impedance matrix but appears to be unduly complicated because of the need to find the transformation required to transform a generally unrealizable coupling network into a realizable one.

A simple algebraic method stemming from ideas in minimal state-variable realization theory and having similar beginnings as that of Youla[9] is developed here for the synthesis of two-variable reactance matrices. The method rests mainly on the factorization of a one-variable polynominal matrix which is para Hermitian* and positive semidefinite on the imaginary axis. Such a factorization is well known in n-port network theory and once it is accomplished, the coupling network is obtained by simple matrix operations. Furthermore the method always yields a network minimal in both types of elements.

We first introduce some basic definitions and necessary theorems, and later we add more as the need arises. The synthesis procedure is developed in Section III. Since the various proofs involved are rather indirect and tend to cloud the simplicity of the actual procedure, the synthesis procedure is outlined in Section IV. The reader interested only in the procedure and not in the theory behind it may go directly to Section IV where step-by-step instructions are given for the synthesis of any two variable reactance matrix. In Section V an example is worked out. The notation used in this paper is almost the same as found in earlier work to assure easy reading for those familiar with it.[9] Capital letters indicate matrices; bold face letters indicate matrix transposition. A superscript dagger indicates the substitution of $-s$ or $-p$ for s and p respectively, in the case of two-variable functions.

II. BASIC DEFINITIONS AND THEOREMS

The basic notion in the two variable theory is that of a two variable positive real matrix, which is a straightforward extension of the same notion in the one variable theory (See p. 96 of Ref. 11 and p. 32 of Ref. 8).

* A matrix $A(p)$ is said to be para-Hermitian if $A(p) = \mathbf{A}^{\dagger}(p)$ where the bold face letter denotes matrix transposition and the superscript dagger denotes replacement of p by $-p$.

Definition 1: An $n \times n$ matrix $W(p, s)$ is said to be a two variable positive real matrix if

(*i*) W is real for real p and s.
(*ii*) W is analytic in the domain $\mathrm{Re}\, p > 0$ and $\mathrm{Re}\, s > 0$.
(*iii*) $W + \mathbf{W}^*$ is positive semidefinite in the domain $\mathrm{Re}\, p > 0$ and $\mathrm{Re}\, s > 0$.

By statements such as: "W is analytic" in the definition and in what follows, we mean, "each element of W is analytic." A two variable function is said to be analytic at a point if it has a total differential at the point. The bold face letter indicates matrix transposition, and the superscript star indicates the complex conjugation of each element.

If $W(p, s)$ satisfies conditions (ii) and (iii) of the definition and not necessarily condition (i), it will be called a two variable positive matrix.

Definition 2: An $n \times n$ matrix $W(p, s)$ is said to be a two variable reactance matrix if

(*i*) W is a two variable positive real matrix.
(*ii*) $W + \mathbf{W}^{\dagger} \equiv 0$.

The superscript dagger indicates the operation of substituting $-p$ and $-s$ for p and s in the original matrix. This definition of a two variable reactance matrix is similar to the corresponding one in the one variable theory. (See p. 102 of Ref. 11 and p. 32 of Ref. 8.) Analogously, as in the one variable case (p. 117 of Ref. 11), it is generally hard to check if condition (*iii*) of Definition 1, which involves the whole domain $\mathrm{Re}\, p > 0$ and $\mathrm{Re}\, s > 0$, is satisfied for a given two variable matrix; we would like to find an equivalent set of conditions that are easier to check. In the case of two variable reactance matrices, the following theorem proved by Ozaki and Kasami[2] in the scalar case, and extended to nonsymmetric matrices by Koga, (p. 33 of Ref. 8) serves this purpose.

Theorem 1: The necessary and sufficient conditions for an $n \times n$ matrix $W(p, s)$ to be a two variable reactance matrix are:

(*i*) *W is rational in p and s, and real for real p and s.*
(*ii*) *W is analytic in the domain $Re\, p > 0$; $Re\, s > 0$.*
(*iii*) $W \equiv -\mathbf{W}^{\dagger}$.
(*iv*) *For any (p_0, s_0) with $Re\, p_0 = Re\, s_0 = 0$, which is a regular*

point of W, poles of $W(p_0, s)$ and $W(p, s_0)$ are simple and restricted to the imaginary s and p axes respectively.

(v) $\partial W/\partial p$ and $\partial W/\partial s$ are positive semidefinite Hermitian for $Re\ p = Re\ s = 0$, except at poles.

The proof of this theorem can be found on p. 33 of Ref. 8. We will interpret the above conditions on a physical basis. Assuming that a network realization consisting of reactances in the p and s-planes exists for $W(p, s)$, condition *i* is fairly obvious, since the general loop impedance will be a real rational function in p and s. Condition *iii* is also an obvious consequence of this reason, since the substitution of $-p$ and $-s$ in $W(p, s)$ is equivalent to changing the sign of all element values and hence of every branch and loop impedance. Under the assumption of existence of a two element kind of reactance network corresponding to the given $W(p, s)$, condition *iv* is also clear, because p is fixed as a pure imaginary number, the p-type elements can be considered as "frequency insensitive reactances," and their presence in a network consisting of pure reactances in the s-plane cannot create poles off the imaginary s axis. Similar reasoning justifies condition v for s fixed at any imaginary number, the positive semidefiniteness of $\partial W/\partial p$ can be considered as an extension of the positive slope of a reactance function in the one variable theory.

The necessary and sufficient conditions for a two variable reactance function are not discussed separately, since scalars can be considered as a special case of a reactance matrix.

If $W(p, s)$ has a pole at $p = p_0$, independent of the value of s, p_0 is said to be an s-independent pole of W. The following theorem (see p. 34 of Ref. 8) concerning such poles is important for the synthesis method to be given.

Theorem 2: A two variable reactance matrix $W_0(p, s)$ can be decomposed as

$$W_0(p, s) = W_1(p) + W_2(s) + W(p, s)$$

where W_1 and W_2 are reactance matrices in p and s, respectively, and W is a two variable reactance matrix with no p-independent or s-independent poles.

Any given two variable reactance matrix $W_o(p, s)$, by virtue of the above theorem, can be realized as a series connection of networks having W_1, W_2, and W as their impedance matrices, as shown in Fig. 1. Since W_1 and W_2 can be realized by existing techniques (See chapter 7 of Ref. 11) the given W_o can be realized if a method of synthesis is

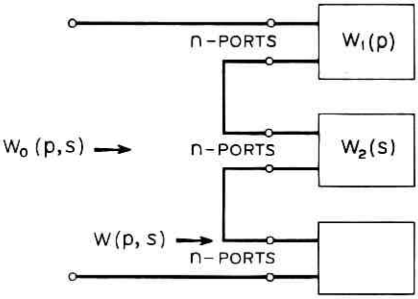


Fig. 1 — Interpretation of theorem 2.

found for W. Henceforth we assume that the given reactance matrix has no p-independent or s-independent poles.

III. SYNTHESIS OF TWO VARIABLE REACTANCE MATRICES

Let us assume that there is a passive n-port network representation, consisting of p- and s-type reactances, gyrators, and ideal transformers, for a given two variable $n \times n$ reactance matrix $W(p, s)$. In such a network it is always possible to replace each s-type capacitor by a gyrator-s-type inductor combination and then isolate all the s-type inductors, of which we assume there are k, as shown in Fig. 2, without changing the impedance seen at the prescribed ports. If we further assume that the $(n + k)$-port coupling network, consisting of p-type reactances, ideal transformers, and gyrators has a Z matrix, then the impedance matrix $W(p, s)$ seen at the first n ports is given by

$$W(p, s) = z_{11}(p) - z_{12}(p)[z_{22}(p) + sl_k]^{-1}z_{21}(p) \tag{1}$$

where $Z(p)$, the impedance matrix of the coupling network is given by

$$Z(p) = \begin{bmatrix} z_{11}(p) & z_{12}(p) \\ z_{21}(p) & z_{22}(p) \end{bmatrix}. \tag{2}$$

Since the coupling network is a lossless network in the p-plane

$$Z = -\mathbf{Z}^{\dagger} \tag{3}$$

and we have

$$W(p, s) = z_{11}(p) + z_{12}(p)[z_{22}(p) + sl_k]^{-1}\mathbf{z}_{12}^{\dagger}(p). \tag{4}$$

Next we show, by algebraic means, that every two variable reactance matrix can be decomposed into the form in equation (4), such that

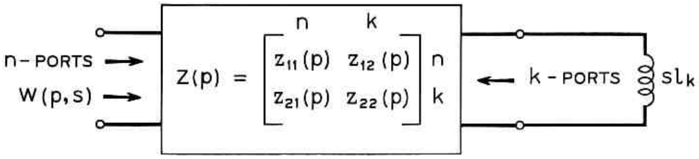


Fig. 2 — Extraction of S-type inductors.

$Z(p)$ of equation (2) describes a lossless network. Once such a decomposition is found, we can realize the given $W(p, s)$ by realizing $Z(p)$ by any of the existing techniques (see chapter 7 of Ref. 11) and terminating it at its last k ports with unit inductors in the s-plane.

To establish that any given two variable reactance matrix $W(p, s)$ can be decomposed as shown in equation (4), we first expand $W(p, s)$ and the expression on the right side of equation (4) about $s = \infty$ and find the expressions that relate z_{11}, z_{12}, and z_{22} with the expansion coefficients of $W(p, s)$. We then show that a set z_{11}, z_{12}, and z_{22}, which satisfies the above relations and at the same time guarantees that the $Z(p)$ of equation (2) is a reactance matrix in p, can always be found.

The given two variable reactance matrix $W(p, s)$ can be assumed to have no p-independent or s-independent poles by virtue of Theorem 2 and hence can be written in the form

$$W(p, s) = \frac{B_0(p)s^r + B_1(p)s^{r-1} + \cdots + B_r(p)}{a_0(p)s^r + a_1(p)s^{r-1} + \cdots + a_r(p)} \tag{5}$$

where the $B_i(p)$ are real polynomial matrices in p and the scalar

$$g(p, s) = a_0(p)s^r + a_1(p)s^{r-1} + \cdots + a_r(p) \tag{6}$$

is the least common denominator of the entries in $W(p, s)$. For any ordinary value of p, $W(p, s)$ can be expanded in the neighborhood of $s = \infty$ as[9]

$$W(p, s) = A_{-1}(p) + \sum_{l=0}^{\infty} \frac{A_l(p)}{s^{l+1}}. \tag{7}$$

Expanding the right side of equation (4) in the neighborhood of $s = \infty$

$$z_{11}(p) + z_{12}(p)[z_{22}(p) + s1_k]^{-1}\mathbf{z}_{12}^{\dagger}(p) = z_{11} + (-1)^l \sum_{l=0}^{\infty} \frac{z_{12}z_{22}^{l}\mathbf{z}_{12}^{\dagger}}{s^{l+1}}. \tag{8}$$

For the equality in equation (4) to hold, we identify

$$z_{11}(p) = A_{-1}(p) = W(p, \infty) \tag{9}$$

and

$$A_l(p) = (-1)^l z_{12} z_{22}^l \mathbf{z}_{12}^{\dagger} \qquad l = 0, 1, 2, \cdots . \tag{10}$$

Since the $Z(p)$ formed out of z_{11}, z_{12} and z_{22}

$$Z(p) = \begin{bmatrix} z_{11} & z_{12} \\ -\mathbf{z}_{12}^{\dagger} & z_{22} \end{bmatrix} \tag{11}$$

has to describe a lossless network in the p plane, we must have

$$Z = -\mathbf{Z}^{\dagger}$$

as given by equation (3), and hence

$$z_{11} = -\mathbf{z}_{11}^{\dagger} \tag{12}$$

and

$$z_{22} = -\mathbf{z}_{22}^{\dagger} . \tag{13}$$

With the identification in equation (9), equation (12) is always satisfied, since by equation (9)

$$z_{11} = W(p, \infty) = -\mathbf{W}(-p, -\infty),$$

and thus z_{11} is uniquely determined. The problem is to chose a pair z_{12}, z_{22} to satisfy equation (10) and at the same time guarantee that equation (11) describes a lossless network in the p-plane. For $Z(p)$ to describe a lossless network, it must be positive real and satisfy equation (3).

Before proceeding further, we would like to know more about $A_l(p)$, the expansion coefficients in equation (7). By equating the right sides of equations (5) and (7),

$$\begin{aligned} & B_0(p)s^r + B_1(p)s^{r-1} + \cdots + B_r(p) \\ & = [a_0(p)s^r + a_1(p)s^{r-1} + \cdots + a_r(p)]\left[A_{-1}(p) + \sum_{l=0}^{\infty} \frac{A_l(p)}{s^{l+1}}\right] . \end{aligned} \tag{14}$$

Equating coefficients of like powers of s on both sides of equation (14), (see p. 207 of Ref. 12, Vol. II).

$$\begin{aligned} & a_0(p)A_{-1}(p) = B_0(p) \\ & a_1(p)A_{-1}(p) + a_0(p)A_0(p) = B_1(p) \\ & a_2(p)A_{-1}(p) + a_1(p)A_0(p) + a_0(p)A_1(p) = B_2(p) \\ & \vdots \end{aligned} \tag{15}$$

$$a_r(p)A_{-1}(p) + a_{r-1}(p)A_0(p) + \cdots + a_0(p)A_{r-1}(p) = B_r(p)$$

and

$$a_0(p)A_i(p) + a_1(p)A_{i-1}(p) + \cdots + a_r(p)A_{i-r}(p) = 0_n \text{ for } i \geqq r.$$

From equation (15) an expression for $A_l(p)$ can be written* in the convenient form (see p. 14 Ref. 9)

$$A_l(p) = \frac{(-1)^{l+1}}{a_0^{l+2}(p)} \begin{vmatrix} B_0(p) & a_0(p) & 0 & 0 & \cdots & 0 \\ B_1(p) & a_1(p) & a_0(p) & 0 & \cdots & 0 \\ B_2(p) & a_2(p) & a_1(p) & a_0(p) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B_l(p) & a_l(p) & a_{l-1}(p) & a_{l-2}(p) & \cdots & a_0(p) \\ B_{l+1}(p) & a_{l+1}(p) & a_l(p) & a_{l-1}(p) & \cdots & a_1(p) \end{vmatrix} \tag{16}$$

$$l = -1, 0, 1, 2, \cdots$$

$$B_l = 0_n \qquad \text{for} \quad l > r$$

$$a_l = 0 \qquad \text{for} \quad l > r$$

where the $(l + 2) \times (l + 2)$ determinant is expanded formally in terms of its first column. In equation (16) the B_i are matrices, the a_i are scalars, and the determinant is not a determinant in the usual sense. From equation (16), it can be seen that $A_l(p)$ is of the form

$$A_l(p) = \frac{\text{real polynomial matrix in } p}{a_0^{l+2}(p)}. \tag{17}$$

Another important property of the $A_l(p)$'s is obtained from the relation

$$W(p, s) = -\mathbf{W}(-p, -s)$$

which implies

$$A_{-1}(p) + \sum_{l=0}^{\infty} \frac{A_l(p)}{s^{l+1}} = -\mathbf{A}_{-1}(-p) - \sum (-1)^{l+1} \frac{\mathbf{A}_l(-p)}{s^{l+1}}. \tag{18}$$

Hence by equating like powers of s

*Alternate methods of obtaining these $A_l(p)$'s are by differentiation of $W(p, s)$

$$A_l(p) = \left. \frac{\partial^{l+1} W(p, s)}{\partial s^{l+1}} \right|_{s=\infty}$$

or by long division.

$$A_l = (-1)^l \mathbf{A}_l^\dagger . \tag{19}$$

If, for the purpose of choosing a pair z_{12}, z_{22} that satisfies equation (10) and, at the same time, guarantees that the $Z(p)$ of equation (11) describes a lossless network in the p-plane, we define $P_l(p)$ as

$$P_l(p) = \begin{bmatrix} z_{12} \\ z_{12}z_{22} \\ z_{12}z_{22}^2 \\ \vdots \\ z_{12}z_{22}^l \end{bmatrix} . \tag{20}$$

Then

$$P_l\mathbf{P}_l^\dagger = \begin{bmatrix} z_{12}\mathbf{Z}_{12}^\dagger & z_{12}\mathbf{Z}_{22}^\dagger\mathbf{Z}_{12}^\dagger & z_{12}\mathbf{Z}_{22}^{2\dagger}\mathbf{Z}_{12}^{2\dagger} & \cdots & z_{12}\mathbf{Z}_{22}^{l\dagger}\mathbf{Z}_{12}^\dagger \\ z_{12}z_{22}\mathbf{Z}_{12}^\dagger & z_{12}z_{22}\mathbf{Z}_{22}^\dagger\mathbf{Z}_{12}^\dagger & z_{12}z_{22}\mathbf{Z}_{22}^{2\dagger}\mathbf{Z}_{12}^\dagger & \cdots & z_{12}z_{22}\mathbf{Z}_{22}^{l\dagger}\mathbf{Z}_{12}^\dagger \\ z_{12}z_{22}^2\mathbf{Z}_{12}^\dagger & z_{12}z_{22}^2\mathbf{Z}_{22}^\dagger\mathbf{Z}_{12}^\dagger & z_{12}z_{22}^2\mathbf{Z}_{22}^{2\dagger}\mathbf{Z}_{22}^\dagger & \cdots & z_{12}z_{22}^2\mathbf{Z}_{22}^{l\dagger}\mathbf{Z}_{12}^\dagger \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{12}z_{22}^l\mathbf{Z}_{12}^\dagger & z_{12}z_{22}^l\mathbf{Z}_{22}^\dagger\mathbf{Z}_{12}^\dagger & z_{12}z_{22}^l\mathbf{Z}_{22}^{2\dagger}\mathbf{Z}_{22}^\dagger & \cdots & z_{12}z_{22}^l\mathbf{Z}_{21}^{l\dagger}\mathbf{Z}_{12}^\dagger \end{bmatrix} \tag{21}$$

In the above matrix, the entry in the ith row and jth column is $z_{12}z_{22}^i\mathbf{Z}_{22}^{j\dagger}\mathbf{Z}_{12}^\dagger$ and by equation (13)

$$z_{12}z_{22}^i\mathbf{Z}_{22}^{j\dagger}\mathbf{Z}_{12}^\dagger = (-1)^j z_{12}z_{22}^{i+j}\mathbf{Z}_{12}^\dagger . \tag{22}$$

Since we wish the equality in equation (10) to hold

$$z_{12}z_{22}^i\mathbf{Z}_{22}^{j\dagger}\mathbf{Z}_{12}^\dagger = (-1)^j z_{12}z_{22}^{i+j}\mathbf{Z}_{12}^\dagger = (-1)^j A_{i+j} . \tag{23}$$

If we define $T_l(p)$ as

$$T_l(p) = \begin{bmatrix} A_0(p) & A_1(p) & A_2(p) & \cdots & A_l(p) \\ -A_1(p) & -A_2(p) & -A_3(p) & \cdots & -A_{l+1}(p) \\ A_2(p) & A_3(p) & A_4(p) & \cdots & A_{l+2}(p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ (-1)^l A_l(p) & (-1)^l A_{l+1}(p) & (-1)^l A_{l+2}(p) & \cdots & (-1)^l A_{2l}(p) \end{bmatrix} \tag{24}$$

from equation (23), we can see that

$$T_l = P_l\mathbf{P}_l^\dagger . \tag{25}$$

Equation (25) suggests that a way of obtaining a pair z_{12}, z_{22} would be to form the matrix $T_l(p)$, factor it in the form of equation (25), and then try to identify z_{12} and z_{22} from these factors. We do not know in advance if the matrix $T_l(p)$ formed from the expansion coefficients of W about $s = \infty$ can always be factored as indicated in equation (25); hence we first study the properties of $T_l(p)$, to see if it can be factored in the desired form.

Consider the matrix $T_l(p)$ when $l = r$, r being the s-degree of $g(p, s)$, as given in equation (6),

$$T_r = \begin{bmatrix} A_0 & A_1 & A_2 & \cdots & A_{r-1} & A_r \\ -A_1 & -A_2 & -A_3 & \cdots & -A_r & -A_{r+1} \\ A_2 & A_3 & A_4 & \cdots & A_{r+1} & A_{r+2} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ (-1)^{r-1}A_{r-1} & (-1)^{r-1}A_r & (-1)^{r-1}A_{r+1} & \cdots & (-1)^{r-1}A_{2r-1} & (-1)^{r-1}A_{2r-1} \\ (-1)^r A_r & (-1)^r A_{r+1} & (-1)^r A_{r+2} & \cdots & (-1)^r A_{2r-1} & (-1)^r A_{2r} \end{bmatrix}. \tag{26}$$

The matrix obtained by deleting the last column and row in equation (26) is T_{r-1}, and by equation (15) it is easy to see that the last column is a linear combination of the first r columns. Hence*

$$\text{rank } T_r = \text{rank } T_{r-1}$$

and

$$\text{rank } T_l = \text{rank } T_{r-1} \qquad \text{for} \quad l \geqq r - 1.$$

The rank of T_{r-1} is connected with the s-degree $\delta_s[W(p, s)]$ of $W(p, s)$ which is defined in Definition 3 (see p. 10 of Ref. 9).

Definition 3: The s-degree of a rational two variable matrix $W(p, s)$ is obtained from the rule

$$s = \text{degree of} \quad W(p, s) = \delta_s[W(p, s)] = \max_{p_0} \delta[W(p_0, s)]$$

where $\delta[W(p_o, s)]$ is the McMillan degree (see part II of Ref. 13) of $W(p_0, s)$. For any fixed p_o, $W(p_o, s)$ is a matrix of rational functions in s with its McMillan degree uniquely specified; hence the above definition uniquely specifies the s-degree of $W(p, s)$. The relationship between the s-degree of $W(p, s)$ and the rank of T_{r-1} is stated formally in the following lemma.

Lemma 1: The rank of $T_{r-1}(p)$ is equal to the s-degree of $W(p, s)$.

* By the rank of rational or polynomial matrix we mean the "normal rank," which is defined to be the rank everywhere except at a finite number of values of the variable.

The proof of this lemma for the one variable case can be found in Ref. 14 and on p. 200 of Ref. 10, and for the two variable case on p. 17 of Ref. 9.

To show that the matrix $T_{r-1}(p)$ can always be factored in the form of equation (25), we need the following lemma.

Lemma 2: The matrix $T_{r-1}(p)$ *defined by equation* (24) *for* $l = r - 1$ *satisfies*

(*i*) $T_{r-1} = \mathbf{T}_{r-1}^{\dagger}$
(*ii*) $T_{r-1}(j\omega)$ *is Hermitian and positive semidefinite.*

Proof:

Since $A_l = (-1)^l \mathbf{A}_l^{\dagger}$ by equation (19), the proof of i is readily seen from equation (26)

$$T_{r-1} = \mathbf{T}_{r-1}^{\dagger}. \tag{27}$$

To prove (*ii*), we first notice that by Theorem 1, for any real ω, $W(j\omega, s)$ has only simple poles, which are restricted to the imaginary axis in the s-plane. Hence $W(j\omega, s)$ can be expressed in the partial fraction form

$$W(j\omega, s) = A_{-1}(j\omega) + \sum_{i=1}^{r} \frac{R_i(j\omega)}{s - j\alpha_i(\omega)} \tag{28}$$

where, $R_i(j\omega)$ are the residue matrices at the poles $j\alpha_i(\omega)$, and the $\alpha_i(\omega)$ are real.

It is shown in Appendix A that the $R_i(j\omega)$ are Hermitian and positive semidefinite for each ω. Now, if each term in the sum on the right side of equation (28) is expanded about $s = \infty$, we have

$$W(j\omega, s) = A_{-1}(j\omega) + \sum_{i=1}^{r} \sum_{q=0}^{\infty} \frac{(j\alpha_i)^q}{s^{q+1}} R_i(j\omega). \tag{29}$$

For the purpose of comparison, equation (7), written with $p = j\omega$, is

$$W(j\omega, s) = A_{-1}(j\omega) + \sum_{q=0}^{\infty} \frac{A_q(j\omega)}{s^{q+1}}. \tag{30}$$

The right sides of equation (29) and (30) are expansions of $W(j\omega, s)$ about $s = \infty$, and because of the uniqueness of a power series expansion

$$A_q(j\omega) = \sum_{i=1}^{r} (j\alpha_i)^q R_i(j\omega). \tag{31}$$

By noting that the α_i are real and the $R_i(j\omega)$ are Hermitian and positive semidefinite for each ω, we have

$$A_0(j\omega) = \sum_{i=1}^{r} R_i(j\omega) \qquad \geqq 0 \tag{32}^*$$

$$A_1(j\omega) = j \sum_{i=1}^{r} \alpha_i R_i(j\omega) \tag{33}$$

$$A_2(j\omega) = -\sum_{i=1}^{r} \alpha_i^2 R_i(j\omega) \qquad \leqq 0 \tag{34}$$

$$\vdots$$

$$A_{4m-3}(j\omega) = -j \sum_{i=1}^{r} \alpha_i^{4m-3} R_i(j\omega) \tag{35a}$$

$$A_{4m-2}(j\omega) = -\sum_{i=1}^{r} \alpha_i^{4m-2} R_i(j\omega) \qquad \leqq 0 \tag{35b}$$

$$A_{4m-1}(j\omega) = -j \sum_{i=1}^{r} \alpha_i^{4m-1} R_i(j\omega) \tag{35c}$$

$$A_{4m}(j\omega) = \sum_{i=1}^{r} \alpha_i^{4m} R_i(j\omega) \qquad \geqq 0. \tag{35d}$$

By direct substitution of equation (33) into equation (24), $T_{r-1}(j\omega)$ can be written as

$$T_{r-1}(j\omega) = \sum_{i=1}^{r} \begin{bmatrix} R_i & j\alpha_i R_i & -\alpha_i^2 R_i & \cdots & (j\alpha_i)^{r-1} R_i \\ -j\alpha_i R_i & \alpha_i^2 R_i & j\alpha_i^3 R_i & \cdots & -(j\alpha_i)^r R_i \\ -\alpha^2 R_i & -j\alpha_i^3 R_i & \alpha_i^4 R_i & \cdots & (j\alpha)^{r+1} R_i \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ (-1)^{r-1}(j\alpha)^{r-1} R_i & (-1)^{r-1}(j\alpha)^r R_i & (-1)^{r-1}(j\alpha)^{r+1} R_i & \cdots & (-1)^{r-1}(j\alpha)^{2r-2} R_i \end{bmatrix}. \tag{36}$$

The matrix sum on the right side of equation (36) can be written

$$T_{r-1}(j\omega) = \sum_{i=1}^{r} L_i \begin{bmatrix} R_i & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \mathbf{L}_i^* \tag{37}$$

where

* By the notation $A \geqq 0$ or $A \leqq 0$, we mean that the associated Hermitian form of A is positive semidefinite or negative semidefinite.

$$L_i = \begin{bmatrix} 1_n & 0 & 0 & \cdots & 0 \\ j\alpha_i 1_n & 1_n & 0 & \cdots & 0 \\ \alpha_i^2 1_n & 0 & 1_n & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ (-1)^r (j\alpha)^{r-1} 1_n & 0 & 0 & \cdots & 1_n \end{bmatrix}^{-1} . \tag{38}$$

Since each $R_i(j\omega)$ is Hermitian and positive semidefinite for each ω, the sum on the right side of equation (37) is also Hermitian and positive semidefinite. Hence, we have proved the lemma.

We have shown that T_{r-1}, a matrix of rational functions, is para-Hermitian and positive semidefinite on the imaginary axis. Such a matrix can always be factored in the form shown in equation (25), (see p. 133 of Ref. 15). It is tempting to factor T_{r-1} at this stage and find z_{12}, z_{22} to satisfy the required conditions, but we will factor $a_0^{2r}\ T_{r-1}$ instead of T_{r-1} for the reason that the factors would be polynomial matrices.

From equation (17) we can see that $a_0^{2r}\ T_{r-1}$ is a polynomial matrix in p. To be able to factor $a_0^{2r}\ T_{r-1}$ in the required fashion, we have to show that $\hat{T} = a_0^{2r}\ T_{r-1}$ is para Hermitian and positive semidefinite on the $j\omega$ axis. To do this, we obtain the required additional information about the polynomial $a_0(p)$ from the following theorem. Since the theorem contains more information than we need at this point, we will only state it here; a proof is given in Appendix B.

Theorem 3: If

$$W(p, s) = \frac{B_0(p)s^r + B_1(p)s^{r-1} + \cdots + B_r(p)}{a_0(p)s^r + a_1(p)s^{r-1} + \cdots + a_r(p)}$$

is a two variable reactance matrix, then for all $i = 0, 1, \cdots, r$

(i) $\dfrac{B_i}{a_i}$ is a reactance matrix in p

(ii) $\dfrac{a_i}{a_{i+1}}$ is a reactance function in p

(iii) a_i has all its zeros on the $j\omega$ axis and these are simple

(iv) $\dfrac{\mathbf{X}B_iX}{\mathbf{X}B_{i+1}X}$ for all constant real $n \times 1$ vectors, X, is a reactance function in p

From this theorem, $a_o(p)$ can be represented as

$$a_0(p) = p^\nu \prod_i (p^2 + \omega_i^2) = \pm a_0(-p) \tag{39}$$

where $\nu = 0$ or 1. Hence

$$\hat{T}_{r-1} = a_0^{2r} T_{r-1} = \hat{\mathbf{T}}_{r-1}^\dagger . \tag{40}$$

From the form of a_o shown in equation (39) and Lemma 1, it can be seen that

$$\hat{T}_{r-1}(j\omega) \geqq 0 \tag{41}$$

except when simultaneously, $\nu = 1$ and r is odd; in which case

$$\hat{T}_{r-1}(j\omega) \leqq 0. \tag{42}$$

We will assume that $\hat{T}_{r-1}(j\omega) \geqq 0$ in developing the synthesis procedure and discuss the needed modification when $\hat{T}_{r-1}(j\omega) \leqq 0$ later.

If the s-degree of $W(p, s)$ is equal to k, by Lemma 1 the rank of $T_{r-1}(p)$ and hence of $\hat{T}_{r-1}(p)$ is k. Since $\hat{T}_{r-1} = \hat{\mathbf{T}}_{r-1}^\dagger$ and $\hat{T}_{r-1}(j\omega) \geqq 0$ there exists a factorization [16,17]

$$\hat{T}_{r-1}(p) = M(p)\mathbf{M}^\dagger(p) \tag{43}$$

where $M(p)$ is an $nr \times k$ polynomial matrix and has a left inverse $M^{-1}(p)$ which is analytic in Re $p > 0$.

From the definition of $\hat{T}_{r-1}$, we have

$$T_{r-1}(p) = \frac{M(p)\mathbf{M}^\dagger(p)}{a_0^{2r}}. \tag{44}$$

$M(p)$ can be partitioned into $n \times k$ blocks $M_i(p)$

$$M(p) = \begin{bmatrix} M_0(p) \\ \hdashline M_1(p) \\ \hdashline \vdots \\ \hdashline M_{r-1}(p) \end{bmatrix} \tag{45}$$

and hence

$$\mathbf{M}^\dagger(p) = [\mathbf{M}_0^\dagger(p) \,\vdots\, \mathbf{M}_1^\dagger(p) \,\vdots\, \cdots \,\vdots\, \mathbf{M}_{r-1}^\dagger(p)]. \tag{46}$$

Now by comparison of equation (45) with equation (20), we can immediately identify a suitable z_{12} as

$$z_{12} = \frac{M_0}{a_0^r}. \tag{47}$$

To find a suitable z_{22}, if we define T_d as

$$T_d(p) = \begin{bmatrix} -A_1 & -A_2 & -A_3 & \cdots & -A_r \\ A_2 & A_3 & A_4 & \cdots & A_{r+1} \\ -A_3 & -A_4 & -A_5 & \cdots & -A_{r+2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ (-1)^r A_r & (-1)^r A_{r+1} & (-1)^r A_{r+2} & \cdots & (-1)^r A_{2r} \end{bmatrix} \tag{48}$$

from equations (20), (21), and (25) we see that z_{22} must satisfy

$$\frac{1}{a_0^{2r}} M z_{22} \mathbf{M}^\dagger = T_d \, . \tag{49}$$

Even though equation (49) does not uniquely specify z_{22}, we can choose for z_{22}

$$z_{22} = a_0^{2r} M^{-1} T_d \mathbf{M}^{-1\dagger}. \tag{50}$$

From equation (19) and the definition of T_d, we see that $T_d = -\mathbf{T}_d^\dagger$ and hence

$$z_{22} = -\mathbf{z}_{22}^\dagger. \tag{51}$$

We now notice that by construction, the pair z_{12}, z_{22} defined by equations (47) and (50) satisfies

$$(-1)^l z_{12} z_{22}^l \mathbf{z}_{12}^\dagger = A_l \tag{10}$$

for all $0 \leqq l \leqq 2r - 2$. Our aim is to find z_{12} and z_{22} that satisfy equation (10) for all $l \geqq 0$. It is not immediately clear that the pair z_{12}, z_{22} defined by equations (47) and (50) satisfy equation (10) for all $l \geqq 0$.

To see that the chosen pair z_{12}, z_{22} does indeed satisfy equation (10) for all $l \geqq 0$ and not just for $0 \leqq l \leqq 2r - 2$, we introduce the generalized companion matrix $\Omega(p)$ defined by[10]

$$\Omega(p) = \begin{bmatrix} 0_n & 1_n & 0_n & \cdots & \cdots & 0_n \\ 0_n & 0_n & 1_n & \cdots & \cdots & 0_n \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_n & 0_n & 0_n & \cdots & 0_n & 1_n \\ -\frac{a_r}{a_0} 1_n & -\frac{a_{r-1}}{a_0} 1_n & -\frac{a_{r-2}}{a_0} 1_n & \cdots & -\frac{a_2}{a_0} 1_n & -\frac{a_1}{a_0} 1_n \end{bmatrix}. \tag{52}$$

From equation (15) it can be seen that

$$T_d = -T_{r-1}\mathbf{\Omega}. \tag{53}$$

Hence, by equation (43)

$$\begin{aligned} z_{22} &= -a_0^{2r} M^{-1} T_{r-1} \mathbf{\Omega} \mathbf{M}^{-1^\dagger} \\ &= -M^{-1} M \mathbf{M}^\dagger \mathbf{\Omega} \mathbf{M}^{-1^\dagger} \\ &= -\mathbf{M}^\dagger \mathbf{\Omega} \mathbf{M}^{-1^\dagger}, \end{aligned}$$

and by equation (51)

$$z_{22} = -\mathbf{M}^\dagger \mathbf{\Omega} \mathbf{M}^{-1^\dagger} = M^{-1} \Omega^\dagger M. \tag{54}$$

Hence

$$\begin{aligned} z_{22}^2 &= -M^{-1} \Omega^\dagger M \mathbf{M}^\dagger \mathbf{\Omega} \mathbf{M}^{-1^\dagger} \\ &= -a_0^{2r} M^{-1} \Omega^\dagger T_{r-1} \mathbf{\Omega} \mathbf{M}^{-1^\dagger} \\ &= -a_0^{2r} M^{-1} \Omega^{2^\dagger} T_{r-1} \mathbf{M}^{-1^\dagger} \\ &= -M^{-1} \Omega^{2^\dagger} M \end{aligned}$$

and

$$z_{22}^l = (-1)^{l-1} M^{-1} \Omega^{l^\dagger} M; \qquad l > 0. \tag{55}$$

From the definition of Ω, we see that $g(p, \xi)$ is its minimal polynomial, and hence the matrix polynomial

$$g(p, \Omega) = a_o \Omega^r + a_1 \Omega^{r-1} + \cdots + a_r 1_{nr} \equiv 0_{nr} \tag{56}$$

and hence

$$g(-p, \Omega^\dagger) = 0_{nr}. \tag{57}$$

By equation (55)

$$g(p, z_{22}) = M^{-1}\ (-1)^{r-1} a_o \Omega^{r\dagger} + (-1)^{r-2} a_1 \Omega^{r-1\dagger} + \cdots + a_r 1_{nr}\ M,$$

and by equation (57) and Theorem 3

$$g(p, z_{22}) = \pm M^{-1}[g(-p, \Omega^\dagger)]M \equiv 0_k. \tag{58}$$

From the last equation in equation (15), from equation (58), and from equation (10), which holds for $0 \leqq l \leqq 2r - 2$,

$$\begin{aligned} a_0 A_{2r-1} &= -a_1 A_{2r-2} - a_2 A_{2r-3} - \cdots - a_r A_{r-1} \\ &= -z_{12}[a_1 z_{22}^{2r-2} + a_2 z_{22}^{2r-2} + \cdots + a_r z_{22}^{r-1}]\mathbf{z}_{12}^\dagger \\ &= -a_0 z_{12} z_{22}^{2r-1} \mathbf{z}_{12}^\dagger . \end{aligned}$$

Hence

$$A_{2r-1} = (-1)^{2r-1} z_{12} z_{22}^{2r-1} \mathbf{z}_{12}^{\dagger} ,$$

and by induction

$$A_{l} = (-1)^{l} z_{12} z_{22}^{l} \mathbf{z}_{12}^{\dagger}$$

for all $l \geqq 0$, which is the same as equation (10).

We thus have a set of three matrices z_{11} , z_{12} , z_{22} such that the infinite set of equations obtained by equating the right sides of equations (7) and (8) are satisfied. Hence the right side of equation (4) and $W(p, s)$ have the same Taylor's series expansion in the neighborhood of $s = \infty$. By analytic continuation, for all p and s

$$W(p, s) = z_{11}(p) + z_{12}(p)[z_{22}(p) + s1_{k}]^{-1} \mathbf{z}_{12}^{\dagger}(p),$$

where z_{11} , z_{12} , and z_{22} are defined by equations (9), (47), and (50), respectively.

We have thus succeeded in decomposing $W(p, s)$ as shown in equation (4). It now remains to show that $Z(p)$ formed from the chosen z_{11} , z_{12} , and z_{22}

$$\begin{aligned} Z(p) &= \begin{bmatrix} z_{11}(p) & z_{12}(p) \\ -\mathbf{z}_{12}^{\dagger}(p) & z_{22}(p) \end{bmatrix} \\ &= \begin{bmatrix} W(p, \infty) & \dfrac{M_{0}(p)}{a_{0}^{r}(p)} \\ -\dfrac{\mathbf{M}_{0}^{\dagger}(p)}{a_{0}^{r\dagger}(p)} & a_{0}^{2r}(p) M^{-1}(p) T_{d}(p) \mathbf{M}^{-1\dagger}(p) \end{bmatrix} \end{aligned} \tag{59}$$

is a reactance matrix.

To show that the $Z(p)$ in equation (59) is a reactance matrix, we may choose any standard test, but we will choose the one given below since it is particularly suited for the problem at hand (see pp. 117 and 123 of Ref. 11):

Lemma 3: The necessary and sufficient conditions for a square matrix $Z(p)$ to be a reactance matrix are:

(*i*) *Z is rational and real for real p.*

(*ii*) *Poles of $Z(p)$ are simple and restricted to the imaginary axis.*

(*iii*) $Z + \mathbf{Z}^{\dagger} \equiv 0$.

(*iv*) *Residue matrices are positive semidefinite Hermitian.*

Since all the entries of $Z(p)$ in equation (59) are real and rational, condition i is satisfied.

From equations (13a) and (15), $z_{11} = A_{-1} = B_0/a_0$ is a reactance matrix by Theorem 3; hence its poles are simple and restricted to the imaginary axis. Also, the pole at $p = \infty$, if any exists, is simple for this block. Since M_0 is a polynomial matrix, it is clear that the poles of the off diagonal blocks z_{12} and $-\mathbf{z}_{12}^\dagger$ are in the zeros of a_0 and hence by Theorem 3 the poles of z_{12} are restricted to the $j\omega$ axis. However, it is not clear that these poles are simple. To show that they are indeed simple we will use the fact that $\hat{A}_0$ defined by

$$\hat{A}_0 = a_0^2 A_0 \tag{60}$$

is a polynomial matrix. From equations (10) and (47)

$$A_0 = \frac{M_0 \mathbf{M}_0^\dagger}{a_0^{2r}} \tag{61}$$

and from equation (39) $a_0 = \pm a_0^\dagger$. We first consider $a_0 = a_0^\dagger$ in which case

$$\hat{A}_0 = \left[\frac{M_0}{a_0^{r-1}}\right]\left[\frac{\mathbf{M}_0}{a_0^{r-1}}\right]^\dagger . \tag{62}$$

Equation (54) then shows that $\hat{A}_0 = \hat{\mathbf{A}}_0^\dagger$ and $\hat{A}_0(j\omega) \geqq 0$. Hence there exists an $n \times q$ polynomial matrix, Q, such that

$$\hat{A}_0 = Q\mathbf{Q}^\dagger \tag{63}$$

where q is the rank of $\hat{A}_0$. Equations (62) and (63) are two different factorizations of $\hat{A}_0$, hence:[17]

$$\frac{M_0}{a_0^{r-1}} = Q[1_q \vdots 0_{k\times(r-q)}]V \tag{64}$$

where $V(p)$ is a $k \times k$ para unitary matrix, that is, $V\mathbf{V}^\dagger = 1_k$. Since Q is a polynomial matrix, and $V(p)$ being para unitary can have no poles on the imaginary axis (see p. 186 of Ref. 11), the left side of equation (64) can have no poles on the imaginary axis. Hence a_0^{r-1}, which has all its zeros on the $j\omega$ axis, must divide M_0. Thus z_{12} has all its finite poles in the zeros of a_0. By Theorem 3, the zeros of a_0 are simple and restricted to the $j\omega$ axis. In the above, we have assumed that $a_0 = a_0^\dagger$; if $a_0 = -a_0^\dagger$ and r is odd, the same proof holds; if r is even we can construct a similar proof by factoring $-\hat{A}_0$ instead of $\hat{A}_0$.

To show that the pole of z_{12} at $p = \infty$, if any, is simple. Consider the following representation for A_0 obtained from equations (15) and (17)

$$A_0 = \frac{a_0 B_1 - a_1 B_0}{a_0^2} = \frac{B_1}{a_1}\cdot\frac{a_1}{a_0} - \frac{B_0}{a_0}\cdot\frac{a_1}{a_0}. \tag{65}$$

Since B_1/a_1 and B_0/a_0 are reactance matrices and a_1/a_0 is a reactance function, according to Theorem 3, the right side of equation (65) behaves as Kp^ν near $p = \infty$, where K is a constant matrix and ν is an integer such that $-2 \leqq \nu \leqq 2$. But z_{12} satisfies

$$A_0 = z_{12}\mathbf{z}_{12}^{\dagger}$$

and hence the pole of z_{12} at $p = \infty$, if any, must be simple.

We now have to show that the z_{22} block also satisfies condition *(ii)* of the lemma. By equation (54)

$$z_{22} = -\mathbf{M}^{\dagger}\mathbf{\Omega}\mathbf{M}^{\dagger^{-1}} = M^{-1}\Omega^{\dagger}M. \tag{54}$$

Since M^{-1} is analytic in the open right-half plane and Ω has all its poles in the zeros of a_0, by equation (54) the poles of z_{22} are restricted to the $j\omega$ axis. To show that these poles are simple we will prove by contradiction that $a_0 z_{22}$ is polynomial.

From equation (52), the definition of Ω, and equation (54) we see that if $a_0 z_{22}$ has a pole of multiplicity α at $p = j\omega_0$. In the neighborhood of this pole, we have the approximation

$$a_0 z_{22} \approx \frac{K}{(p - j\omega_0)^{\alpha}} \tag{66}$$

where K is a constant matrix and α is a positive integer, and

$$a_0^2 z_{22}^2 \approx \frac{K^2}{(p - j\omega_0)^{2\alpha}}. \tag{67a}$$

Now by equation (55) $z_{22}^2 = -M^{-1}\Omega^{2\dagger}M$, and hence in the neighborhood of $p = j\omega_0$

$$a_0^2 z_{22}^2 \approx \frac{K_1}{(p - j\omega_0)^{\beta}} \tag{67b}$$

where K_1 is a constant matrix and β is a positive integer. Since the poles of $a_0 z_{22}$ are contained in the poles of M^{-1}, $\beta \leqq 2\alpha$. By comparison of equations (67a) and (67b), which must be equal, it is clear that either $\alpha = \beta = 0$ or $K_1 = K^2 = 0$. Since $z_{22} = -\mathbf{z}_{22}^{\dagger}$, $K = \mathbf{K}^*$, and hence $K^2 = K\mathbf{K}^* = 0$ implies that $K = 0$. Thus $a_0 z_{22}$ can have no poles on the $j\omega$ axis and this, coupled with the fact that z_{22} can have poles only on the $j\omega$ axis, guarantees that $a_0 z_{22}$ is always polynomial. We therefore conclude that all the finite poles of z_{22} are in the zeros of a_0, and their multiplicity cannot exceed that of the corresponding zeros of a_0. Hence, again by Theorem 3, all the finite poles of z_{22} are simple and restricted to the $j\omega$ axis.

To show that the pole at $p = \infty$ of z_{22}, if any, is simple, consider equation (15) written in this form:

$$\begin{aligned} A_{-1} &= \frac{B_0}{a_0} \\ A_0 &= \frac{B_1}{a_1}\cdot\frac{a_1}{a_0} - \frac{B_0}{a_0}\cdot\frac{a_1}{a_0} \\ A_1 &= \frac{B_2}{a_2} - \frac{a_2}{a_1}\cdot\frac{a_1}{a_0} - \frac{a_1}{a_0}\left[\frac{B_1}{a_1}\cdot\frac{a_1}{a_0} - \frac{B_0}{a_0}\cdot\frac{a_1}{a_0}\right] - \frac{a_2}{a_1}\cdot\frac{a_1}{a_0}\cdot\frac{B_0}{a_0}. \\ &\vdots \end{aligned} \tag{68}$$

Owing to the reactance nature of B_i/a_i and a_i/a_{i+1} by Theorem 3, and from the form of A_i shown in equation (68), near $p = \infty$, A_i behaves as

$$A_i \approx K_i p^{\nu_i} \tag{69}$$

where K_i is a constant matrix and ν_i is an integer such that

$$i + 2 \geqq \nu_i \geqq -(i + 2). \tag{70}$$

Also from equation (10)

$$z_{12} z_{22}^{i} \mathbf{z}_{12}^{\dagger} = (-1)^i A_i \approx \pm K_i p^{\nu_i}. \tag{71}$$

Since z_{12} has at most a simple pole at $p = \infty$, in the neighborhood of $p = \infty$

$$z_{12} \approx K p^{l} \tag{72}$$

where K is a constant matrix and l is an integer such that $l \leqq 1$. If z_{22} behaves as $K_{22}p^m$ near $p = \infty$, where K_{22} is a constant matrix and m an integer, then by equation (70), (71), and (72), $(i + 2) \geqq im + 2l \geqq -(i + 2)$. For such to be true for any fixed l and all integral $i \geqq 0$, m has to be less than or equal to unity. Hence the pole of z_{22} at $p = \infty$, if any, is simple.

We have thus shown that condition *ii* of Lemma 1 is satisfied for each block in $Z(p)$, and hence $Z(p)$ also satisfies it.

Since z_{11} is a reactance matrix, $z_{11} = -\mathbf{z}_{11}^{\dagger}$ and $z_{22} = -\mathbf{z}_{22}^{\dagger}$ by equation (51), we have $Z = -\mathbf{Z}^{\dagger}$ and thus condition (*iii*) of the lemma is also satisfied.

Now to complete the proof that $Z(p)$ is a reactance matrix, we have to show that the residue matrices at the poles are positive semidefinite Hermitian. To do this we need Lemma 4, which follows from the definitions of a two variable positive real and two variable reactance matrices (see p. 34 of Ref. 8).

Lemma 4. If $W(p, s)$ is a two variable reactance matrix with no p-independent or s-independent poles, $W[p, s(p)]$ is a reactance matrix in p for any reactance function $s(p)$.

To prove that $Z(p)$ satisfies condition *iv* of Lemma 1, which requires that the residue matrix of $Z(p)$ at any of its simple poles on the $j\omega$ axis is positive semidefinite Hermitian, we note that at any pole, $p = j\omega$, of $Z(p)$, if we set

$$\begin{aligned} s(p) &= \frac{2lp}{p^2 + \omega_0^2} \qquad \text{for} \quad |\,\omega_0\,| < \infty \\ &= lp \qquad\qquad \text{for} \quad \omega_0 = \infty \end{aligned}$$

in

$$W(p, s) = z_{11} + z_{12}(z_{22} + sl_k)^{-1}\mathbf{z}_{12}^{\dagger}$$

[which is equation (4)] then by Lemma 4, $W[p, s(p)]$ is a reactance matrix in p for all positive l. Since $Z(p)$ is real for real p and $Z = -\mathbf{Z}^{\dagger}$, the residue matrix H at the pole $p = j\omega$ is Hermitian; if we write it as

$$H = \begin{bmatrix} H_{11} & H_{12} \\ \mathbf{H}_{12}^{*} & H_{22} \end{bmatrix} \tag{73}$$

then, K, the residue matrix of $W[p, s(p)]$ at $p = j\omega_0$ is given by

$$K = H_{11} - H_{12}(H_{22} + l1_k)^{-1}\mathbf{H}_{12}^{*}\,. \tag{74}$$

Since H, H_{11}, and H_{22} are Hermitian, there exist unitary matrices U_1 and U_2 such that

$$\Lambda_{11} = \mathbf{U}_1^{*}H_{11}U_1 = \operatorname{diag}\,[d_1, d_2, \cdots, d_n] \tag{75}$$

and

$$\Lambda_{22} = \mathbf{U}_2^{*}H_{22}U_2 = \operatorname{diag}\,[\lambda_1, \lambda_2, \cdots, \lambda_k]. \tag{76}$$

Hence

$$\mathbf{U}_1^{*}KU_1 = \Lambda_{11} - J_{12}(\Lambda_{22} + l1_k)^{-1}\mathbf{J}_{12}^{*} \tag{77}$$

where

$$J_{12} = \mathbf{U}_1^{*}H_{12}U_2\,. \tag{78}$$

If J_{12i} denotes the ith column of J_{12}, the right side of equation (77) can be rewritten as

$$\mathbf{U}_1^{*}KU_1 = \Lambda_{11} - \sum_{i=1}^{k} \frac{1}{\lambda_i + l} J_{12i}\mathbf{J}_{12i}^{*}\,. \tag{79}$$

Since K is the residue matrix of a reactance matrix, for all $l > 0$, K is positive semidefinite. Λ_{11} is also positive semidefinite, since H_{11} is the residue matrix of the reactance matrix z_{11}. $J_{12i}\mathbf{J}^*_{12i}$ is obviously positive semidefinite and the left side of (79) can be positive semidefinite for all positive l only if all the λ_i are nonnegative. Hence Λ_{22} and H_{22} are positive semidefinite.

To show that H is positive semidefinite, we will show that H' defined by

$$H' = (\mathbf{U}^*_1 \dotplus \mathbf{U}^*_2)H(U_1 \dotplus U_2) = \begin{bmatrix} \Lambda_{11} & J_{12} \\ \mathbf{J}^*_{12} & \Lambda_{22} \end{bmatrix} \tag{80}$$

is positive semidefinite. For this purpose, consider the Hermitian form

$$[\mathbf{X}^*_1 \ \mathbf{X}^*_2] \begin{bmatrix} \Lambda_{11} & J_{12} \\ \mathbf{J}^*_{12} & D_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

$$= \mathbf{X}^*_1\Lambda_{11}X_1 + \mathbf{X}^*_2\mathbf{J}^*_{12}X_1 + \mathbf{X}^*_1J_{12}X_2 + \mathbf{X}^*_2D_{22}X_2 \tag{81}$$

where

$$D_{22} = \Lambda_{22} + l1_k \,; \qquad l > 0.$$

Since

$$\mathbf{U}^*_1KU_1 = \Lambda_{11} - J_{12}D^{-1}_{22}\mathbf{J}^*_{12}$$

is positive semidefinite, we obtain from equation (81) the following inequality:

$$[\mathbf{X}^*_1 \ \mathbf{X}^*_2] \begin{bmatrix} \Lambda_{11} & J_{12} \\ \mathbf{J}^*_{12} & D_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

$$\geqq \mathbf{X}^*_1J_{12}D^{-1}_{22}\mathbf{J}^*_{12}X_1 + \mathbf{X}^*_1J_{12}X_2 + \mathbf{X}^*_2\mathbf{J}^*_{12}X_1 + \mathbf{X}^*_2\Lambda_{22}X_2 \,. \tag{82}$$

Since the right side of equation (82) can be expressed as $\mathbf{G}^*G$, where $G = [D^{1/2}_{22}J^*_{12}X_1 + \Lambda^{1/2}_{22}X_2]$, the Hermitian form in equation (81) is positive semidefinite for all $l > 0$; by a continuity argument we can see that H' and consequently H are positive semidefinite.

We have thus shown that $Z(p)$ does indeed describe a lossless network in the p-plane and thus $W(p, s)$ has the network representation shown in Fig. 2.

In the development of the synthesis procedure we assumed that $a^{2r}_0(j\omega)T_{r-1}(j\omega) = \hat{T}_{r-1}(j\omega) \geqq 0$. If simultaneously, $a_0(p)$ is an odd function of p [in other words $\nu = 1$ in equation (39)] and r, the s-degree of the least common denominator of $W(p, s)$ is odd, then $\hat{T}_{r-1}(j\omega) \leqq 0$. In this case we factor $-\hat{T}_{r-1}(p)$ which is para Hermitian and positive semidefinite on the $j\omega$ axis. We will then have

$$-\hat{T}_{r-1} = M\,\mathbf{M}^{\dagger} \tag{83}$$

and hence, as before equation (44),

$$T_{r-1} = \frac{M}{a_0^r}\,\frac{\mathbf{M}^{\dagger}}{a_0^{r\dagger}}.$$

It is then clear that the identification of z_{12} and z_{22} can be done in exactly the same way as when $\hat{T}_{r-1}(j\omega) \geqq 0$.

It is of importance to notice that the number of s-plane inductors used in the realization of Fig. 2 is equal to the s-degree, $\delta_s[W(p, s)]$ which in general is smaller than the number required in Koga's technique. Appendix C shows that $\delta_s[W(p, s)]$ is the minimum number of s-plane inductors required in any realization, and that if a realization is minimal in the variable s it is automatically minimal in the variable p, the minimum number of p-type reactances needed in any realization being the p-degree, $\delta_p[W(p, s)]$.[18]

The main result of this section can be conveniently put in the form of a theorem:

Theorem 4: Every two variable reactance matrix $W(p, s)$ can be realized as the impedance seen at the first n-ports of a lossless $(n + k)$-port consisting of $\delta_p[W(p, s)]$ reactances in the p-plane, terminated at its last k ports with $\delta_s[W(p, s)]$ unit inductors in the s-plane. Furthermore, such a realization uses the minimum possible number of reactances of each kind. (The roles of p and s are completely interchangeable.)

Since several of the proofs involved in establishing Theorem 4 were rather indirect and lengthy, while the procedure for synthesis, summarized in Section IV, is itself rather simple.

IV. SUMMARY OF SYNTHESIS PROCEDURE

Given an $(n \times n)$ two variable reactance matrix $W_o(p, s)$, decompose it as

$$W_0(p, s) = W_1(p) + W_2(s) + W(p, s)$$

where W_1 and W_2 are reactance matrices in p and s, and $W(p, s)$ is a two variable reactance matrix with no p-independent or s-independent poles. Such a decomposition is always possible by Theorem 2.

Expand $W(p, s)$ as

$$W(p, s) = A_{-1}(p) + \sum_{l=0}^{\infty} \frac{A_l(p)}{s^{l+1}}$$

[which is the same as equation (7)] where the $A(p)$'s may be obtained by equations (16) or (16a) or by long division.

Find $g(p, s)$, the least common denominator of the entries in $W(p, s)$ and express it in the form

$$g(p, s) = a_0(p)s^r + a_1(p)s^{r-1} + \cdots + a_0(p).$$

[which is the same as equation (6)].

Form the $(nr \times nr)$ matrix $T_{r-1}(p)$, defined by

$$T_{r-1}(p) = \begin{bmatrix} A_0(p) & A_1(p) & A_2(p) & \cdots & A_{r-1}(p) \\ -A_1(p) & -A_2(p) & -A_3(p) & \cdots & -A_r(p) \\ A_2(p) & A_3(p) & A_4(p) & \cdots & A_{r+1}(p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ (-1)^{r-1}A_{r-1}(p) & (-1)^{r-1}A_r(p) & (-1)^{r-1}A_{r+1}(p) & \cdots & (-1)^{r-1}A_{2r-2}(p) \end{bmatrix}$$

which is equation (24).

Factor $\hat{T}_{r-1}(p) = a_0^{2r}T_{r-1}(p)$, a polynomial matrix, as

$$\hat{T}_{r-1}(p) = M\mathbf{M}^{\dagger} \tag{43}$$

unless simultaneously, $a_0 = -a_0^{\dagger}$ and r in equation (6) is odd, in which case factor $-\hat{T}_{r-1}(p)$. The factorization must be such that M is a $(k \times nr)$ polynomial matrix with k = rank of $T_{r-1}(p)$ and M^{-1}, the left inverse of M analytic in the open right plane. The existence of such a factorization is guaranteed by Lemmas 1 and 2.

Partition $M(p)$ into $(n \times k)$ blocks of equation (45)

$$M(p) = \begin{bmatrix} M_0(p) \\ \hdashline M_1(p) \\ \hdashline \vdots \\ \hdashline M_{r-1}(p) \end{bmatrix}.$$

Form the $(nr \times nr)$ matrix $\Omega(p)$ defined by equation (52)

With the identification of equations (9), (45), and (54)[*]

$$z_{11}(p) = A_{-1}(p),$$

$$\text{and} \qquad z_{12} = \frac{M_0(p)}{a_0(p)},$$

$$\text{and} \qquad z_{22} = M^{-1}(p)\Omega^{\dagger}(p)M(p),$$

the decomposition

$$W(p, s) = z_{11}(p) + z_{12}(p)[z_{22}(p) + s\mathbf{1}_k]^{-1}\mathbf{z}_{12}^{\dagger} \tag{4}$$

is obtained. Notice that this is equation (4). It should also be noticed that $W(p, s)$ can be decomposed as in equation (4) even if it has s-independent poles, since the assumption that $W(p, \infty)$ is finite is enough to guarantee the validity of the procedure. For network realization it is usually more convenient to remove both p-independent and s-independent poles; we therefore removed them at the start of the procedure.

To realize $W(p, s)$ as the impedance of a passive network, we perform the following operations.

Form the $(n+k \times n+k)$ impedance matrix $Z(p)$ of the coupling network

$$Z(p) = \begin{bmatrix} A_{-1}(p) & \dfrac{M_0(p)}{a_0(p)} \\[2ex] -\dfrac{\mathbf{M}_0^{\dagger}(p)}{a_0^{\dagger}(p)} & M^{-1}(p)\Omega^{\dagger}(p)M(p) \end{bmatrix}. \tag*{(84)†}$$

Realize $Z(p)$ as a lossless $(n+k)$ port network in the p-plane and terminate its last k-ports with unit inductors in the s-plane. Also realize the reactance matrices $W_1(p)$ and $W_2(s)$ as lossless p-plane and s-plane n-ports, and connect all three networks in series as shown in Fig. 1. The given $W_o(p, s)$ is thus realized as a passive network.

V. AN EXAMPLE

It is desired to synthesize the two variable reactance matrix[‡]

[*] Equation (54) is used to determine $z_{22}(p)$, in preference to equation (50) since equation (54) is easier to compute.

[†] Equation (84) is the same as equation (59) except that for the z_{22} block equation (54) is used instead of equation (50) for the reason mentioned in the previous note.

[‡] This example was given by Koga, (see p. 50 of Ref. 8).

$$W_0(p,\, s) = \begin{bmatrix} \dfrac{(p^2 + 1)(s^2 + 1)}{(p + s)(ps + 1)} & \dfrac{ps - 1}{p + s} \\[2ex] \dfrac{ps - 1}{p + s} & \dfrac{ps + 1}{p + s} \end{bmatrix}.$$

Since $W_o(p, s)$ has no p-independent or s-independent poles the first step 1 of Section IV need not be performed, and $W_o(p, s) = W(p, s)$. The least common denominator of the elements of $W(p, s)$ is

$$g(p,\, s) = ps^2 + (p^2 + 1)s + p$$

[which is equation (6)], and hence

$$a_0(p) = p, \quad a_1(p) = (p^2 + 1),\, a_2(p) = p, \quad \text{and} \quad r = 2.$$

The least common denominator of the minors of $W(p, s)$ is also $g(p, s)$ and hence

$$k = \delta_s[W(p,\, s)] = 2.$$

In the expansion, equation (7),

$$W(p,\, s) = A_{-1}(p) + \sum_{l=0}^{\infty} \frac{A_l(p)}{s^{l+1}}$$

by the formula of equation (16) or by long division

$$A_{-1}(p) = \frac{1}{p}\begin{bmatrix} p^2 + 1 & p^2 \\ p^2 & p^2 \end{bmatrix},$$

$$A_0(p) = -\frac{1}{p^2}\begin{bmatrix} (p^2 + 1)^2 & p^2(p^2 + 1) \\ p^2(p^2 + 1) & p^2(p^2 - 1) \end{bmatrix},$$

$$A_1(p) = \frac{1}{p^3}\begin{bmatrix} (p^2 + 1)^3 & p^4(p^2 + 1) \\ p^4(p^2 + 1) & p^4(p^2 - 1) \end{bmatrix},$$

$$A_2(p) = -\frac{1}{p^4}\begin{bmatrix} p^8 + 3p^6 + 4p^4 + 3p^2 + 1 & p^6(p^2 + 1) \\ p^6(p^2 + 1) & p^6(p^2 - 1) \end{bmatrix}.$$

$T_{r-1}(p) = T_1(p)$ defined by equation (24) is

$$T_{r-1}(p) = \frac{1}{p^4}\left[\begin{array}{cc:cc} -p^2(p^2+1)^2 & -p^4(p^2+1) & p(p^2+1)^3 & p^5(p^2+1) \\ -p^4(p^2+1) & -p^4(p^2-1) & p^5(p^2+1) & p^5(p^2-1) \\ \hdashline -p(p^2+1)^3 & -p^5(p^2+1) & p^8+3p^6+4p^4+3p^2+1 & p^6(p^2+1) \\ -p^5(p^2+1) & -p^5(p^2-1) & p^6(p^2+1) & p^6(p^2-1) \end{array}\right].$$

The polynomial matrix $T_1(p) = p_0^4 T_1(p)$ is factored by the method in Ref. 16 as equation (43)

$$T_1(p) = M(p)\mathbf{M}^{\dagger}(p) = \frac{1}{(2)^{\frac{1}{2}}}\begin{bmatrix} p(p^2+1) & -p(p^2+1) \\ p^2(p-1) & -p^2(p+1) \\ p^4-p^3+2p^2-p+1 & -(p^4+p^3+2p^2+p+1) \\ p^4-p^3 & -(p^4+p^3) \end{bmatrix}$$

$$\times \frac{1}{(2)^{\frac{1}{2}}}\begin{bmatrix} -p(p^2+1) & -p^2(p+1) & p^4+p^3+2p^2+p+1 & p^4+p^3 \\ p(p^2+1) & p^2(p-1) & -(p^4-p^3+2p^2-p+1) & -(p^4-p^3) \end{bmatrix}.$$

The (4×2) matrix $M(p)$ is partitioned as equation (45)

$$M(p) = \left[\begin{array}{c} M_0(p) \\ \hdashline M_1(p) \end{array}\right]$$

$$= \frac{1}{2}\left[\begin{array}{cc} p^2(p^2+1) & -p^2(p^2+1) \\ p^2(p-1) & -p^2(p+1) \\ \hdashline p^4 - p^3 + 2p^2 - p + 1 & -(p^4 + p^3 + 2p^2 + p + 1) \\ p^4 - p^3 & -(p^4 + p^3) \end{array}\right].$$

To find $M^{-1}(p)$, a left inverse of $M(p)$, it is enough to find a left inverse of M_0 if it exists, since

$$[M_0^{-1} \mid 0]\left[\begin{array}{c} M_0 \\ \hdashline M_1 \end{array}\right] = 1_k \ .$$

In our example $k = 2$ and M_o is a nonsingular matrix and hence $M^{-1}(p)$ is given by

$$M^{-1}(p) = \frac{-1}{2p^3(p^2+1)}\left[\begin{array}{cc:cc} -p^2(p+1) & p(p^2+1) & 0 & 0 \\ -p^2(p-1) & p(p^2+1) & 0 & 0 \end{array}\right].$$

From the definition of Ω, equation (52)

$$\Omega(p) = \left[\begin{array}{cc:cc} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hdashline -1 & 0 & -\dfrac{p^2+1}{p} & 0 \\ 0 & -1 & 0 & -\dfrac{p^2+1}{p} \end{array}\right].$$

Using equation (54)

$$z_{22} = M^{-1}\Omega^{\dagger}M$$
$$= \begin{bmatrix} \dfrac{p^2+1}{2p} & -\dfrac{(p+1)^2}{2p} \\[2ex] -\dfrac{(p-1)^2}{2p} & \dfrac{p^2+1}{2p} \end{bmatrix}.$$

Hence the coupling network formed by p-type elements has the (4×4) matrix of equation (84)

$$Z(p) = \begin{bmatrix} A_{-1} & \dfrac{M_0}{a_0} \\[2ex] -\dfrac{\mathbf{M}_0^{\dagger}}{a_0^{\dagger}} & M^{-1}\Omega^{\dagger}M \end{bmatrix}$$
$$= \begin{bmatrix} \dfrac{p^2+1}{2p} & p & \dfrac{p^2+1}{(2)^{\frac{1}{2}}p} & -\dfrac{p^2+1}{(2)^{\frac{1}{2}}p} \\[2ex] p & p & \dfrac{p-1}{(2)^{\frac{1}{2}}} & \dfrac{p+1}{(2)^{\frac{1}{2}}} \\[2ex] \dfrac{p^2+1}{(2)^{\frac{1}{2}}p} & \dfrac{p+1}{(2)^{\frac{1}{2}}} & \dfrac{p^2+1}{2p} & -\dfrac{(p+1)^2}{2p} \\[2ex] -\dfrac{p^2+1}{(2)^{\frac{1}{2}}p} & -\dfrac{p-1}{(2)^{\frac{1}{2}}} & -\dfrac{(p-1)^2}{2p} & \dfrac{p^2+1}{2p} \end{bmatrix}$$

$Z(p)$ can be verified to be lossless, and the given $W_o(p, s)$ can of course be realized as the impedance seen at the first two ports of $Z(p)$ when it is terminated at its last two ports by unit s-plane inductors.

VI. CONCLUSIONS

The synthesis method for two-variable reactance matrices developed here, in general yields a nonreciprocal coupling network even when the given two-variable reactance matrix is symmetric, and if a reciprocal coupling network is desired, Koga's method for generating a reciprocal network from the nonreciprocal one can be used.[8] This procedure generally yields a reciprocal network at the cost of increased numbers of elements of both kinds.

This method of synthesis of two-variable reactance matrices has been successfully applied to the synthesis of lumped-distributed RC networks which are important in microelectronics circuits.[7] In practice, the only laborious step in the synthesis procedure is the factorization of polynomial matrix in the desired form. Of great importance is the approximation of desired characteristics by rational functions in two-variables; any work in this area would greatly enhance the usefulness of the two-variable theory. The synthesis problem of n-variable positive real functions, for which many applications can be found,[7] can be reduced to the synthesis of $(n+1)$-variable reactance matrices.[21, 22] when $n = 1$ the two-variable method developed here gives rise to a new method of passive RLC synthesis, which is no more complex than the existing methods.

VII. ACKNOWLEDGEMENTS

The author expresses his deep gratitude to Professor R. W. Newcomb under whose guidance this work was carried out.

APPENDIX A

Partial Fraction Expansion of $W(j\omega, s)$

To guarantee the factorization of $T_{r-1}(j\omega)$ as $M\mathbf{M}^\dagger$ we needed Lemma 2, which asserts that T_{r-1} is para Hermitian and that $T_{r-1}(j\omega) \geqq 0$. In the proof of Lemma 2 we used the fact that $R_i(j\omega)$, the residue matrices of $W(j\omega, s)$, are positive semidefinite. The proof is given below.

Under the assumption that $W(p, s)$ has no p-independent of s-independent poles, for each real ω the s-plane poles of $W(j\omega, s)$ are simple and restricted to the imaginary s-axis by Theorem 1. Hence, for any fixed ω, we can write $W(j\omega, s)$ as

$$W(j\omega, s) = A_{-1}(j\omega) + \sum_{i=1}^{r} \frac{R_i(\omega)}{s - j\alpha_i(\omega)} \tag{85}$$

where the $\alpha_i(\omega)$ are real and the $R_i(\omega)$ are the residue matrices at the poles $j\alpha_i(\omega)$. As in equation (9), r is the s-degree of $g(p, s)$, the least common denominator of the elements of W.

By complex conjugation on both sides of equation (85)

$$W^*(j\omega, s) = A_{-1}^*(j\omega) + \sum_{i=1}^{r} \frac{R_i^*(\omega)}{s^* + j\alpha_i(\omega)}. \tag{86}$$

Since W and A_{-1} are rational in $j\omega$,

$$W^*(j\omega, s) = W(-j\omega, s^*)$$

and

$$A_{-1}^{*}(j\omega) = A_{-1}(-j\omega).$$

Hence equation (86) becomes

$$W(-j\omega, s^*) = A_{-1}(-j\omega) + \sum_{i=1}^{r} \frac{R_i^*(\omega)}{s^* + j\alpha_i(\omega)} \tag{87}$$

and

$$-\mathbf{W}(-j\omega, -s^*) = -\mathbf{A}_{-1}(-j\omega) + \sum_{i=1}^{r} \frac{\mathbf{R}_i^*(\omega)}{s^* - j\alpha_i(\omega)}. \tag{88}$$

Since equation (88) is an identity for s^*, we have

$$-\mathbf{W}(-j\omega, -s) = -\mathbf{A}_{-1}(-j\omega) + \sum_{i=1}^{r} \frac{\mathbf{R}_i^*(\omega)}{s - j\alpha_i(\omega)}. \tag{89}$$

From the definition of a two variable reactance matrix,

$$W(j\omega, s) = -\mathbf{W}(-j\omega, -s)$$

and by equation (19)

$$A_{-1}(j\omega) = -\mathbf{A}_{-1}(-j\omega).$$

Hence, by comparison of equations (85) and (89) we have the desired result

$$R_i(\omega) = \mathbf{R}_i^*(\omega). \tag{90}$$

To show that the $R_i(\omega)$ are positive semidefinite for each ω, we first notice that if

$$W(p, s) = \frac{\psi(p, s)}{g(p, s)}$$

where $\psi(p, s)$ is a polynomial matrix and $g(p, s)$ is the least common denominator of the entries in W, $R_i(\omega)$ in equation (85) is given by (see p. 308 of Ref. 19)

$$R_i(\omega) = \left. \frac{\psi(p, s)}{\dfrac{\partial g(p, s)}{\partial s}} \right|_{\substack{p = j\omega \\ s = j\alpha_i(\omega)}} . \tag{91}$$

Denoting $\partial g/\partial s$ by g_s and $\partial \psi/\partial s$ by ψ_s, for any $n \times 1$ constant matrix X, (p. 39 of Ref. 8)

$$\begin{aligned}\mathbf{X}^*R_iX &= \frac{\mathbf{X}^*\psi X}{g_s}\bigg|_{\substack{p=j\omega\\ s=j\alpha_i(\omega)}}\\ &= \left[\frac{(\mathbf{X}^*\psi X)g_s - g(\mathbf{X}^*\psi_s X)}{(\mathbf{X}^*\psi X)^2}\right]^{-1}_{\substack{p=j\omega\\ s=j\alpha_i(\omega)}}\\ &= \left[\frac{\partial}{\partial s}\left(\frac{g}{\mathbf{X}^*\psi X}\right)\right]^{-1}_{\substack{p=j\omega\\ s=j\alpha_i(\omega)}}.\end{aligned}$$

Hence, if $X^*WX \not\equiv 0$

$$\mathbf{X}^*R_i(\omega)X = \left[\frac{\partial}{\partial s}(\mathbf{X}^*WX)^{-1}\right]^{-1}_{\substack{p=j\omega\\ s=j\alpha_i(\omega)}}. \qquad (92)$$

From definitions 1 and 2, and Theorem 1, $\mathbf{X}^*WX$ is a two variable positive function and for $\operatorname{Re} p = \operatorname{Re} s = 0$

$$\operatorname{Re}\,[\mathbf{X}^*WX] \equiv 0$$

and

$$\frac{\partial}{\partial s}(\mathbf{X}^*WX) \geqq 0.$$

Hence

$$\frac{\partial}{\partial s}(\mathbf{X}^*WX)^{-1} = -\frac{1}{(\mathbf{X}^*WX)^2}\cdot\frac{\partial}{\partial s}(\mathbf{X}^*WX) \geqq 0$$

for

$$\operatorname{Re} p = \operatorname{Re} s = 0$$

and consequently the left side of equation (90) is nonnegative.

Thus we have proved that the residue matrices, $R_i(\omega)$, are positive semidefinite Hermitian for each ω.

APPENDIX B

Proof of Theorem 3

Theorem 3: If

$$W(p, s) = \frac{B_0(p)s^r + B_1(p)s^{r-1} + \cdots + B_r(p)}{a_0(p)s^r + a_1(p)s^{r-1} + \cdots + a_r(p)}$$

is a two variable reactance matrix, then for all $i = 0, 1, \cdots r$

B_i/a_i is a reactance matrix in p.

a_i/a_{i+1} is a reactance function in p.

a_i has all its zeros on the j axis, and these are simple.

$\mathbf{X}B_iX/\mathbf{X}B_{i+1}X$ for all constant $n \times 1$ vectors, X, is a reactance function in p.

Proof: For any constant $n \times 1$ matrix X,

$$\mathbf{X}^*WX = \frac{(\mathbf{X}^*B_0X)s^r + (\mathbf{X}^*B_ix)s^{r-1} + \cdots + (\mathbf{X}^*B_rX)}{a_0s^r + a_1s^{r-1} + \cdots + a_r} \tag{93}$$

is a rational function in p and s with possible complex coefficients. For convenience, if we define

$$\begin{aligned} b_i &= \mathbf{X}^*B_iX \\ w(p, s) &= \mathbf{X}^*WX \\ f(p, s) &= b_0s^r + b_is^{r-1} + \cdots + b_r \end{aligned}$$

and as before

$$g(p, s) = a_0s^r + a_1s^{r-1} + \cdots + a_r$$

equation (93) can be written as

$$w(p, s) = \frac{f(p, s)}{g(p, s)}. \tag{94}$$

From the definition of a two variable reactance matrix, $w(p, s)$ is a two variable positive function, and hence for any p_0 with Re $p_0 > 0$, $w(p_0, s)$ is a positive function of s.[20] Consequently, for all s with Re $s > 0$

$$\text{Re}\,\frac{f(p_0, s)}{g(p_0, s)} \geqq 0. \tag{95}$$

Since equation (95) has to be satisfied for all s with Re $s > 0$ and hence for arbitrarily small s, it can be seen from equation (93) that

$$\text{Re}\,\frac{b_r(p_0)}{a_r(p_0)} \geqq 0$$

for all p_0 with Re $p_0 > 0$. Hence, $B_r(p)/a_r(p)$ is a positive real matrix and since $W = -\mathbf{W}^\dagger$

$$\left[\frac{B_r}{a_r}\right] = -\left[\frac{\mathbf{B}_r}{a_r}\right]^\dagger$$

and thus B_r/a_r is a reactance matrix in p.

If instead of starting from the positive function $f(p_0, s)/g(p_0, s)$, we start from $\partial^\nu f(p_0, s)/\partial s^\nu/\partial^\nu g(p_0, s)/\partial s^\nu$, which is also a positive function[20] for all $0 \leqq \nu \leqq r$, the same arguments used in proving that B_r/a_i is a reactance matrix can be repeated to show that B_i/a_i is a reactance matrix in p for all $0 \leqq i \leqq r$.

Now to show that a_{i+1}/a_i is a reactance function in p, we can use a similar proof based on the fact that $\partial^{\nu-1}g(p_0, s)/\partial s^{\nu-1}/\partial^\nu g(p_0, s)/\partial s^\nu$ is a positive function.[20]

Again, it can be seen from the fact that

$$\partial^{\nu-1}g(p_0, s)/\partial s^{\nu-1}/\partial^\nu g(p_0, s)/\partial s^\nu$$

is a positive function.[20] that b_{i+1}/b_i is a positive function satisfying

$$\left[\frac{b_{i+1}}{b_i}\right] = -\left[\frac{b_{i+1}}{b_i}\right]^\dagger.$$

If X in equation (93) is chosen real b_{i+1}/b_i will be real for real p and hence $\mathbf{X}B_{i+1}X/\mathbf{X}B_iX$ for any real $n \times 1$ matrix, X, is a reactance function.

To see that the zeros of a_i are all simple and restricted to the imaginary axis: if any one of the a_i has a double zero on the $j\omega$ axis or a zero off the $j\omega$ axis, from the reactance nature of a_{i+1}/a_i for all $0 \leqq i \leqq r$, all the a_i must have the same zero, and, consequently, $W(p, s)$ will have an s-independent pole contradicting our original assumption that W has no such poles.

We have thus proved all the claims of Theorem 3.

APPENDIX C

Proof of the Minimality of the Realization of $W(p, s)$ in Both Variables

In this appendix we show that the realization of $W(p, s)$ that Section III gives is minimal in both the p and s variables. From the definitions of $\delta_s[W(p, s)]$ and $\delta_p[W(p, s)]$, it can be shown that if $W(p, s)$ is finite at $p = \infty$ and $s = \infty$,

$$\delta_s[W(p, s)] = \delta_s[\eta(p, s)]$$

$$\delta_p[W(p, s)] = \delta_p[\eta(p, s)]$$

where the two variable real polynomial

$$\eta(p, s) = d_0(p)s^k + d_1(p)s^{k-1} + \cdots + d_k(p) \tag{96}$$

is the least common denominator of all the minors of $W(p, s)$. The form in which $\eta(p, s)$ is written in equation (96) immediately reveals that

$$\delta_s[W(p, s)] = k.$$

And if $\eta(p, s)$ is written as

$$\eta(p, s) = c_0(p)p^m + c_1(p)p^{m-1} + \cdots + c_m(p), \tag{97}$$

it can be seen that

$$\delta_p[W(p, s)] = m.$$

C.1 *Minimum Elements*

Next we would like to find the minimum number of elements of each kind needed in the realization of $W(p, s)$.

Lemma 1 states that k, the rank of $T_{r-1}(p)$, is equal to the s-degree of $W(p, s)$, and the realization obtained there uses exactly k s-type elements. By equation (4)

$$W(p, s) = z_{11}(p) + z_{12}(p)[z_{22}(p) + s\mathbf{1}_k]^{-1}\mathbf{z}_{12}^{\dagger}(p).$$

Suppose that there exists a realization with k_0 s-type elements, where $k_0 < k = \text{rank } T_{r-1}(p)$. Then,

$$W(p, s) = z_{11}(p) + \underline{z}_{12}(p)[\underline{z}_{22}(p) + s\mathbf{1}_{k_0}]^{-1}\underline{\mathbf{z}}_{12}^{\dagger}(p)$$

where the matrices $\underline{z}_{12}(p)$ and $\underline{z}_{22}(p)$ are $n \times k_0$ and $k_0 \times k_0$, respectively. Then by equation (25), $T_{r-1}(p) = N(p)\mathbf{N}^{\dagger}(p)$ where

$$N(p) = \begin{bmatrix} \underline{z}_{12}(p) \\ \underline{z}_{12}(p)\underline{z}_{22}(p) \\ \vdots \\ \underline{z}_{12}(p)\underline{z}_{22}^{r-1}(p) \end{bmatrix}$$

is an $nr \times k_0$ matrix and hence, rank $N(p) \leqq k_0$. Also, we have

$$\text{rank } T_{r-1}(p) \leqq \text{rank } N(p) \leqq k_0 < k = \text{rank } T_{r-1}(p)$$

which is a contradiction, and hence $k = \text{rank } T_{r-1}(p) = \delta_s[W(p, s)]$ is the minimum number of s-type elements required in any realization. Now by repeating the same argument with a realization of $W(p, s)$ where p-type elements are extracted instead of s-type elements, we can see that any realization must contain at least m p-type elements where $m = \delta_p[(p, s)]$.

C.2 *Minimality of the Realization in Section III*

We next discuss the minimality of the realization in both p-type and s-type elements. For the purpose of realization, the reactance matrix $W(p, s)$ was decomposed as

$$W(p, s) = z_{11}(p) + z_{12}(p)[z_{22}(p) + sl_k]^{-1}\mathbf{z}_{12}^{\dagger}(p) \tag{4}$$

where

$$Z(p) = \begin{bmatrix} z_{11}(p) & z_{12}(p) \\ -\mathbf{z}_{12}^{\dagger}(p) & z_{22}(p) \end{bmatrix} \tag{11}$$

can be realized as the impedance matrix of a lossless $(n + k)$ port in the *p-plane*. $W(p, s)$ is the impedance seen at the first n ports when the above $(n + k)$ port network is terminated with unit s-plane inductors at its last k-ports. Since k is the s-degree of $W(p, s)$, the realization uses the minimum number of s-type elements. To show that the realization uses the minimum number of p-type elements, we have to show that $\delta[Z(p)] = \delta_p[W(p, s)]$. For this we need a relationship that exists between the least common denominator of the minors of $W(p, s)$ and the determinant $|z_{22}(p) + s1_k|$.

Every minor of $[z_{22}(p) + s1_k]^{-1}$ can be expressed as $\mu(p, s)/\varphi(p, s)$ (See p. 21, of Ref. 12, Vol. 1) where $\mu(p, s)$ and $\varphi(p, s)$ are polynomials in s with coefficients from the field of rational functions in p. Furthermore,

$$\varphi(p, s) = | z_{22}(p) + s1_k |$$

is a monic polynomial in s of degree k.

Since $W(p, s)$ has no p-independent or s-independent poles, every zero of $\eta(p, s)$ is a zero of $\varphi(p, s)$, and since $k = \delta_s[\varphi(p, s)] = \delta_s[\eta(p, s)]$, $\varphi(p, s)$ and $\eta(p, s)/d_o(p)$, which are monic polynomials in s with rational functions of p as coefficients, must be identical. Hence

$$| z_{22}(p) + s1_k | = \frac{\eta(p, s)}{d_0(p)}. \tag{98}$$

To show that $\delta[Z(p)] = \delta_p[W(p, s)]$ (since we already know that $\delta[Z(p)] \geqq \delta_p[W(p, s)]$) it is sufficient to show that $\delta[Z(p)] \leqq \delta_p[W(p, s)]$. To establish this inequality, consider the matrix $S(p, s)$ defined by

$$S(p, s) = [Z(p) - s1_{n+k}][Z(p) + s1_{n+k}]^{-1}. \tag{99}$$

When $s = 1$, $S(p, s)$ is the scattering matrix of a lossless network, since $Z(p)$ describes a lossless network and (see p. 184 of Ref. 11)

$$\delta[Z(p)] = \delta[S(p, 1)]. \tag{100}$$

Since $S(p, 1)$ is para unitary (see p. 131 of Ref. 15)

$$\delta[S(p, 1)] = \delta[| S(p, 1) |]. \tag{101}$$

Equating the determinants of matrices on both sides of equation (99)

$$| S(p, s) | = \frac{| Z(p) - s\mathbf{1}_{n+k} |}{| Z(p) + s\mathbf{1}_{n+k} |}.$$

Using a formula from the theory of determinants (see p. 46 of Ref. 12, Vol. I)

$$\begin{aligned} | S(p, s) | &= \frac{| (z_{11} - s\mathbf{1}_n) + z_{12}(z_{22} - s\mathbf{1}_k)^{-1}\mathbf{z}_{12}^{\dagger} | \cdot | z_{22} - s\mathbf{1}_k |}{| (z_{11} + s\mathbf{1}_n) + z_{12}(z_{22} + s\mathbf{1}_k)^{-1}\mathbf{z}_{12}^{\dagger} | \cdot | z_{22} + s\mathbf{1}_k |} \\ &= \frac{| W(p, -s) - s\mathbf{1}_n |}{| W(p, s) + s\mathbf{1}_n |} \cdot \frac{| z_{22} - s\mathbf{1}_k |}{| z_{22} + s\mathbf{1}_k |}. \end{aligned} \tag{102}$$

Now if $|W(p, s) + s\mathbf{1}_n|$ is written as

$$| W(p, s) + s\mathbf{1}_n | = \frac{h(p, s)}{\eta(p, s)} \tag{103}$$

where $h(p, s)$ is a real polynomial in p and s, since the left side of equation (102) is finite at $p = \infty$

$$\delta_p[h(p, s)] \leqq \delta_p[\eta(p, s)] = \delta_p[W(p, s)]. \tag{104}$$

Substituting equations (98) and (103) in equation (102), we have

$$\begin{aligned} | S(p, s) | &= \frac{h(p, -s)}{\eta(p, -s)} \cdot \frac{\eta(p, s)}{h(p, s)} \cdot \frac{\eta(p, -s)}{d_0(p)} \cdot \frac{d_0(p)}{\eta(p, s)} \\ &= \frac{h(p, -s)}{h(p, s)} \end{aligned} \tag{105}$$

and by equations (100), (101), and (104)

$$\delta[Z(p)] = \delta[S(p, 1)] \leqq \delta_p[W(p, s)].$$

We have thus shown that $\delta[Z(p)] = \delta_p[W(p, s)]$.

It should be noted that $Z(p)$ is the impedance matrix of any lossless coupling network in a realization of $W(p, s)$, minimal in s, and hence we come to the important conclusion that if a realization of $W(p, s)$ is minimal in one of the variables it is automatically minimal in the other variable.

REFERENCES

1. Levenstein, H., "Theory of Networks of Linearly Variable Resistances," Proceedings of the IRE, *46,* No. 2 (February 1958), pp. 486–493.
2. Ozaki, H., and Kasami, T., "Positive Real Functions of Several Variables and their Applications to Variable Networks," IRE Transactions on Circuit Theory, *CT-7,* No. 3 (September 1960), pp. 251–260.
3. Ansell, H. G., "Networks of Transmission Lines and Lumped Reactances," IRE Transactions on Circuit Theory, *CT-11,* No. 2 (June 1964), pp. 214–223.
4. Shaffer, C. V., "Transformerless n-Port Symmetrical-Transmission-Line Synthesis," Ph.D. Dissertation, Stanford University, August 1965.
5. Rao, T. N. and Newcomb, R. W., "Synthesis of Lumped-Distributed RC n-Ports," IEEE Transactions on Circuit Theory, *CT-12,* No. 4 (December 1966), pp. 458–460.
6. Kaufman, W. M., "The Theory of a Monolithic Null Device and Some Novel Circuits," Proceedings of the IRE, *48,* No. 9 (September 1960), pp. 1540–1545.
7. Rao, T. N., "Synthesis of Lumped-Distributed RC Networks," Ph.D. Dissertation, Stanford University, Stanford, California.
8. Koga, T., "Synthesis of Finite Passive n-Ports with Prescribed Two-Variable Reactance Matrices," IEEE Transactions on Circuit Theory, *CT-12,* No. 1 (March 1966), pp. 31–52.
9. Youla, D. C., "The Synthesis of Networks Containing Lumped and Distributed Elements—Part I," Report No. PIBMRI-1323-66, New York; Polytechnic Institute of Brooklyn, March 1966. Also published in the Proceedings of the Polytechnique Symposium on Generalized Networks, New York, 1966, pp. 289–343.
10. Youla, D. C. and Tissi, P., "N-Port Synthesis via Reactance Extraction," IEEE International Convention Record, Part 7, 1966, pp. 183–208.
11. Newcomb, R. W., *Linear Multiport Synthesis,* New York: McGraw-Hill Book Company, 1966.
12. Gantmacher, F. R., *The Theory of Matrices,* Vols. I and II, New York: Chelsea Publishing Company, 1959.
13. McMillan, B., "Introduction to Formal Realizability Theory I and II," B.S.T.J., *31,* No. 2 (March 1952), pp. 217–279; No. 3 (April 1952), pp. 541–600.
14. Ho, B. L. and Kalman, R. E., "Effective Construction of Linear State Variable Models from Input-Output Data," Presented at the Allerton Conference on Circuit Theory, University of Illinois, Monticello, Illinois, October 20–22, 1965.
15. Oono, Y. and Yasuura, K., "Synthesis of Finite Passive 2n-Terminal Networks with prescribed Scattering Matrices," Memoirs of the Faculty of Engineering, Kyushu University, Fukuoka, Japan, *14,* No. 2 (May 1954), pp. 125–177.
16. Davis, M. C., "Factoring the Spectral Matrix," IEEE Transactions on Automatic Control, *AC-8* (October 1963), pp. 296–305.
17. Youla, D. C., "On the Factorization of Rational Matrices," IRE Transactions tions on Information Theory, *IT-7,* No. 3 (July 1961), pp. 296–305.
18. Newcomb, R. W., "On the Realization of Multivariable Transfer Functions," Research Report EERL 58, Cornell University, December 1966.
19. Guillemin, E. A., *The Mathematics of Circuit Analysis,* New York: John Wiley and Sons, 1949.
20. Talbot, A., "Some Theorems on Positive Functions," IEEE Transactions on Circuit Theory, *CT-12,* No. 4 (December 1965), pp. 607–608.
21. Rao, T. N. and Newcomb, R. W., "New Synthesis method for positive real matrices," *Electronics Letters, 3,* No. 8 (August 1967), pp. 349–350.
22. Koga, T., "Synthesis of Passive n-Ports with prescribed positive real matrices of several variables," IEEE Transactions on Circuit Theory, *CT-15,* No. 1 (March 1968), pp. 2–23.

Computation of the Noncentral Chi-Square Distribution*

By G. H. ROBERTSON

(Manuscript received July 17, 1968)

This article gives a formula that allows accurate values of the cumulative noncentral chi-square distribution to be computed. Although this distribution has been recognized for a long time, none of the standard references give formulae that are suitable for computing accurate values over an extensive range of the parameters; approximations in terms of the chi-square distribution are usually recommended. A program written by the author, based on the formula given here, has been successful for computations involving more than 10,000 degrees of freedom. Since many steps are required when the degrees of freedom are as large as this, the program is not "fast" but it is believed to be accurate.

I. INTRODUCTION

The Non-Central Chi-Square Distribution is encountered in many statistical problems, one of the most important in communications studies being the detection of signals in noise using a square-law detector.[1] Marcum discussed this application but concluded that a satisfactory algorithm for computing system performance could not be based on the formula he used.[2] This article shows that a satisfactory algorithm can be based on the formula that Marcum derived if the expression is expanded in a power series and the terms are properly grouped before being evaluated.

More recently Urkowitz[3] discussed detection system performance in which the above distribution arose and recommended that approximations in terms of the chi-square distribution, given by Patnaik,[1] be used for computation. While these approximations are adequate for some purposes, it is desirable to have a reliable and accurate method of computing values, if only to check the approximations.

* This work was supported by the U. S. Navy under contract N00039-68-C-3584.

II. INTEGRATION OF NONCENTRAL CHI-SQUARE DISTRIBUTION

If the signal-to-noise power ratio is x for the sum of η independent samples of the output of a square-law detector, the following integral* gives the probability that the sum will be y or more. The variables are normalized to the variance of the individual noise samples, so the average signal-to-noise power ratio for one sample is x/η and the average output per sample is y/η. Considering one sample of noise to be the sum of the squares of two independent gaussian variables of unit variance, the integral is related to the noncentral chi-square distribution by the conversions given in equation (8).

$$Q = \int_y^\infty \left(\frac{z}{x}\right)^{(\eta-1)/2} \exp\,(-z - x) I_{\eta-1}[2(zx)^{\frac{1}{2}}]\, dz. \tag{1}$$

From (Ref. 4, Section 8.445) $I_{\eta-1}\ [2\ (zx)^{\frac{1}{2}}]$ is the modified Bessel function

$$I_m(t) = \sum_{k=0}^{\infty} \frac{(t/2)^{m+2k}}{k!\ \Gamma(m+k+1)}. \tag{2}$$

Thus

$$Q = \frac{\exp\,(-x)}{\Gamma(\eta)} \int_y^\infty \exp\,(-z) z^{\eta-1}\, dz + \frac{\exp\,(-x)}{\Gamma(\eta)} \int_y^\infty \exp\,(-z) z^{\eta-1}$$
$$\cdot \left[\frac{xz}{1!\ \eta} + \frac{(xz)^2}{2!\ \eta(\eta+1)} + \frac{(xz)^3}{3!\ \eta(\eta+1)(\eta+2)} + \cdots\right] dz. \tag{3}$$

Notice that

$$\int_y^\infty \exp\,(-z) z^{\eta-1}\, dz = \Gamma(\eta, y) \tag{4}$$

the incomplete gamma function (Ref. 5, Section 6.5.3). Since

$$\int_y^\infty t^p \exp\,(-t)\, dt = -t^p \exp\,(-t) \Big|_y^\infty + p \int_y^\infty t^{p-1} \exp\,(-t)\, dt \tag{5}$$

equation (3) can be written

$$Q = \frac{\exp\,(-x)}{\Gamma(\eta)} \Bigg[\Gamma(\eta, y)$$
$$+ \frac{x}{1!} \left\{ \Gamma(\eta, y) + \left(\frac{y}{\eta}\right) y^{\eta-1} \exp\,(-y) \right\}$$

* Sometimes called the generalized Marcum Q-function. See Ref. 2.

$$+ \frac{x^2}{2!}\left\{\Gamma(\eta, y) + \left(\frac{y}{\eta} + \frac{y^2}{\eta(\eta+1)}\right)y^{\eta-1}\exp(-y)\right\}$$
$$+ \frac{x^3}{3!}\left\{\Gamma(\eta, y) + \left(\frac{y}{\eta} + \frac{y^2}{\eta(\eta+1)} + \frac{y^3}{\eta(\eta+1)(\eta+2)}\right)y^{\eta-1}\exp(-y)\right\}$$
$$\left. + \cdots \text{ and so on}\right]. \tag{6}$$

Summing the terms by columns gives

$$Q = \frac{\Gamma(\eta, y)}{\Gamma(\eta)} + \frac{y^{\eta-1}\exp(-y)}{\Gamma(\eta)}\left[\frac{y}{\eta}\exp(-x)\sum_{r=1}^{\infty}\frac{x^r}{r!}\right.$$
$$+ \frac{y^2}{\eta(\eta+1)}\exp(-x)\sum_{r=2}^{\infty}\frac{x^r}{r!}$$
$$+ \frac{y^3}{\eta(\eta+1)(\eta+2)}\exp(-x)\sum_{r=3}^{\infty}\frac{x^r}{r!}$$
$$\left. + \cdots \text{ and so on}\right] \tag{7}$$

A satisfactory computing algorithm can be based on equation (7) where we notice that Q can be expressed as the sum of two parts, $Q_1 = \Gamma(\eta, y)/\Gamma(\eta)$ which is independent of x, and another part which we call Q_2.

III. DISCUSSION

The noncentral chi-square cumulative distribution can be written $Q(\chi'^2|\nu, \lambda)$ (see Ref. 5, Section 26.4.25), where the distribution is integrated from χ'^2 to infinity, the number of degrees of freedom is ν, and the noncentral parameter is λ. This integral is the same as that given in equation (7) if we put

$$\begin{aligned} \nu &= 2\eta \\ \chi'^2 &= 2y \\ \lambda &= 2x \end{aligned} \tag{8}$$

so that

$$\begin{aligned} Q(2y \mid 2\eta, 2x) &= Q_1 + Q_2 \\ &= Q(2y \mid 2\eta) + Q_2 \end{aligned} \tag{9}$$

where $Q(2y|2\eta) = Q(\chi^2|\nu)$, the cumulative chi-square distribution (see Ref. 4, Section 26.4.2).

If M independent samples of the output of a square-law detector are averaged, when the input is narrowband gaussian noise plus a CW signal at the center of the band, Q can be used to find the probability that a threshold value will be exceeded. Expressing all parameters in units of the narrowband noise power, the desired threshold is y/η, x/η is the signal-to-noise power ratio, and $M = \eta$.

It is interesting that the Rayleigh distribution, and the Rice distribution, are equivalent to the chi-square and non-central chi-square distributions respectively, when the latter are expressed in terms of a parameter χ equal to the square root of χ^2, and $\eta = 1$.

Marcum[2] gave an expression of the form shown in equation (3) for the output of a square-law detector. He stated that it could only be used satisfactorily for values of η up to about 10. More recently Urkowitz[3] has discussed the integration of a square-law detector output and recommends that the noncentral chi-square distribution be computed using an approximation given by Patnaik[1] in terms of the chi-square distribution. Patnaik compares with exact values some results computed using the approximation and finds errors of the order of 1% around $Q = 0.5$. The accuracy is much less for values around unity and for values less than 0.01.

Brennan and Reed have shown that, when the order of the Bessel function in equation (1) is zero, corresponding to one sample, a straightforward recursive method applied to the resulting equation (6) can be used to compute the integral.[6] They suggested that a similar procedure could be used even on the form of equation (1) given here. However, as pointed out by Marcum, such a technique rapidly becomes useless as η increases above about 10.

A program written by the author, based on equation (7), has been used satisfactorily for η as large as 8192, and simultaneously for values of x/η up to 0.1. The exact values given by Patnaik were checked. Further checks were made possible by the development of a uniform asymptotic expansion by S. O. Rice, with which it is possible to get results outside the useful range of the algorithm given here.[7]

Table 1 compares values obtained with the author's program (CHISQ) and corresponding values supplied by S. O. Rice using his uniform asymptotic expansion (UAE), with results obtained using the Patnaik[1] and Gauss approximations (Ref. 5, Section 26.4.29).

The accuracy of the algorithm given in equation (7) decreases as x/η increases in the table, and the value in the last entry depended quite sensitively on the last digit of a 18 digit double precision con-

TABLE I — COMPARISON OF COMPUTATION METHODS

x/η	y/η	η	1-CHISQ	UAE	PATNAIK	GAUSS
0.01	1.05	8192	0.999801547E-00	0.9998015E-00	0.999801544E-00	0.999809E-00
0.05	1.05	8192	0.501464546E-00	0.5014645E-00	0.501467E-00	0.50110E-00
0.08	1.05	8192	0.552623909E-02	0.5526235E-02	0.552472E-02	0.56050E-02
0.1	1.05	8192	0.138627645E-04	0.1386275E-04	0.138700E-04	0.14803E-04
0.11	1.05	8192	0.281186446E-06	0.2811860E-06	0.280145E-06	0.31438E-06
0.12	1.05	8192	0.316387190E-08	0.3163860E-08	0.314279E-08	0.37651E-08
0.13	1.05	8192	0.20004E-10	0.199969E-10	0.197750E-10	0.25799E-10

stant used in the program. Notice that even for the last entry, the value actually computed, (CHISQ), appears to be correct to 14 places after the decimal point.

IV. EXTENSION TO A MORE GENERAL INTEGRAL

A more general integral is obtained by writing, for example, the βth moment of the partial noncentral chi-square distribution,

$$Q_\beta = \int_y^\infty z^\beta \left(\frac{z}{x}\right)^{(\eta-1)/2} \exp(-z - x) I_{\eta-1}[2(zx)^{\frac{1}{2}}]\, dz. \tag{11}$$

The corresponding form of equation (4) is

$$\int_y^\infty z^\beta \exp(-z) z^{\eta-1}\, dz = \Gamma(\xi, y) \tag{12}$$

where

$$\xi = \eta + \beta, \tag{13}$$

and the corresponding form of equation (7) becomes

$$\begin{aligned} Q = {} & \frac{\Gamma(\xi, y)}{\Gamma(\eta)} \exp(-x) {}_1F_1(\xi; \eta; x) \\ & + \frac{y^{\xi-1} \exp(-y)}{\Gamma(\eta)} \Bigg[\frac{y}{\eta} \exp(-x) \sum_{r=1}^{\infty} \frac{x^r}{r!} \frac{(\xi+1)_{r-1}}{(\eta+1)_{r-1}} \\ & \qquad + \frac{y^2}{\eta(\eta+1)} \exp(-x) \sum_{r=2}^{\infty} \frac{x^r}{r!} \frac{(\xi+2)_{r-1}}{(\eta+2)_{r-1}} \\ & \qquad + \frac{y^3}{\eta(\eta+1)(\eta+2)} \exp(-x) \sum_{r=3}^{\infty} \frac{x^r}{r!} \frac{(\xi+3)_{r-1}}{(\eta+3)_{r-1}} \\ & \qquad + \cdots \text{ and so on} \Bigg]. \end{aligned} \tag{14}$$

The confluent hypergeometric function ${}_1F_1(a; b; x)$ (Ref. 5, Section 13.1.10) is defined by

$$\begin{aligned} {}_1F_1(a; b; x) &= 1 + \frac{ax}{b1!} + \frac{a(a+1)x^2}{b(b+1)2!} + \frac{a(a+1)(a+2)x^3}{b(b+1)(b+2)3!} + \cdots \\ &= \sum_{r=0}^{\infty} \frac{(a)_r x^r}{(b)_r r!}. \end{aligned} \tag{15}$$

Equation (15) conveniently gives an example of Pochhammer's symbol $(\alpha)_r$ (Ref. 5, Section 6.1.22), also used in equation (14).

The structure of equation (14) is closely related to that of equation (7), so it can form the basis for a useful algorithm to compute the integral given in equation (11).

V. ACKNOWLEGEMENT

The interest of Mr. Stephen O. Rice in this work and his encouragement to publish the results, is gratefully acknowledged.

REFERENCES

1. Patnaik, P. B., "The Noncentral χ^2 and F Distributions and Their Applications," Biometrika, *36* (1949), pp. 202–232.
2. Marcum, J. I.: "A Statistical Theory of Target Detection by Pulsed Radar: Mathematical Appendix," Research Memorandum *RM-753,* The Rand Corporation, July 1, 1948. Published also as: Marcum, J. I. and Swerling, P., IRE Trans., PGIT, *IT-4,* No. 2 (April 1960).
3. Urkowitz, H., "Energy Detection of Unknown Deterministic Signals," Proc. IEEE, *55,* No. 4 (April 1967), pp. 523–531.
4. Gradshteyn, I. S. and Ryzhik, I. M., *Table of Integrals Series, and Products,* New York: Academic Press, 1965.
5. Abramowitz, M. and Stegun, I. A., eds., *Handbook of Mathematical Functions,* National Bureau of Standards, Applied Mathematics Series 55, Washington, D. C., 1964.
6. Brennan, L. E. and Reed, I. S., "A Recursive Method of Computing the Q Function," IEEE Trans. on Inf. Th. (Correspondence), *IT-11,* No. 2 (April 1965), pp. 312–313.
7. Rice, S. O., "Uniform Asymptotic Expansions for Saddle Point Integrals—Application to a Probability Distribution Occurring in Noise Theory," B.S.T.J., *47,* No. 9 (November 1968), pp. 1971–2013.

Uniform Approximation of Linear Systems*

By HARRY HEFFES and PHILIP E. SARACHIK

(Manuscript received August 6, 1968)

A method for reducing the complexity of the class of linear, time-varying, dynamic control systems is developed where the approach taken is that of uniform approximation (that is, modeling for a region of initial conditions). The objective of the modeling procedure is to choose a linear time-invariant system of given dimension, that minimizes a "worst-case" type of error criterion. Some results from the theory of widths of sets in Banach space are used to obtain bounds on the optimal approximation error as a function of the dimension of the approximating system. The use of these bounds in choosing the order of the approximation is discussed. An example illustrates the use of the derived results.

I. INTRODUCTION

In the analysis and design of control systems it is often useful to have low order constant coefficient models for the system. The problem of modeling linear systems by lower order linear systems has received considerable attention, but these analyses have usually been restricted to the modeling of constant coefficient systems.

References 1 through 5 contain various approaches to the system approximation problem; however, these analyses are generally restricted to the modeling of constant coefficient systems or systems which are forced with a given input or initial condition.

The control system analyst often finds himself dealing with nonstationary systems, but little work has been done in the area of optimally modeling this class of systems. The emphasis here is on modeling the class of linear, homogeneous time-varying systems with constant coefficient models. Reference 6 considers approximation of forced systems. Rather than design the model requiring solutions of the actual and approximate systems be "close" for a prescribed initial

* From a dissertation written as part of the requirements for a Ph.D. degree, New York University, 1968.

condition, the approach taken here is that of uniform approximation. Initial conditions are assumed to lie in some set in Euclidean space and a "worst-case" type of error criterion is defined. This eliminates tuning the model to specific conditions which may not be met when using the model. The material presented here thus generalizes previous work in that it extends the class of systems considered to time-varying systems and generalizes the error criterion to handle the more realistic problem of modeling for regions of initial conditions.

The problem is of importance, for example, in trajectory analysis where the linear time-varying system is obtained by linearizing a set of nonlinear equations about a nominal trajectory. In this case the time-varying nature of the system is described by partial derivatives evaluated along the nominal trajectory. Solutions to the resulting equations require simulation for each set of initial conditions. Using a constant coefficient model eliminates the need for repeated simulation.

The above example illustrates the use of a simplified model in analysis. The designer is interested in reducing the complexity of high-order nonstationary control system plants since this provides a means for designing simpler controllers based upon the model description. The results presented here not only allow one to obtain stationary models but simultaneously offer the opportunity to obtain lower order models of the original system.

II. PROBLEM DEFINITION AND FORMULATION

The system we are considering is described by the linear, time-varying, homogeneous vector differential equation

$$\dot{x}(t) = A(t)x(t) \tag{1}$$

with the outputs given by

$$y(t) = C(t)x(t) \tag{2}$$

where

$x(t)$ is an n-vector

$A(t)$ is an $n \times n$ matrix whose elements are bounded and piecewise continuous on $[t_o, t_f]$.

$C(t)$ is an $m \times n$ matrix whose elements are bounded and piecewise continuous on $[t_o, t_f]$.

It is desired to obtain a constant coefficient system of kth order* $(k \geqq m)$

$$\dot{\tilde{x}}(t) = \tilde{A}\tilde{x}(t) \tag{3}$$

such that the first m components of the state vector $\tilde{x}(t)$ closely approximate the components of $y(t)$ over the finite time interval $[t_o, t_f]$. Writing

$$\tilde{y}(t) = \tilde{C}\tilde{x}(t) \tag{4}$$

with

$$\tilde{C} = [I_{m \times m} \vdots 0]$$

the approximation problem can be viewed as choosing the elements of the $k \times k$ matrix $\tilde{A}$ such that $\tilde{y}(t)$ approximates $y(t)$ over $[t_o, t_f]$.

Since, in general, it is not known at the time of modeling what initial conditions will exist in the system, it is desirable to have the approximating system depend on a prescribed range of initial conditions rather than being tuned to any specific initial condition. The initial conditions are considered to lie in a closed, bounded convex subset of Euclidean n-space. That is,

$$x(t_o) \in R \subset E_n,$$

and the performance criterion is given by

$$J_k(\tilde{A}) = \max_{x_0 \in R} \min_{\tilde{x}_0 \in E_k} \int_{t_o}^{t_f} (y - \tilde{y})' W(t)(y - \tilde{y})\, dt \tag{5}$$†

where

$[t_o, t_f]$ is bounded
$y(t)$ is the solution of (1) and (2) with $x(t_o) = x_o$
$\tilde{y}(t)$ is the solution of (3) and (4) with $\tilde{x}(t_o) = \tilde{x}_o$
$W(t)$ is positive definite and bounded for all $t \in [t_o, t_f]$.

The above performance criterion corresponds to the worst case error in the approximation, corresponding to a given model, when the initial condition on the model, $\tilde{x}(t_o)$, is chosen optimally in terms of the initial conditions on the actual system. The modeling objective is to choose $\tilde{A}$ to minimize $J_k(\tilde{A})$(that is, minimize the maximum approximation error).

The approximation problem will be cast into a Hilbert space setting

* Notice that k is not restricted from above. It may be desirable to have $k > n$ if the original system is time-varying.

† In all that follows the prime denotes transpose.

which will permit the use of many of the general results to be presented in the next section. Vector spaces of solutions of the original system equations and any member of the class of approximate system equations are established. These spaces are then imbedded into an encompassing Hilbert space. It then is shown that the problem of finding an optimal approximation can be viewed as a problem of finding the "best" subspace (of a given form) of the Hilbert space to use in approximating solutions of the original system. Writing the output vector of the original system in terms of the transition matrix leads to

$$y(t) = C(t)\Phi(t, t_o)x(t_o) \tag{6}$$

where the transition matrix $\Phi(t, t_o)$ satisfies

$$\frac{d}{dt}\Phi(t, t_o) = A(t)\Phi(t, t_o) \tag{7}$$

with initial conditions

$$\Phi(t_o, t_o) = I. \tag{8}$$

Now if the original system is completely observable[8,9] on the finite interval $[t_o, t_f]$ the columns of the $m \times n$ matrix $C(t)\Phi(t, t_o)$ are linearly independent as vector-valued time functions. That is,

$$C(t)\Phi(t, t_o)x(t_o) \equiv 0 \quad \text{for all} \quad t \in [t_o, t_f]$$

implies $x(t_o) = 0$. For an observable system, the initial state can be determined uniquely from knowledge of the output. Since $x(t_o) = 0 \Rightarrow y(t) \equiv 0$ and, from observability, $y(t) \equiv 0 \Rightarrow x(t_o) = 0$ the linear independence of the columns of $C(t)\Phi(t, t_o)$ follows.

Let $\bar{\mathcal{Y}}$ be the linear space spanned by the n columns of $C(t)\Phi(t, t_o)$. The solutions of the original system lie in $\bar{\mathcal{Y}}$, which is of dimension n for an observable system. Notice that the number of components (m) in the vector y and the dimension of the space $\bar{\mathcal{Y}}$ need not be the same. If the system is not completely observable on $[t_o, t_f]$ the dimension of $\bar{\mathcal{Y}}$ is less than n.

The solutions to equations (3) and (4) can be written as

$$\tilde{y}(t) = \tilde{C}e^{\tilde{A}(t-t_o)}\tilde{x}(t_o) \tag{9}$$

where

$$e^{\tilde{A}(t-t_o)} = \sum_{i=0}^{\infty} \frac{\tilde{A}^i(t - t_o)^i}{i!}$$

and

$$\frac{d}{dt} e^{\tilde{A}(t-t_o)} = \tilde{A} e^{\tilde{A}(t-t_o)}. \tag{10}$$

It is thus seen that solutions $\tilde{y}(t)$ lie in the vector space spanned by the k columns of the $m \times k$ matrix $\tilde{C}e^{\tilde{A}(t-t_o)}$. Denote this vector space as $\mathcal{Y}_k$. If $\tilde{A}$ is such that the approximate system is observable then $\mathcal{Y}_k$ is of dimension k and the k columns of $\tilde{C}e^{\tilde{A}(t-t_o)}$ form a basis

$$\{\mathbf{g}_i ; i = 1, \cdots , k\}$$

for the k-dimensional vector space $\mathcal{Y}_k$ of approximating solutions. These basis elements can be written as

$$g_i(t) = \tilde{C}e^{\tilde{A}(t-t_o)}K_i \tag{11}$$

$$\mathbf{g}_i = \{g_i(t); t \in [t_o , t_f]\} \tag{12}$$

where K_i is the ith column of the $k \times k$ identity matrix. If the approximation is not observable the dimension of $\mathcal{Y}_k$ is less than k. In any case vector spaces $\mathcal{Y}_k$ with basis elements of the form (11) characterize the approximating systems where $\tilde{A}$ is a $k \times k$ real matrix. Defining

$$\mathfrak{D}_k = \{\mathcal{Y}_k ; \mathbf{g}_1 , \cdots , \mathbf{g}_k \text{ span } \mathcal{Y}_k\} \tag{13}$$

where $g_i(t)$ is given by equation (11) and $\tilde{A}$ is any real constant $k \times k$ matrix casts the problem into finding an element of $\mathfrak{D}_k$ minimizing J_k.

The problem of finding an optimal approximation has been cast into the problem of finding an extremal space $\mathcal{Y}_k^* \in \mathfrak{D}_k$ of approximating solutions. A Hilbert space $\mathcal{H}$ containing $\bar{\mathcal{Y}}$ and all members of $\mathfrak{D}_k$ will now be constructed.

Recall that the elements of $\bar{\mathcal{Y}}$ and $\mathcal{Y}_k$ are real, vector-valued, time functions having m components. Thus each element of the Hilbert space $\mathcal{H}$ to be constructed will have m components. The inner product in $\mathcal{H}$ is defined by

$$(\mathbf{y}_1 , \mathbf{y}_2) = \int_{t_o}^{t_f} y_1'(t)W(t)y_2(t)\,dt \tag{14}$$

where $W(t)$ is a real symmetric $m \times m$ matrix which is positive definite for $t \in [t_o , t_f]$ and whose elements are bounded for $t \in [t_o , t_f]$. Notice that this is the same matrix appearing in the performance criterion given by equation (5). The norm of an element in $\mathcal{H}$ is given by

$$||\mathbf{y}|| = (\mathbf{y}, \mathbf{y})^{\frac{1}{2}}. \tag{15}$$

The Hilbert space $\mathcal{H}$ is defined as

$$\mathcal{H} = \{\mathbf{y}; \mathbf{y} \text{ has } m \textbf{ components}, ||\,\mathbf{y}\,|| < \infty\}$$

where $\| \mathbf{y} \|$ is given by (15) and the inner product given by (14).

Since

$$t_f - t_o < \infty$$

and the elements of $A(t)$ and $C(t)$ are bounded it follows that solutions of equations (1) and (2) are bounded thus yielding

$$\bar{\mathcal{Y}} \subset \mathcal{H}.$$

Since elements of $\mathcal{Y}_k$ are bounded over the finite interval $[t_o , t_f]$

$$\mathcal{Y}_k \subset \mathcal{H}.$$

That $\mathcal{Y}_k$ and $\bar{\mathcal{Y}}$ are subspaces of $\mathcal{H}$ follows from the fact that any finite-dimensional linear set in a normed space is closed[10].

The set of functions to be approximated are solutions to the original system equations with the initial conditions $x(t_o)$ satisfying

$$x(t_o) \ \varepsilon \ R \subset E_n$$

where R is a closed, bounded convex subset of Euclidean n-space. Writing

$$\mathfrak{F} = \{\mathbf{y};\, y(t) = C(t)\Phi(t, t_o)x(t_o),\, x(t_o) \ \varepsilon \ R\} \tag{16}$$

gives

$$J_k(\tilde{A}) = \max_{\mathbf{y} \varepsilon \mathfrak{F}} \min_{\tilde{\mathbf{y}} \varepsilon \mathcal{Y}_k} \| \mathbf{y} - \tilde{\mathbf{y}} \|^2 \tag{17}$$

where the modeling objective is to find

$$\bar{d}_k^2 \triangleq \inf_{\mathcal{Y}_k \varepsilon \mathfrak{D}_k} \max_{\mathbf{y} \varepsilon \mathfrak{F}} \min_{\tilde{\mathbf{y}} \varepsilon \mathcal{Y}_k} \| \mathbf{y} - \tilde{\mathbf{y}} \|^2. \tag{18}$$

Before proceding to solve the formulated approximation problem, some results from the theory of widths in Banach space are outlined. Lower bounds on the optimal performance are found as a function of the dimension of the approximating system.

III. WIDTHS OF SETS IN BANACH SPACE AND LOWER BOUNDS*

Classically, approximation theory was concerned with the following problem. Given a function to approximate and a set of approximating functions (sinusoids, exponentials, and polynomials, for example) find that linear combination of approximating functions which

* Ref. 7 contains an excellent treatment of widths of sets in Banach space.

minimizes some distance function. Notice that here the approximating functions are given as part of the problem statement.

Rather than approximate a single function, the problem under consideration is to approximate the class of functions $\mathfrak{F}$ given by (16). For a given class of functions $\mathfrak{F}$ it is desired to obtain a "best" set of approximating functions rather than to choose the set arbitrarily. A measure of comparison is introduced which enables one to evaluate the efficiency of different sets of approximating functions. The following definitions serve to illustrate these ideas.

Let $\mathfrak{B}$ be a Banach space containing a set of functions $\mathfrak{F}$ to be approximated by elements of an n-dimensional subspace, X_n, of $\mathfrak{B}$. It is desired to find the "best" n-dimensional subspace, or equivalently the "best" set of approximating functions to use in approximating elements of $\mathfrak{F}$.

For a given $f \in \mathfrak{F}$ and $X_n \subset \mathfrak{B}$

$$\inf_{x \in X_n} \| f - x \|$$

represents how well one can do in approximating a given f with elements of X_n. Taking the supremum of the above quantity over all elements in $\mathfrak{F}$ leads to the following definition.

Definition 1: The *deviation* of $\mathfrak{F}$ from X_n is given by

$$E_{X_n}(\mathfrak{F}) = \sup_{f \in \mathfrak{F}} \inf_{x \in X_n} \| f - x \|.$$

The deviation represents the worst case approximation error over the class $\mathfrak{F}$ when using elements of X_n. Notice that the deviation serves as a performance measure of X_n. Taking the infimum of the deviation over all n-dimensional subspaces of $\mathfrak{B}$ leads to the following definition.

Definition 2: The nth *width* of $\mathfrak{F}$ is given by

$$\begin{aligned} d_n(\mathfrak{F}) &= \inf_{X_n \subset \mathfrak{B}} E_{X_n}(\mathfrak{F}) \\ &= \inf_{X_n \subset \mathfrak{B}} \sup_{f \in \mathfrak{F}} \inf_{x \in X_n} \| f - x \|. \end{aligned}$$

Some of the elementary results following from the above definitions are

(*i*) The monotonicity of the width:

$$d_0(\mathfrak{F}) \geqq d_1(\mathfrak{F}) \geqq d_2(\mathfrak{F}) \geqq \cdots$$

and

(*ii*) The nested property: If $\mathfrak{F}_1 \subset \mathfrak{F}_2 \subset \cdots$ then

$$d_n(\mathfrak{F}_1) \leqq d_n(\mathfrak{F}_2) \leqq \cdots.$$

Notice that

$$J_k(\tilde{A}) = E^2_{\mathcal{Y}_k}(\mathfrak{F}). \tag{19}$$

In defining $\bar{d}^2_k$ the infimum of the square of the deviation was taken over the j-dimensional ($j \leqq k$) subspaces in $\mathfrak{D}_k$ whereas in defining d_k the infimum was taken over all k-dimensional subspaces of $\mathcal{B}$. Using the monotonicity property of the width, with $\mathcal{H}$ serving as the required Banach space, gives

$$J_k(\tilde{A}) \geqq d^2_k(\mathfrak{F}) \tag{20}$$

for any $k \times k$ matrix $\tilde{A}$.

Definition 3: U_n is a *closed ball of radius* r in X_n if

$$U_n = \{x \varepsilon X_n ; \| x \| \leqq r\}.$$

The following theorem, by Gohberg and Krein, is proved in Ref. 7 and will be found useful.

Theorem: If X_{n+1} is an $(n + 1)$-dimensional subspace of a Banach space $\mathcal{B}$ and if U_{n+1} is the closed ball of radius r in X_{n+1} then $d_n(U_{n+1}) = r$.

This theorem and the nested property of widths can be used to obtain lower bounds on $d_n(\mathfrak{F})$. This lower bound can be obtained by constructing a ball in an $(n + 1)$-dimensional subspace and choosing r such that $U_{n+1} \subset \mathfrak{F}$. Using the nested property then leads to

$$r = d_n(U_{n+1}) \leqq d_n(\mathfrak{F}). \tag{21}$$

Since

$$\bar{d}_k \geqq d_k \tag{22}$$

the radius of ball also serves as a lower bound on $(J_k)^{\frac{1}{2}}$.

Lemma 1: *Let $\Phi(t, t_o)$ and $C(t)$ be the transition matrix and output matrix, respectively, of the original system* (1) *and* (2). *Assume this system to be completely observable on $[t_o , t_f]$. Let $W(t)$ satisfy the previously stated conditions. Then the matrix*

$$M = \int_{t_o}^{t_f} \Phi'(t, t_o)C'(t)W(t)C(t)\Phi(t, t_o) \, dt \tag{23}$$

is positive definite.

Proof: Consider the quadratic form $x_0'Mx_0 = \| \mathbf{y} \|^2 \geqq 0$ where

$$x(t_o) = x_0 , \quad \text{that is,} \quad y(t) = C(t)\Phi(t, t_o)x_0 .$$

Now $\| \mathbf{y} \|^2 = 0 \Rightarrow y(t) \equiv 0$ on $[t_o , t_f]$.

Since the system is observable $y \equiv 0 \Rightarrow x_0 = 0$. Thus M is positive definite.

The following theorem provides the lower bound on the performance.

Theorem 1: *Let R be the closed region of initial conditions on the original system and let $x(t_o) = 0$ be an interior point of R. Assume the system to be completely observable on $[t_o, t_f]$. Denote the boundary of R by ∂R and let*

$$\rho^2 \triangleq \min_{x(t_o) \varepsilon \partial R} x'(t_o)x(t_o). \tag{24}$$

Let the eigenvalues of the positive definite matrix M be ordered $\lambda_1(M) \geqq \lambda_2(M) \geqq \cdots \geqq \lambda_n(M)$. Then the performance, for any k-dimensional approximating system, satisfies $J_k(\tilde{A}) \geqq \rho^2\lambda_{k+1}(M)$ for $k < n$.

Proof: Let

$$\mathfrak{F} = \{\mathbf{y};\, y(t) = C(t)\Phi(t, t_o)x(t_o),\, x(t_o) \,\varepsilon\, R\}.$$

A $k + 1$ dimensional ball will now be constructed which is a subset of $\mathfrak{F}$. Consider the $k + 1$ dimensional ball of radius r

$$U_{k+1} = \{\mathbf{y};\, y(t) = C(t)\Phi(t, t_o)x(t_o),\, x(t_o) \,\varepsilon\, E_{k+1} \subset E_n,\, \|\mathbf{y}\| \leqq r\}$$

E_{k+1} and r will be chosen such that $U_{k+1} \subset \mathfrak{F}$. Since M is real and symmetric it can be diagonalized with an orthogonal matrix T. Thus $M = T'\Lambda T$ and

$$\|\mathbf{y}\|^2 = [Tx(t_o)]'\Lambda[Tx(t_o)] = z'\Lambda z$$

where

$$T' = T^{-1}$$

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

and

$$z = Tx(t_o).$$

Defining

$$E_{k+1} = \{x(t_o);\, [Tx(t_o)]_i = 0, \qquad i = k + 2, \cdots, n\}.$$

and

$$r^2 = \rho^2\lambda_{k+1}(M)$$

gives

$$U_{k+1} = \{\mathbf{y};\ \ y(t) = C(t)\Phi(t,\,t_o)x(t_o),$$
$$\|\,\mathbf{y}\,\|^2 \leqq \rho^2\lambda_{k+1}(M),\ \ [Tx(t_o)]_i = 0$$
$$i = k+2,\ \cdots,\ n\}.$$

Thus for $\mathbf{y} \,\varepsilon\, U_{k+1}$

$$\|\,\mathbf{y}\,\|^2 = x'(t_o)Mx(t_o) = \sum_{i=1}^{n} z_i^2\lambda_i \leqq \rho^2\lambda_{k+1}\,.$$

Since

$$z_i = 0 \qquad i = k+2,\ \cdots,\ n$$

and

$$\frac{\lambda_i}{\lambda_{k+1}} \geqq 1 \qquad \text{for} \quad i \leqq k+1$$

we have

$$\sum_{i=1}^{n} z_i^2 = x_0'x_0 \leqq \rho^2.$$

It then follows, from the definition of ρ^2 and the fact that zero is an interior point of R, that $x_0 \,\varepsilon\, R$ and therefore $\mathbf{y} \,\varepsilon\, \mathfrak{F}$. Thus $U_{k+1} \subset \mathfrak{F}$ and the desired result

$$J_k(\tilde{A}) \geqq d_k^2(\mathfrak{F}) \geqq \rho^2\lambda_{k+1}(M) \qquad k < n$$

follows.

Remarks: Recalling that the eigenvalues of M are ordered, we notice that the lower bound is a decreasing function of the dimension of the approximating system. This result can be used to determine what order aproximating system (at least) need be considered to achieve a given performance. We emphasize that the bound depends on the original system and is obtainable prior to the modelling procedure. From an engineering viewpoint, if one has an approximating system whose performance is "close" to the bound it may not be necessary to seek the minor improvement. Notice that the only property of R appearing in the lower bound is ρ and no attempt was made to take the orientation of the set into account. The bound will therefore be least conservative when R is a hypersphere of radius ρ.

IV. EVALUATING THE PERFORMANCE FUNCTION

In this section the problem of finding the performance (or equivalently the deviation) of a given approximating system is considered. The optimal choice of initial conditions on the approximating system is obtained using some elementary Hilbert space concepts and it is shown that

$$\inf_{\tilde{\mathbf{y}} \varepsilon \mathcal{Y}_k} || \mathbf{y} - \tilde{\mathbf{y}} ||^2$$

is a positive semidefinite quadratic form in $x(t_o)$. Next, properties of convex functions are used to evaluate the performance for different classes of regions of initial conditions; namely, for ellipsoids and convex polyhedra. The Powell algorithm for minimizing a function of several variables, without calculating derivatives, is then outlined and applied to the system approximation problem.

The problem of finding

$$\delta^2 = \inf_{\tilde{\mathbf{y}} \varepsilon \mathcal{Y}_k} || \mathbf{y} - \tilde{\mathbf{y}} ||^2 \tag{25}$$

is equivalent to finding the best choice of initial conditions on a given approximating system characterized by $\mathcal{Y}_k \ \varepsilon \ \mathfrak{D}_k$. It can be shown[11] that there exists a unique $\tilde{\mathbf{y}}^* \ \varepsilon \ \mathcal{Y}_k$ ($\mathbf{y}^*$ is called the projection of $\mathbf{y}$ in the space $\mathcal{Y}_k$) such that

$$\delta^2 = || \mathbf{y} - \tilde{\mathbf{y}}^* ||^2 = || \mathbf{y} ||^2 - || \tilde{\mathbf{y}}^* ||^2. \tag{26}$$

Furthermore, since $\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_k$ spans $\mathcal{Y}_k$, $\tilde{\mathbf{y}}^*$ has the representation

$$\tilde{\mathbf{y}}^* = \sum_{i=1}^{k} \mathbf{g}_i \tilde{x}_i^*$$

where

$$G(\mathbf{g}_1, \cdots, \mathbf{g}_k)\tilde{x}^* = \begin{bmatrix} (\mathbf{y}, \mathbf{g}_1) \\ \vdots \\ (\mathbf{y}, \mathbf{g}_k) \end{bmatrix} \tag{27}$$

$$\tilde{x}^* = \begin{bmatrix} \tilde{x}_1^* \\ \vdots \\ \tilde{x}_k^* \end{bmatrix}$$

and G is the Grammian of $\{\mathbf{g}_i\,; i = 1, \cdots, k\}$, that is,

$$[G(\mathbf{g}_1, \cdots, \mathbf{g}_k)]_{i,j} = (\mathbf{g}_i, \mathbf{g}_j) \qquad i, j = 1, \cdots, k.$$

Any solution to (27) results in an optimum choice of initial conditions on the approximate system. If the $\mathbf{g}_i$'s are linearly independent (this corresponds to the system being observable) the Grammian is invertible and $\tilde{x}^*$ is unique. Thus

$$\tilde{x}^*(t_o) = G^{\dagger}(\mathbf{g}_1, \cdots, \mathbf{g}_k)\begin{bmatrix}(\mathbf{y}, \mathbf{g}_1)\\ \vdots \\ (\mathbf{y}, \mathbf{g}_k)\end{bmatrix} \tag{28}$$

where $G^{\dagger}$ is the pseudoinverse[12] of G.

The Grammian is given by

$$G(\mathbf{g}_1, \cdots, \mathbf{g}_k) = \int_{t_o}^{t_f} e^{\tilde{A}'(t-t_o)}\tilde{C}'W(t)\tilde{C}e^{\tilde{A}(t-t_o)}\,dt \tag{29}$$

and

$$(\mathbf{y}, \mathbf{g}_i) = K_i'Fx(t_o) \tag{30}$$

where F is given by

$$F = \int_{t_o}^{t_f} e^{\tilde{A}'(t-t_o)}\tilde{C}'W(t)C(t)\Phi(t, t_0)\,dt. \tag{31}$$

Using (30) in (28) gives

$$\tilde{x}^*(t_o) = G^{\dagger}Fx(t_o). \tag{32}$$

Thus the optimal initial condition on the approximating system is obtained by linearly transforming the actual initial condition with the $(k \times n)$ matrix $G^{\dagger}F$. Using the orthogonality property (26) yields

$$\| \mathbf{y} - \tilde{\mathbf{y}}^* \|^2 = \| \mathbf{y} \|^2 - \tilde{x}^{*\prime}(t_o)G\tilde{x}^*(t_o).$$

Letting

$$M = \int_{t_o}^{t_f} \Phi'(t, t_o)C'(t)W(t)C(t)\Phi(t, t_o)\,dt \tag{33}$$

and using (32) and the symmetry of G (and thus $G^{\dagger}$) gives

$$\| \mathbf{y} - \tilde{\mathbf{y}}^* \|^2 = x'(t_o)(M - F'G^{\dagger}F)x(t_o). \tag{34}$$

In summary,

$$\delta^2 = \inf_{\tilde{\mathbf{y}} \in \mathcal{Y}_k} \| \mathbf{y} - \tilde{\mathbf{y}} \|^2 = x'(t_o)\, Dx(t_o) \tag{35}$$

with

$$D = M - F'G^{\dagger}F. \tag{36}$$

Thus, finding the optimal initial condition on the approximating systems leads to the positive semidefinite quadratic form (35) for the approximation error. The above represents the first step in evaluating the performance of any given approximating system.

Since D is a positive semidefinite matrix, δ^2 defined by (35) is a convex function of the initial state $x(t_o)$. The following theorem from Ref.13 is useful in maximizing δ^2.

Theorem: If the absolute maximum of a convex function, defined on a closed, bounded, convex set, is finite then the absolute maximum is taken on at an extreme point of the set.

Remarks: An extreme point of a convex set is a point in the set that cannot be written as a convex combination of two other points in the set. Notice that an extreme point is a boundary point; however, generally not every boundary point is an extreme point. Thus, if one is seeking the absolute maximum of a convex function defined on a closed, bounded, convex set only boundary points need be considered. Also if the domain of definition is a convex polyhedron (a closed, bounded, convex set with a finite number of extreme points) the absolute maximum can be obtained by simply evaluating the function at the extreme points and choosing the largest value.

Two general classes of closed, bounded, convex regions of initial conditions are considered in this paper, the ellipsoid and the convex polyhedron.

Let the region under consideration be an ellipsoid defined by

$$R = \{x(t_o);\quad x'(t_o)Bx(t_o) \leqq r^2\} \tag{37}$$

where B is a positive definite, symmetric matrix and r is finite. Notice that R is closed, bounded, and convex. Now the constrained maximization problem is one with an inequality constraint. Using the convexity of R and δ^2, the absolute maximum of the quadratic form is seen to take place on the boundary of the set R. Thus the performance can be written

$$J_k(\bar{A}) = \max_{x(t_o)} x'(t_o)\, Dx(t_o)$$

subject to the constraint

$$x'(t_o)Bx(t_o) = r^2.$$

It can easily be shown that the $x(t_o)$ maximizing the quadratic form is the eigenvector of the matrix $B^{-1}D$ corresponding to the largest eigenvalue and the maximum is given by

$$J_k(\tilde{A}) = \lambda_{\max}(B^{-1}D)r^2. \tag{38}$$

A convex polyhedron is usually representative of the type of information one has as to the range of initial conditions. As an example of this situation consider the original system to represent linearized equations of motion of a space vehicle. Suppose it is known that the range of initial conditions are in terms of bounds on position, velocity deviations, and so on. For example,

$$| x_1(t_o) | \leqq 100 \text{ feet.}$$

$$| x_2(t_o) | \leqq 5 \text{ feet per second.}$$

This particular region is described by a rectangular region in state space with the extreme points being the corners

$$\begin{bmatrix} 100 \\ 5 \end{bmatrix}, \begin{bmatrix} 100 \\ -5 \end{bmatrix}, \begin{bmatrix} -100 \\ 5 \end{bmatrix}, \begin{bmatrix} -100 \\ -5 \end{bmatrix}.$$

In general for this type of initial condition region, that is,

$$| x_i(t_o) | \leqq b_i \qquad i = 1, \cdots, n,$$

the region has 2^n extreme points. Since δ^2 is an even function of $x(t_o)$ it is only necessary to consider 2^{n-1} extreme points eliminating from consideration the negative of any point considered.

The convex polyhedron region also is important, for example, since it may be used to simply approximate a more complex region. In general, let

$$x^{(i)} \; i = 1, 2, \cdots, N$$

be the extreme points of the convex polyhedron R. Using the convexity of δ^2 in the initial state $x(t_o)$ the absolute maximum δ^2 over R takes place at one of the $x^{(i)}$. Letting

$$\delta_i^2 = x^{(i)\prime}Dx^{(i)}$$

where D is given by equation (36) leads to

$$J_k(\tilde{A}) = \max [\delta_1^2, \delta_2^2, \cdots, \delta_N^2]. \tag{39}$$

V. MINIMIZING THE PERFORMANCE FUNCTION

Since it is a fairly simple matter to evaluate the performance, whereas evaluating the gradient of the performance function requires significant computational effort, it is desirable to use a numerical procedure not requiring a gradient computation. Notice that $J_k(\tilde{A})$ is not generally differentiable. Here, for completeness, the Powell method of minimizing a function of several variables without calculating derivatives is presented.[14] Reference 15 contains a summary of the various minimization techniques available not requiring the computation of a derivative. See Refs. 14 and 15 for a more detailed description of the methods and their convergence properties.

Consider a real, scalar, valued function of N real variables $a_1, \cdots, a_N$ written $f(a)$. Powell's iterative scheme concerns itself with finding the minimum of $f(a)$ without computing its derivative.

Each iteration of the modified Powell procedure starts with a search down N linearly independent directions

$$\eta_1, \eta_2, \cdots, \eta_N$$

starting with an initial guess a_o and defines a new set of directions for the next iteration.

An iteration of the recommended procedure, suggested by Powell, is:

(*i*) for $j = 1, 2, \cdots, N$ calculate λ_j such that $f(a_{j-1} + \lambda_j \eta_j)$ is minimum and define $a_j = a_{j-1} + \lambda_j \eta_j$.

(*ii*) Find the integer m, $1 \leqq m \leqq N$, such that $f(a_{m-1}) - f(a_m)$ is a maximum and define $\Delta = f(a_{m-1}) - f(a_m)$.

(*iii*) Calculate $f_3 = f(2a_N - a_o)$ and define

$$f_1 = f(a_o)$$

$$f_2 = f(a_N).$$

(*iv*) If either $f_3 \geqq f_1$ or

$$(f_1 - 2f_2 + f_3)(f_1 - f_2 - \Delta)^2 \geqq \tfrac{1}{2}\Delta(f_1 - f_3)^2$$

use the old directions $\eta_1, \cdots, \eta_N$ for the next iteration and use a_N for the next a_o, otherwise

(*v*) define $\eta = a_n - a_o$ and calculate λ such that $f(a_N + \lambda\eta)$ is minimum. Use

$$\eta_1, \cdots, \eta_{m-1}, \eta, \eta_{m+1}, \cdots, \eta_N$$

as the new directions and $a_N + \lambda\eta$ as the new a_o.

The performance functions, for the two classes of initial conditions being considered are given by (38) and (39) in terms of the matrix D defined in (36). The major effort in computing the performance function is seen to lie in the computation of D. Sylvester's expansion (see page 83 of Ref. 16) for computing $e^{\tilde{A}t}$ is useful in the computation of the matrices F and G.

The basic procedure can be outlined as follows:

(*i*) Compute and store $C(t)\Phi(t, t_o)$ for $t \varepsilon [t_o, t_f]$ using (7) and (8).

(*ii*) Evaluate M using (23).

(*iii*) If it is desired to compute the lower bounds to aid in choosing the dimension of the approximating system, compute the eigenvalues of M and obtain the bounds from the result of Theorem 1.

(*iv*) Choose starting values for $\tilde{A}$ and choose the directions for the initial search in the modified Powell method to be

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \cdots, \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \\ 1 \end{bmatrix}$$

where the above are k^2 vectors.

(*v*) Use modified Powell method to determine the minimum of the performance function. Each element of the vector a in the Powell method corresponds to an element of $\tilde{A}$.

VI. EXAMPLE

A linearized missile guidance loop may be expressed in the form

$$\dot{x}_1 = x_2, \qquad \dot{x}_2 = \frac{H}{m - t} x_3, \qquad \dot{x}_3 = u \tag{40}$$

where x_1 is the lateral position deviation from a nominal trajectory, x_2 is the lateral velocity deviation, x_3 is the attitude deviation in the given direction and u is the control signal. The relationship between the attitude and lateral acceleration is given through the time-varying gain $H/(m - t)$ which accounts for the loss of mass because of fuel consumption.

Suppose it is desired to approximate homogeneous solutions to (40) for initial conditions (at beginning of a stage) lying in a set R (R is defined later) with solutions of a constant coefficient system. The actual system (40) can be written in the vector-matrix form

$$\dot{x}(t) = A(t)x(t) \tag{41}$$

with output

$$y(t) = [1 \ 0 \ 0]x(t) = Cx(t) \tag{42}$$

where

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix}$$

and

$$A(t) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & \dfrac{H}{m - t} \\ 0 & 0 & 0 \end{bmatrix}. \tag{43}$$

Let

$$\| y \|^2 = \int_0^T y^2(t)\, dt. \tag{44}$$

Before proceeding to find the approximation it is instructive to determine the lower bounds on the optimal performance. This will naturally aid in choosing the dimension of the approximating system. The matrix, M, defined by (33), is given by

$$M = \int_0^T \Phi'(t, o)C'C\Phi(t, o)\, dt \tag{45}$$

with

$$\frac{d}{dt} \Phi(t, o) = A(t)\Phi(t, o). \tag{46}$$

The transition matrix, which is the solution to (46) with the identity initial condition, is given by

$$\Phi(t,\, o) = \begin{bmatrix} 1 & t & H\left\{(m - t)\ln\left(\dfrac{m-t}{m}\right) + t\right\} \\ 0 & 1 & -H\ln\left(\dfrac{m-t}{m}\right) \\ 0 & 0 & 1 \end{bmatrix}.$$

Evaluating M leads to

$$M = \begin{bmatrix} T & \dfrac{T^2}{2} & H\left\{\dfrac{T^2}{2} - \dfrac{(m-T)^2}{2}\ln\left(\dfrac{m-T}{m}\right) + \dfrac{1}{4}(T^2 - 2mT)\right\} \\ \dfrac{T^2}{2} & \dfrac{T^3}{3} & M_{23} \\ M_{13} & M_{23} & M_{33} \end{bmatrix}$$

with

$$\begin{aligned} M_{23} = H\Bigg[\frac{T^3}{3} - \frac{5}{36}m^3 - \frac{(2T+m)(m-T)^2}{6} \\ \cdot\ln\left(\frac{m-T}{m}\right) + \frac{(m-T)^2(4T+5m)}{36}\Bigg] \end{aligned}$$

and

$$\begin{aligned} M_{33} = H^2\Bigg[\frac{T^3}{3} - \frac{(m-T)^3}{3}\ln^2\left(\frac{m-T}{m}\right) + \frac{2}{9}(m-T)^3 \\ \cdot\ln\left(\frac{m-T}{m}\right) + \frac{2}{27}\{m^3 - (m-T)^3\} \\ - \frac{10}{36}m^3 - \frac{(2T+m)}{3}(m-T)^2 \\ \cdot\ln\left(\frac{m-T}{m}\right) + \frac{(m-T)^2(4T+5m)}{18}\Bigg]. \end{aligned}$$

Let the constants defining the problem be given by

$m = 15$ seconds (normalized mass)

$T = 10$ seconds

$H = 15$ (pound-seconds per slug) $\times 10^{-3}$

and let the region of initial conditions be given by

$$R = \{x(o);\ |x_1(o)| \leqq 30 \text{ feet},\quad |x_2(o)| \leqq 2 \text{ feet per second},\ |x_3(o)| \leqq 1 \text{ milliradian}\}.$$

Evaluating M for the above values of the constants leads to

$$M = \begin{bmatrix} 10 & 50 & 206 \\ 50 & 333 & 1570 \\ 206 & 1570 & 8082 \end{bmatrix}$$

with eigenvalues

$$\lambda_1 = 8393, \qquad \lambda_2 = 31, \qquad \lambda_3 = 1.1.$$

We have

$$J_0 \geqq 8,393$$
$$J_1 \geqq 31$$

and

$$J_2 \geqq 1.1.$$

Here J_0 represents

$$\max_{x_0 \varepsilon R} \| \mathbf{y} \|^2.$$

The second order approximation thus has the possibility of yielding a negligible approximation error. Thus in the remainder of this paper the optimal second order approximation will be sought. Thus

$$\dot{\tilde{x}} = \tilde{A}\tilde{x} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \tilde{x}$$

and

$$\tilde{y} = [1 \ \ 0]\tilde{x}.$$

The initial choice for $\tilde{A}$ in the iterative procedure is

$$\tilde{A}_o = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

which represents polynomial approximations to solutions of the original system.

The extreme points of R are given by

$$x^{(1)} = \begin{bmatrix} 30 \\ 2 \\ 1 \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} -30 \\ 2 \\ 1 \end{bmatrix}, \quad x^{(3)} = \begin{bmatrix} 30 \\ -2 \\ 1 \end{bmatrix} \text{ and } x^{(4)} = \begin{bmatrix} 30 \\ 2 \\ -1 \end{bmatrix}$$

and their negatives. Thus

$$J_2(\tilde{A}_o) = \max_i \{x^{(i)\prime} Dx^{(i)}\} = 340$$

where D is evaluated from (36). It is thus seen that the performance function is far greater than the lower bound and the possibility exists for a significant improvement. The result of applying the Powell algorithm to this problem yields

$$\tilde{A}^* = \begin{bmatrix} 0.244 & 0.827 \\ 0.177 \times 10^{-3} & 0.629 \times 10^{-3} \end{bmatrix}$$

and

$$J_2(\tilde{A}^*) = 33.4$$

with the eigenvalues of $\tilde{A}^*$ given by

$$\lambda_1(\tilde{A}^*) = 0.245$$

$$\lambda_2(\tilde{A}^*) = 0.30 \times 10^{-4}.$$

The above results are obtained after three iterations of the Powell algorithm. The G, F and D matrices are given by

$$G = \begin{bmatrix} 271 & 771 \\ 771 & 2230 \end{bmatrix}$$

$$F = \begin{bmatrix} 43.14 & 296.0 & 1459 \\ 112.2 & 832.6 & 4241 \end{bmatrix}$$

and

$$D = \begin{bmatrix} 4.4 \times 10^{-6} & -1.8 \times 10^{-4} & 1.2 \times 10^{-4} \\ -1.8 \times 10^{-4} & 7.3 & 3.5 \times 10^{-3} \\ 1.2 \times 10^{-4} & 3.5 \times 10^{-3} & 4.2 \end{bmatrix}.$$

Evaluating

$$\max_i \{x^{(i)} Dx^{(i)}\}$$

gives the maximum approximation error occurring at the extreme point

$$x^{(3)} = \begin{bmatrix} 30 \\ -2 \\ 1 \end{bmatrix}.$$

Figure 1 shows the solution of the actual and approximate system for this worst-case initial condition. The solutions are obtained from

$$y(t) = 30 - 2t + \Phi_{13}(t, o)$$

and

$$\tilde{y}(t) = [1\ 0]e^{\tilde{A}\cdot t}G^{-1}F\begin{bmatrix} 30 \\ -2 \\ 1 \end{bmatrix}.$$

$$\tilde{y}(t) = 4.82\, e^{\lambda_1 t} + 19.78\, e^{\lambda_2 t}.$$

The matrix relating the initial conditions is given by $G^{-1}F$, that is,

$$\tilde{x}(o) = G^{-1}Fx(o).$$

$$\tilde{x}(o) = \begin{bmatrix} 1.00 & 1.86 & -1.68 \\ -0.295 & -0.271 & 2.48 \end{bmatrix} x(o).$$

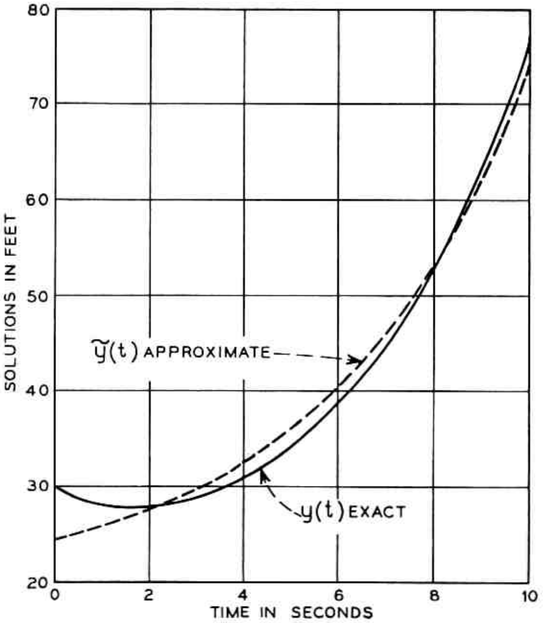


Fig. 1 — Exact and approximate solutions in worst case.

VII. CONCLUSIONS

A method for uniformly approximating solutions of linear, time-varying, homogeneous differential equations has been presented. The problem of approximating systems subject to control or reference inputs is considered in Ref. 6 for the class of exponential polynomial control inputs.

One of the objectives of modeling with constant coefficient systems was to obtain closed form approximations. Use of Sylvester's expansion allows one to derive these closed form expressions. However, more general classes of approximating systems can be sought while still maintaining the property that approximations are in closed form. For example, a general model of the form

$$\dot{\tilde{x}} = p(t)\tilde{A}\tilde{x}$$

where $p(t)$ is a scalar valued function possesses the closed form solution

$$\tilde{x}(t) = \exp\left[\tilde{A}\int_{t_o}^{t} p(\tau)\,d\tau\right]\tilde{x}(t_o)$$

and $p(t)$ as well as $\tilde{A}$ may be sought as part of the modeling procedure. A complexity constraint can be imposed on $p(t)$ by considering it to be a polynomial of given degree and the search for the model reduces to finding the coefficients of the polynomial as well as the elements of $\tilde{A}$.

VIII. ACKNOWLEDGMENTS

We greatly appreciate the valuable suggestions of D. L. Jagerman.

Harry Heffes thanks Bell Telephone Laboratories, Incorporated, for their financial assistance under the doctoral support plan; Philip Sarachik thanks the Air Force Office of Scientific Research for their financial assistance under grant AF OSR 747-66.

REFERENCES

1. McBride, L. E., Jr., Schaefgen, H. W., and Steigler, K., "Time-Domain Approximation by Iterative Methods," IEEE Trans. Circuit Theory *CT-13,* No. 4, (December 1966), pp. 381–387.
2. Mitra, D., "On the Reduction of Complexity of Linear Dynamic Models," United Kingdom Atomic Energy Authority Report AEEW-R520, 1967.
3. Davison, E. J., "A Method for Simplifying Linear Dynamic Systems," IEEE Trans. Automatic Control, Vol. *AC-11,* No. 1, pp. 93–101, January 1966.
4. Nordahl, D. H., and Melsa, J. L., "Modeling with Lyapunov Functions," Proc. Joint Automatic Control Conf., 1967, pp. 208–215.
5. Meier, L., III, "Approximation of Linear Constant Systems by Linear

Constant Systems of Lower Order," Ph.D. Dissertation, Stanford, Calif.: Stanford University, 1965.

6. Heffes, H., "Approximation of Linear Time-Varying Systems by Linear Constant Coefficient Systems Over Finite-Time Intervals," Doctoral Dissertation, New York University, June 1968.
7. Lorentz, G. G., *Approximation of Functions,* New York: Holt, Reinhart and Winston, 1966.
8. Kreindler, E. and Sarachik, P. E., "On the Concepts of Controllability and Observability of Linear Systems," IEEE Tran. Automatic Control, *AC-9,* No. 2, (April 1964), pp. 129–136.
9. Kalman, R. E., "Mathematical Description of Linear Dynamical Systems," *J. SIAM Control, 1,* No. 2 (1963), pp. 152–192.
10. Kantorovich, L. V. and Akilov, G. P., *Functional Analysis in Normed Spaces,* New York: MacMillan, 1964, p. 58.
11. Akhiezer, N. I. and Glazman, I. M., *Theory of Linear Operators in Hilbert Space,* New York: Ungar, 1961.
12. Penrose, R., "A Generalized Inverse for Matrices," Proc. Cambridge Phil. Soc., *51,* pt. 3 (July 1955), pp. 406–413.
13. Hadley, G., *Nonlinear and Dynamic Programming,* New York: Addison Wesley, 1964.
14. Powell, M. J. D., "An Efficient Method of Finding the Minimum of a Function of Several Variables Without Calculating Derivatives," The Computer Journal, *7,* No. 2 (1964), pp. 155–162.
15. Fletcher, R., "Function Minimization Without Evaluating Derivatives—A Review," The Computer Journal, *8,* (1966), pp. 33–41.
16. Frazer, R. A., Duncan, W. J., and Collar, A. R., *Elementary Matrices,* New York: Macmillan, 1946.

A Second Order Statistical Analysis of the Operation of a Limiter-Phase Detector-Filter Cascade

By W. D. WYNN

(Manuscript received August 30, 1968)

This paper presents a second-order statistical analysis for the cascade of a bandpass limiter, and ideal phase detector and a video filter. This cascade forms an important subsystem in the mathematical model of some coherent communication systems where information is transmitted by phase or frequency modulation of the carrier. We derive the autocorrelation function $R(t_1, t_2)$ of the video filter response when the bandpass limiter input is a fixed amplitude-phase modulated carrier plus stationary gaussian noise. The video filter response is wide sense stationary for some nontrivial cases; these include biphase, single tone, and stationary gaussian noise phase modulation. For these cases, we obtain the video filter output average power spectrum as the Fourier transform of $R(\tau)$ for all values of the limiter input signal-to-noise power ratio. An application of the results of this paper is the performance of a FM-PM demodulator for a set of parameters characteristic of one mode of operation of the Apollo Unified S-Band communications system. We present the performance as a family of curves of subcarrier channel output signal-to-noise power ratio as functions of the limiter input signal-to-noise ratio where subcarrier phase modulation index is a parameter. The approach is similar to the analysis by Davenport of the signal-to-noise ratio transfer characteristic of an isolated bandpass limiter.

I. INTRODUCTION

In some coherent communication systems, such as the Apollo Unified S-band system,[1] where information is transmitted by phase modulating a carrier, bandpass limiters[2] are used in the IF channels preceding the coherent demodulators. Ideally the bandpass limiter removes any amplitude modulation that might exist before the signal is demodulated.

Figure 1 shows a typical coherent phase demodulator used in such a system. This demodulator consists of a multiplication operation (a phase detector) with post-video filtering. The phase modulated signal is multiplied by a coherent carrier reference to yield a video signal containing the desired information. The signal into the limiter is usually accompanied by noise that is frequently assumed to be additive and gaussian. The presence of the noise affects the performance of the demodulator in a very complicated way because of the nonlinearity of the limiter. Thus it is difficult to evaluate the corruptive effect of the noise on the demodulated information.

One criterion of performance at points in a communication system is the signal-to-noise power ratio (S/N). For the cascade in Fig. 1, a problem of interest to the systems engineer is the video filter output S/N as a function of the input S/N to the limiter when the input noise is additive, stationary, and gaussian. The relationship is known between input and output S/N for an ideal bandpass limiter where the input is the sum of a stationary gaussian noise and a signal $P(t) \cos (\omega_c t + \phi)$ (see Ref. 2). For the analysis there, $P(t)$ is a random process and is slowly varying compared with $\cos \omega_c t$. The carrier phase ϕ is a random variable independent of $P(t)$ with a uniform distribution over $[0, 2\pi]$.

It is not possible to apply the known S/N transfer characteristic of the ideal bandpass limiter found in Ref. 2 directly to obtain the S/N transfer characteristic for the bandpass limiter-phase detector-video filter cascade. A knowledge of the form of the signal and the noise out of the bandpass limiter, and not just the S/N of this output, is necessary to determine the effect of the phase detector on the bandpass limiter response.

To obtain the cascade S/N transfer characteristic we apply the mathematical tools used in Ref. 2. The form of the signal assumed in the analysis of the cascade is $s(t) = P \cos [\omega_c + \theta(t) + \phi]$ where P is

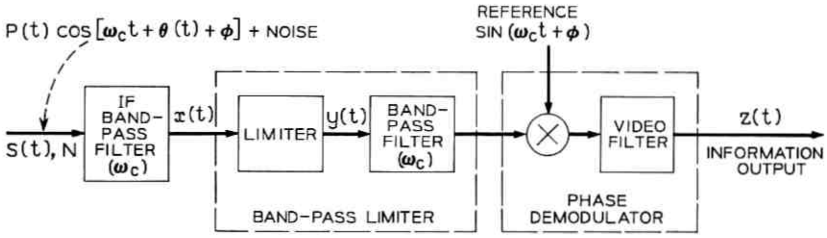


Fig. 1 — A coherent phase demodulator with IF bandpass limiting in the presence of additive noise.

a positive constant, $\theta(t)$ is phase modulation that is slowly varying compared with $\cos \omega_c t$, and ϕ is a random variable representing the arbitrary initial phase of the signal carrier. The probability density function of ϕ is assumed to be uniform in the interval $[0, 2\pi]$. The noise input to the bandpass limiter is assumed to be additive, stationary, and gaussian with zero mean and power spectral density N. The input noise, the modulation $\theta(t)$, and the carrier phase ϕ are assumed to be jointly statistically independent. For the following analysis, the limiter is assumed to be ideal with limit level l. The transfer function of an ideal limiter is defined by

$$y = l(x) = \begin{cases} +l, & x > 0 \\ 0, & x = 0 \\ -l, & x < 0. \end{cases} \tag{1}$$

A coherent carrier reference $\sin(\omega_c t + \phi)$ is assumed to be available for the demodulator where ϕ is the phase of the carrier.

II. THE SECOND ORDER STATISTICAL ANALYSIS

2.1 *A Cascade Model when $s(t)$ is Narrow Band Limited*

In order to obtain a S/N transfer characteristic for Fig. 1, the autocorrelation function of $z(t)$ is derived. When $R_z(t_1, t_2) = R_z(\tau)$ the average power spectrum of $z(t)$ is defined by the Fourier transform of $R_z(\tau)$ and the S/N transfer characteristic can be found. An analysis of the autocorrelation function of $z(t)$ does not seem possible for general $s(t)$. However, if the signal $s(t)$ is a narrow band-limited process such that the bandpass filters are narrow compared with the carrier frequency ω_c, the response $z(t)$ should be the same with or without the post bandpass filter that precedes the phase detector. The response of the nonlinearity $l(x)$ to an input $x(t) = s(t) + n(t)$ that is narrow band-limited about $\pm\omega_c$ is a family of terms narrow band-limited about the frequencies $\pm n\omega_c$ where $n = 0, 1, 2, 3, \cdots$ (see equation 13-53, section 13-1 of Ref. 3). Any narrow band-limited input to the phase detector that is not about $\pm\omega_c$ will generate a phase detector response above the cutoff frequency assumed for the video filter. For a narrow band-limited $x(t)$ the autocorrelation function of $z(t)$ is obtained from the analysis of Fig. 2.

2.2 *The Derivation of the Autocorrelation Function of $z(t)$*

Assume that the input $x(t)$ is narrow band-limited such that Figs. 1 and 2 yield equivalent $z(t)$. The autocorrelation function $R_z(t_1, t_2)$ is

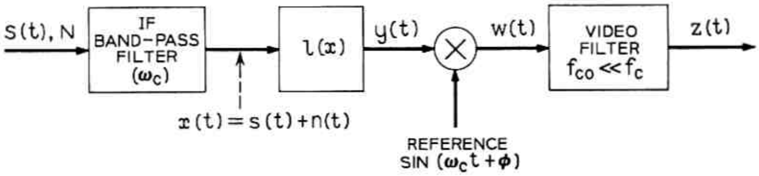


Fig. 2 — The narrow band equivalent receiver for the derivation of $R_z(t_1, t_2)$.

obtained by first deriving $R_w(t_1, t_2)$ from the model in Fig. 2. Since z and w are related by the linear video filter, $R_z(t_1, t_2)$ follows directly from $R_w(t_1, t_2)$.

The Laplace transform solution of a zero memory nonlinearity with stochastic excitation is used to derive $R_w(t_1, t_2)$ (see Chapter 13 of Ref. 3). The limiter characteristic is

$$l(x) = \frac{1}{2\pi j}\left[\int_{C+} f_+(\omega) \exp(x\omega)\, d\omega + \int_{C-} f_-(\omega) \exp(x\omega)\, d\omega\right] \tag{2}$$

where

$$f_+(\omega) = \int_0^{+\infty} l(x) \exp(-\omega x)\, dx = \frac{l}{\omega}, \qquad \text{for} \quad \text{Re}[\omega] > 0$$

and

$$f_-(\omega) = \int_{-\infty}^{0} l(x) \exp(-\omega x)\, dx = \frac{l}{\omega}, \qquad \text{for} \quad \text{Re}[\omega] < 0.$$

The variable $\omega = u + jv$ is complex with $\text{Re}[\omega] = u$. The contours C_+ and C_- are taken parallel to the v axis in the ω plane with $\text{Re}[\omega] > 0$ for C_+ and $\text{Re}[\omega] < 0$ for C_-. For convenience $l(x)$ is written symbolicly as

$$l(x) = \frac{1}{2\pi j}\int_C f(\omega) \exp(x\omega)\, d\omega \tag{3}$$

where equation (3) means the same as equation (2) when C_+ and C_- are not the same contours.

Since $w(t) = \sin(\omega_c t + \phi)\cdot l[x(t)]$, the autocorrelation function of $w(t)$ is

$$R(t_1, t_2) = \left(\frac{1}{2\pi j}\right)^2 \int_C f(\omega_1) \int_C f(\omega_2) E\{\sin(\omega_c t_1 + \phi)\cdot \exp(\omega_1 s_1 + \omega_1 n_1) \cdot\sin(\omega_c t_2 + \phi)\cdot \exp(\omega_2 s_2 + \omega_2 n_2)\}\, d\omega_1\, d\omega_2 \tag{4}$$

where $s_i = s(t_i)$ and $n_i = n(t_i)$, $i = 1, 2$. The order of complex integration and the expectation operation have been interchanged to get equation (4). For the assumed statistical independence of $n(t)$, $\theta(t)$, and ϕ, the expected value in equation (4) factors into

$$E\{\sin(\omega_c t_1 + \phi)\cdot\exp(\omega_1 s_1)\cdot\sin(\omega_c t_2 + \phi)\cdot\exp(\omega_2 s_2)\}$$
$$\cdot\exp\tfrac{1}{2}[\sigma^2\omega_1^2 + 2R_n(\tau)\omega_1\omega_2 + \sigma^2\omega_2^2] \tag{5}$$

where $\tau = t_2 - t_1$. The form for the cross correlation function $E\{\exp(\omega_1 n_1)\exp(\omega_2 n_2)\}$ where $n(t)$ is stationary gaussian noise has been used in equation (5) (see pp. 476–477 of Ref. 4).

For the case where $s(t)$ is narrow band-limited with respect to ω_c, the filter in Fig. 2 is a narrow bandpass filter, and is assumed to be symmetrical about ω_c. Then $n(t)$ can be written as (see pp. 373–374 of Ref. 4)

$$n(t) = x_c \cos\omega_c t - x_s \sin\omega_c t$$

where x_c and x_s are statistically independent stationary gaussian random processes, and

$$R_n(\tau) = R_\nu(\tau)\cos\omega_c\tau \tag{6}$$

where $R_\nu(\tau) = R_{x_c}(\tau) = R_{x_s}(\tau)$. For a narrow bandpass IF filter, the transform of $R_\nu(\tau)$ is lowpass with a narrow bandwidth compared to ω_c.

With the substitution of

$$t_1 = t, \qquad t_2 = t + \tau,$$
$$\phi^* = \phi + \omega_c t,$$
$$\sin\phi^* = \frac{\exp(j\phi^*) - \exp(-j\phi^*)}{2j},$$

and

$$\exp[R_n(\tau)\omega_1\omega_2] = \sum_{m=-\infty}^{+\infty} I_m(\omega_1\omega_2 R_\nu)\exp(jm\omega_c\tau) \tag{7}$$

(see Article 1, Chapter 3 of Ref. 5), equation (5) becomes

$$(-\tfrac{1}{4})\sum_{m=-\infty}^{+\infty} I_m(\omega_1\omega_2 R_\nu)\exp(jm\omega_c\tau)\cdot E\{[\exp(j\omega_c\tau + j2\phi^*)$$
$$+ \exp(-j\omega_c\tau - j2\phi^*) - \exp(j\omega_c\tau) - \exp(-j\omega_c\tau)]$$
$$\cdot\exp[\omega_1 P\cos(\theta_1 + \phi^*) + \omega_2 P\cos(\theta_2 + \phi^* + \omega_c\tau)]\}. \tag{8}$$

Since $\exp(j\omega_c\tau)$ and $\exp[\omega_2 P \cos(\theta_2 + \phi^* + \omega_c\tau)]$ are periodic in $\omega_c\tau$ with the period 2π, the function

$$\exp(jm\omega_c\tau)E\{\quad\} \tag{9}$$

in equation (8) is periodic in $\omega_c\tau$. Since $R_\nu(\tau)$ transforms to a narrow band-limited lowpass spectrum, the autocorrelation function of $z(t)$ corresponds to the dc component of the Fourier expansion of equation (9). With the substitution of $\delta = \omega_c\tau$, the dc component of equation (9) is

$$\begin{aligned}
&\int_0^{2\pi} \frac{d\delta}{2\pi} \exp(jm\delta)\cdot E\{\quad\} \\
&= \sum_{r=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} I_r(\omega_1 P)\cdot I_k(\omega_2 P)\cdot E_\theta\Bigg\{\int_0^{2\pi}\frac{d\delta}{2\pi} \\
&\cdot\Bigg[E_{\phi^*}\{\exp[j(m+1+k)\delta + j(2+r+k)\phi^* + j(r\theta_1 + k\theta_2)]\} \\
&+ E_{\phi^*}\{\exp[j(m-1+k)\delta + j(-2+r+k)\phi^* + j(r\theta_1 + k\theta_2)]\} \\
&- E_{\phi^*}\{\exp[j(m+1+k)\delta + j(r+k)\phi^* + j(r\theta_1 + k\theta_2)]\} \\
&- E_{\phi^*}\{\exp[j(m-1+k)\delta + j(r+k)\phi^* + j(r\theta_1 + k\theta_2)]\}\Bigg]\Bigg\}.
\end{aligned} \tag{10}$$

Since $\phi^* = \omega_c t + \phi$, ϕ^* has a uniformly distributed probability density function on $[0, 2\pi]$. The averages in equation (10) with respect to δ and ϕ^* follow. For example, the first average with respect to δ and ϕ^* is zero if $\omega + k + 1 \neq 0$ or $k + r + 2 \neq 0$, and when $k = -1 - m$ and $r = -2 - k = m - 1$ the double average is $\exp[(m-1)\theta_1 - (m+1)\theta_2]$. Equation (8) reduces to

$$\begin{aligned}
(-\tfrac{1}{4}) \sum_{m=-\infty}^{+\infty} I_m(\omega_1\omega_2 R_\nu)\Bigg[&I_{m-1}^{(\omega_1 P)} I_{-(m+1)}^{(\omega_2 P)} E\{\exp[j(m-1)\theta_1 - j(m+1)\theta_2]\} \\
&+ I_{m+1}^{(\omega_1 P)} I_{-(m-1)}^{(\omega_2 P)} E\{\exp[j(m+1)\theta_1 - j(m-1)\theta_2]\} \\
&- I_{m+1}^{(\omega_1 P)} I_{-(m+1)}^{(\omega_2 P)} E\{\exp[j(m+1)\theta_1 - j(m+1)\theta_2]\} \\
&- I_{m-1}^{(\omega_1 P)} I_{-(m-1)}^{(\omega_2 P)} E\{\exp[j(m-1)\theta_1 - j(m-1)\theta_2]\}\Bigg].
\end{aligned} \tag{11}$$

The terms in equation (11) for positive and negative m can be combined by noting that $I_{-m}(x) = I_m(x)$. With the substitutions

$$I_m(\omega_1\omega_2 R_\nu) = \sum_{q=0}^{+\infty} \frac{\omega_1^{m+2q}\omega_2^{m+2q}R_\nu^{m+2q}}{2^{m+2q}q!\,\Gamma(m+q+1)} \tag{12}$$

and

$$h_{m,k} = \frac{1}{2\pi j} \int_C f(\omega)\omega^k \exp\left(\frac{\sigma^2\omega^2}{2}\right) I_m(\omega P)\, d\omega, \tag{13}$$

the autocorrelation function of $z(t)$ is

$$\begin{aligned} R_z(t_1, t_2) = \tfrac{1}{4} \sum_{m=0}^{\infty} \sum_{q=0}^{\infty} & \frac{\epsilon_m R_v^{2q+m}}{2^{2q+m} q!\,(q+m)!} \\ & \cdot [h_{m+1,2q+m}^2 R_\theta(m+1, m+1, t_1, t_2) \\ & + h_{m-1,2q+m}^2 R_\theta(m-1, m-1, t_1, t_2) \\ & - h_{m+1,2q+m} h_{m-1,2q+m} R_\theta(m+1, m-1, t_1, t_2) \\ & - h_{m+1,2q+m} h_{m-1,2q+m} R_\theta(m-1, m+1, t_1, t_2)] \end{aligned} \tag{14}$$

where

$$\epsilon_m = \begin{cases} 1, & m = 0 \\ 2, & m > 0 \end{cases}$$

and $R_\theta(A, B, t_1, t_2) = E\{\cos[A\theta(t_1) - B\theta(t_2)]\}$ for any integers A and B.

III. THE CLOSED FORM SOLUTION FOR $h_{m,k}$

The autocorrelation function of $z(t)$ given in equation (14) contains the constants $h_{m,k}$ where $m + k$ are odd integers. For the ideal limiter characteristic of equation (1), there are closed form solutions for these parameters. Since $f_+(\omega) = l/\omega$ for Re $[\omega] > 0$ and $f_-(\omega) = l/\omega$ for Re $[\omega] < 0$, equation (13) becomes

$$\begin{aligned} h_{m,k} = \frac{1}{2\pi j} \int_{C_-} & l\omega^{k-1} I_m(\omega P) \exp\left(\frac{\sigma^2\omega^2}{2}\right) d\omega \\ & + \frac{1}{2\pi j} \int_{C_+} l\omega^{k-1} I_m(\omega P) \exp\left(\frac{\sigma^2\omega^2}{2}\right) d\omega \end{aligned} \tag{15}$$

where C_- is the contour $(-\epsilon - j\infty, -\epsilon + j\infty)$ and C_+ is the contour $(+\epsilon - j\infty, +\epsilon + j\infty)$. By the change of variable $\omega = jx$ and the substitution of $I_m(z) = (j)^{-m} J_m(jz)$, analytic continuation can be applied for $m \geqq 0$ and $k \geqq 0$ to give

$$h_{m,k} = \frac{l}{\pi} (j)^{k+m-1} \int_{-\infty}^{\infty} x^{(k-1)} J_m(xP) \exp\left[\frac{-\sigma^2 x^2}{2}\right] dx. \tag{16}$$

When $m + k$ is even, the integrand of equation (16) is odd and $h_{m,k} = 0$.

When $m + k$ is odd, the integrand of equation (16) is even and

$$h_{m,k} = \frac{2l}{\pi} (j)^{k+m-1} \cdot \frac{\Gamma\left(\frac{m+k}{2}\right)\left(\frac{P^2}{2\sigma^2}\right)^{m/2}}{2\Gamma(m+1)\left[\frac{\sigma}{(2)^{\frac{1}{2}}}\right]^k} \, {}_1F_1\left(\frac{m+k}{2}; m+1; \frac{-P^2}{2\sigma^2}\right) \qquad (17)$$

where a solution has been used for the integral

$$\int_0^\infty x^{k-1} J_m(xP) \exp\left[\frac{-\sigma^2 x^2}{2}\right] dx \qquad (18)$$

in terms of the confluent hypergeometric function ${}_1F_1(\alpha; \beta; -x)$ (see equation A.1.49, p. 1079 of Ref.6). For the case when m and k are non-negative integers ${}_1F_1(m + k/2; m + 1; -x)$ can be expressed in closed form in terms of first and second kind modified Bessel functions. A list of these expressions is given by Middleton (see equation A.1.31, section A 1.2 of Ref. 6). A collection of $h_{m,k}$ in closed form for low order indices is given in Table I. For Table I, $x = P^2/2\sigma^2$ is the input signal-to-noise power ratio into the limiter in Fig. (2).

Any of the $h_{m,k}$ in equation (14) can be found in closed form from Table I by using the recurrence relations

$$h_{m+2,k} = h_{m,k} - \frac{2(m+1)}{P} h_{m-1,k-1} + \frac{4(m+1)m}{P^2} h_{m,k-2}, \qquad (19)$$

$$h_{m+1,k+1} = -\frac{P}{\sigma^2} h_{m,k} - \frac{(k-m-2)}{\sigma^2} h_{m-1,k-1} + \frac{2(k-m-2)m}{\sigma^2 P} h_{m,k-2}, \qquad (20)$$

and

$$h_{m,k+2} = \frac{(m-k)}{\sigma^2} h_{m,k} + \frac{P^2}{\sigma^4} h_{m-2,k} - \frac{(m-k)}{\sigma^4} P h_{m-1,k-1}. \qquad (21)$$

Equation (19) is derived from equation (16) by using the Bessel function identity

$$J_{m+2}(xP) = \frac{2(m+1)}{Px} J_{m+1}(xP) - J_m(xP). \qquad (22)$$

Equation (20) is derived through a by-parts integration of equation (16) and the application of equation (19). Equation (21) is derived through by-parts integration of equation (16). In the development of equations (19), (20) and (21), the integral in equation (16) is re-

TABLE I — CLOSED FORM SOLUTIONS OF SOME $h_{m,k}$

m	k	$h_{m,k}$
1	0	$\frac{lP}{(2\pi)^{\frac{1}{2}}\sigma} e^{-x/2}[I_0(x/2) + I_1(x/2)]$
0	1	$\frac{(2)^{\frac{1}{2}} l}{(\pi)^{\frac{1}{2}}\sigma} e^{-x/2} I_0(x/2)$
2	1	$\frac{-(2)^{\frac{1}{2}} l}{(\pi)^{\frac{1}{2}}\sigma} e^{-x/2} I_1(x/2)$
1	2	$\frac{-lP}{(2\pi)^{\frac{1}{2}}\sigma^3} e^{-x/2}[I_0(x/2) - I_1(x/2)]$
3	2	$\frac{-lP}{(2\pi)^{\frac{1}{2}}\sigma^3} e^{-x/2}\left[I_0(x/2) - \left(1 + \frac{4}{x}\right)I_1(x/2)\right]$
0	3	$\frac{-(2)^{\frac{1}{2}} l}{(\pi)^{\frac{1}{2}}\sigma^3} e^{-x/2}[(1 - x)I_0(x/2) + xI_1(x/2)]$
2	3	$\frac{lP^2}{(2\pi)^{\frac{1}{2}}\sigma^5} e^{-x/2}\left[I_0(x/2) - \left(1 + \frac{1}{x}\right)I_1(x/2)\right]$
4	3	$\frac{-lP^2}{(2\pi)^{\frac{1}{2}}\sigma^5} e^{-x/2}\left[\left(1 + \frac{4}{x} + \frac{12}{x^2}\right)I_1(x/2) - \left(1 + \frac{3}{x}\right)I_0(x/2)\right]$
1	4	$\frac{lP}{(2\pi)^{\frac{1}{2}}\sigma^5} e^{-x/2}[(3 - 2x)I_0(x/2) + (2x - 1)I_1(x/2)]$
3	4	$\frac{-lP^3}{(2\pi)^{\frac{1}{2}}\sigma^7} e^{-x/2}\left[\left(1 + \frac{1}{2x}\right)I_0(x/2) - \left(1 + \frac{3}{2x} + \frac{2}{x^2}\right)I_1(x/2)\right]$
5	4	$\frac{-lP^3}{(2\pi)^{\frac{1}{2}}\sigma^7} e^{-x/2}\left[\left(1+\frac{9}{2x}+\frac{12}{x^2}\right)I_0(x/2)-\left(1+\frac{11}{2x}+\frac{18}{x^2}+\frac{48}{x^3}\right)I_1(x/2)\right]$

stricted to the half interval $[0, \infty)$ which is possible since the integrand of equation (16) is even when $m+k$ is odd.

IV. THE AVERAGE POWER SPECTRUM OF $z(t)$

The autocorrelation function of $z(t)$ given in equation (14) becomes time independent such that $z(t)$ has the average power spectrum $S_z(\omega) = F[R_z(\tau)]$ when $R_\theta(A, B, t_1, t_2) = R_\theta(A, B, \tau)$ for integers A and B. There are some important cases of $\theta(t)$ for which R_θ is time independent.

If θ is a biphase modulation with $\theta(t) = \pm \mid \theta \mid$ that has a zero mean and autocorrelation function (see equation 9-42, section 9-2 of Ref. 4)

$$R_\theta(\tau) = \begin{cases} \mid \theta \mid^2 \left(1 - \dfrac{\mid \tau \mid}{T}\right), & \text{for} \quad \mid \tau \mid \leqq T \\ 0, \quad \text{for} \quad \mid \tau \mid > T \end{cases} \tag{23}$$

then

$$R_\theta(A, B, t_1, t_2) \\ = \cos A \mid \theta \mid \cdot \cos B \mid \theta \mid + \sin A \mid \theta \mid \cdot \sin B \mid \theta \mid \cdot r_\theta(\tau) \tag{24}$$

where $r_\theta(\tau) = R_\theta(\tau)/\mid \theta \mid^2$ is the normalized autocorrelation function of $\theta(t)$. Then R_θ is a function of $\tau = t_2 - t_1$.

For a single tone modulation given by $\theta(t) = m_1 \sin(\omega_1 t + \xi)$ where ξ is a random variable with a uniform probability density function on $[0, 2\pi]$, a simple Bessel series expansion gives

$$R_\theta(A, B, t_1, t_2) = \sum_{n=0}^{\infty} \epsilon_n J_{2n}(Am_1) J_{2n}(Bm_1) \cos(2n\omega_1\tau) \\ + \sum_{n=1}^{\infty} \epsilon_n J_{2n-1}(Am_1) J_{2n-1}(Bm_1) \cos[(2n-1)\omega_1\tau] \tag{25}$$

where

$$\epsilon_n = \begin{cases} 1, & n = 0 \\ 2, & n > 0. \end{cases}$$

For the single tone modulation, R_θ depends only on the time difference τ. If $\theta(t)$ is the sum of tones

$$\theta(t) = \sum_{p=1}^{N} m_p \sin(\omega_p t + \xi_p) \tag{26}$$

where ξ_p, $p = 1, \cdots, N$, are independent random variables with uniform probability density functions on $[0, 2\pi]$, R_θ is again independent of time.

If $\theta(t)$ is a stationary gaussian process with zero mean, variance σ_θ^2 and autocorrelation function $K_\theta(\tau)$, then $R_\theta(A, B, t_1, t_2) = R_\theta(A, B, \tau)$. The second-order characteristic function for the stationary gaussian process is defined as (see equation 112, Chapter 7 of Ref. 4)

$$\begin{aligned}\Phi_\theta(\omega_1, \omega_2; \tau) &= E(\exp\{j[\omega_1\theta(t+\tau) + \omega_2\theta(t)]\}) \\ &= \exp[-\tfrac{1}{2}K_\theta(0)(\omega_1^2 + \omega_2^2) - K_\theta(\tau)\omega_1\omega_2].\end{aligned} \tag{27}$$

Then

$$\begin{aligned}R_\theta(A, B, t_1, t_2) &= \text{Real Part } E\{\exp(jA\theta_1 - jB\theta_2)\} \\ &= \exp\left[-\frac{\sigma_\theta^2}{2}(A^2 + B^2)\right]\cdot\exp[ABK_\theta(\tau)] \\ &= R_\theta(A, B, \tau).\end{aligned} \tag{28}$$

The validity of equation (14) depends on the narrow band-limited assumption for the modulated signal $s(t)$ at the carrier frequency w_c. For $s(t)$ to be narrowband limited, the parameter values that the modulation functions can have are restricted.

V. AN APPLICATION OF THE R_z RESULTS TO THE PERFORMANCE OF A SUBCARRIER CHANNEL

A modulation technique sometimes used for communication is FM-PM where the carrier is phase modulated by a subcarrier that is in turn frequency modulated by the information waveform. The FM-PM signal is of the form

$$s(t) = P\cos\{\omega_c t + \phi + m_1\sin[\omega_1 t + \xi + \lambda(t)]\} \tag{29}$$

where P, ω_c, ω_1 and m_1 are constants, ϕ and ξ are independent random variables usually assumed to have uniform probability density functions over $[0, 2\pi]$, and $\lambda(t)$ is the integral of the information waveform. In a typical application, $\omega_c \gg \omega_1$ and $\lambda(t)$ is slowly varying compared with $\cos\omega_1 t$. With these restrictions the information $\dot{\lambda}(t)$ can be recovered from $s(t)$ with the receiver shown in Fig. 3.

The purpose of the bandpass limiter is to remove the effect of variations that might occur in P. For the ideal case where $s(t)$ is not perturbed by noise, the subcarrier filter input $z(t)$ is

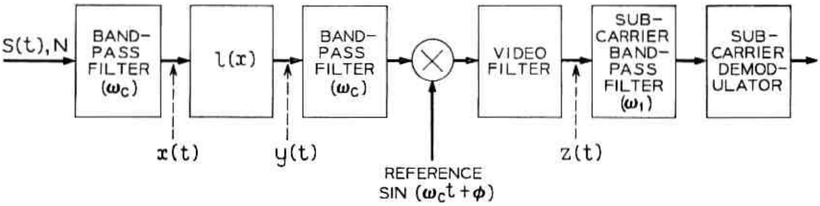

S(t), N
BAND-PASS FILTER (ω_C)
$l(x)$
BAND-PASS FILTER (ω_C)
VIDEO FILTER
SUB-CARRIER BAND-PASS FILTER (ω_1)
SUB-CARRIER DEMODULATOR
$x(t)$
$y(t)$
REFERENCE SIN ($\omega_C t + \phi$)
$z(t)$

Fig. 3 — FM-PM receiver with ideal bandpass limiter.

$$\begin{aligned} z(t) &= -\frac{2l}{\pi} \sin \{m_1 \sin [\omega_1 t + \lambda(t) + \xi]\} \\ &= -\frac{4l}{\pi} \sum_{n=1}^{\infty} J_{2n-1}(m_1) \sin \{(2n - 1)[\omega_1 t + \lambda(t) + \xi]\}. \end{aligned} \tag{30}$$

If $\lambda(t)$ is slowly varying compared with $\cos \omega_1 t$, the information can be recovered with a subcarrier filter that passes only the first component of the sum in equation (30). For the noiseless case the subcarrier filter response is then

$$-\frac{4l}{\pi} J_1(m_1) \sin [\omega_1 t + \lambda(t) + \xi]. \tag{31}$$

After additional processing in a subcarrier demodulator, $\dot{\lambda}(t)$ is obtained from equation (31). One criterion of performance of the receiver is the S/N out of the subcarrier filter as a function of the limiter input S/N, $x = P^2/2\sigma^2$. Since $\lambda(t)$ varies slowly compared with $\cos \omega_1 t$, the output S/N for the subcarrier filter is determined with sufficient accuracy by setting $\lambda(t) \equiv 0$. If $\lambda(t) \equiv 0$, the subcarrier output S/N follows directly from equations (14) and (25). Substitution of equation (25) into equation (14) gives the power spectrum

$$\begin{aligned} S_z(\omega)\Big|_{\theta = m_1 \sin (\omega_1 t + \xi)} &\cong 2h_{10}^2 \sum_{n=1}^{\infty} J_{2n-1}^2(m_1) \cdot F[\cos (2n - 1)\omega_1 \tau] \\ &+ (\tfrac{1}{4})[\sigma h_{01} - \sigma h_{21} J_0(2m_1)]^2 \cdot F[r_v(\tau)] \\ &+ (\tfrac{1}{2})\sigma^2 h_{21}^2 \sum_{n=1}^{\infty} J_n^2(2m_1) \cdot F[r_v(\tau) \cdot \cos (n\omega_1 \tau)] \\ &+ (\tfrac{1}{2})\sigma^4 h_{12}^2 \sum_{n=1}^{\infty} J_{2n-1}^2(m_1) \cdot F[r_v^2(\tau) \cdot \cos (2n - 1)\omega_1 \tau] \\ &+ (\tfrac{1}{16}) \sum_{n=0}^{\infty} \epsilon_n[\sigma^2 h_{12} J_n(m_1) - \sigma^2 h_{32} J_n(3m_1)]^2 \cdot F[r_v^2(\tau) \cdot \cos n\omega_1 \tau] \end{aligned}$$

$$
\begin{aligned}
&+ (\tfrac{1}{32})[\sigma^3 h_{03} - \sigma^3 h_{23} J_0(2m_1)]^2 \cdot F[r_v^3(\tau)] \\
&+ (\tfrac{1}{16})\sigma^6 h_{23}^2 \sum_{n=1}^{\infty} J_n^2(2m_1) \cdot F[r_v^3(\tau) \cdot \cos n\omega_1 \tau] \\
&+ (\tfrac{1}{96}) \sum_{n=0}^{\infty} \epsilon_n [\sigma^3 h_{23} J_n(2m_1) - \sigma^3 h_{43} J_n(4m_1)]^2 \cdot F[r_v^3(\tau) \cdot \cos n\omega_1 \tau] \\
&+ (\tfrac{1}{32})\sigma^8 h_{14}^2 \sum_{n=1}^{\infty} J_{2n-1}^2(m_1) \cdot F[r_v^4(\tau) \cdot \cos (2n - 1)\omega_1 \tau] \\
&+ (\tfrac{1}{192}) \sum_{n=0}^{\infty} \epsilon_n [\sigma^4 h_{14} J_n(m_1) - \sigma^4 h_{34} J_n(3m_1)]^2 \cdot F[r_v^4(\tau) \cdot \cos n\omega_1 \tau] \\
&+ (\tfrac{1}{768}) \sum_{n=0}^{\infty} \epsilon_n [\sigma^4 h_{34} J_n(3m_1) - \sigma^4 h_{54} J_n(5m_1)]^2 \cdot F[r_v^4(\tau) \cdot \cos n\omega_1 \tau] \qquad (32)
\end{aligned}
$$

where $r_v = R_v/R_v(0) = R_v/\sigma^2$. The approximation, equation (32), neglects all the terms of equation (14) containing the factor R_v^{2q+m} where $2q + m > 4$. The terms in equation (32) are the significant terms of $S_z(\omega)$ for the single tone modulation. The spectrum in equation (32) is the weighted sum of terms of the form

$$F[r_v^n(\tau) \cos m\omega_1 \tau] = \frac{1}{2\pi} F[r_v^n(\tau)] * F[\cos m\omega_1 \tau] \qquad (33)$$

where * is the convolution operation. Since $F[\cos m\omega_1 \tau]$ is a pair of impulses of weight π at $\pm m\omega_1$,

$$F[r_v^n(\tau) \cos m\omega_1 \tau] = \tfrac{1}{2}[S_{v,n}(\omega + m\omega_1) + S_{v,n}(\omega - m\omega_1)] \qquad (34)$$

where

$$S_{v,n}(\omega) = F[r_v^n(\tau)].$$

The first term in the spectrum of equation (32) is the signal content of $z(t)$. All other terms of equation (32) correspond to noise alone or a combination of signal and noise. All terms of equation (32) except the first term are usually combined to give the interference (noise) spectrum at the output of the video filter.

A computation was made for the subcarrier filter output S/N as a function of the input S/N x. The following conditions are assumed for the computation.

(*i*) The power spectrum of the input gaussian noise to the cascade in Fig. 1 is uniform over the bandwidth of the prelimiter bandpass filter.

(*ii*) The prelimiter bandpass filter is assumed to have a gaussian transfer function such that

$$r_v(\tau) = \exp\left[\frac{-\tau^2\omega_0^2}{\pi}\right]. \tag{35}$$

(*iii*) The subcarrier amplitude transfer function is

$$|H(j\omega)| = \begin{cases} 1, & \omega_1 - \dfrac{\Delta\omega}{2} < |\omega| < \omega_1 + \dfrac{\Delta\omega}{2} \\ 0, & \text{all other } \omega, \end{cases} \tag{36}$$

where $\Delta\omega \ll \omega_0$. Also, $\omega_0 = 12.566 \times 10^6$ and $\omega_1 = 6.434 \times 10^6$ are assumed. Substitution of equation (35) into equation (34) gives

$$S_{v,n}(\omega) = F[r_v^n(\tau)] = \frac{\pi}{\omega_0(n)^{\frac{1}{2}}} \exp\left[-\frac{\pi}{4n}\left(\frac{\omega}{\omega_0}\right)^2\right]. \tag{37}$$

From condition *iii*, the noise spectrum in the passband of the subcarrier filter is approximately constant when $\omega = \omega_1$. The signal and noise powers out of the subcarrier filter follow from $S_z(\omega_1)$. The signal power is $2h_{10}^2J_1^2(m_1)$; the noise power is

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} [S_z(\omega) - 2h_{10}^2J_1^2(m_1)\cdot F(\cos\omega_1\tau)]\cdot|H(j\omega)|^2\, d\omega \cong [S_z'(\omega_1)]\cdot 2\,\Delta f$$

where Δf is the width of the subcarrier filter and where S_z' is equation (32) with the first term omitted. The function

$$S(m_1) = \frac{2h_{10}^2J_1^2(m_1)}{x[S_z'(\omega_1)]} \tag{38}$$

was computed for x between 0.01 and 100 with m_1 as a parameter. The results of the computation are shown in Fig. (4). For a given m_1 and x, the output S/N for the subcarrier filter is $x/2\Delta f\cdot S(m_1)$.

VI. SUMMARY

A general, second order statistical analysis is presented for the cascade of a narrow bandpass limiter, an ideal phase detector, and a video filter. In this analysis, the input to the limiter is assumed to be the sum of a stationary gaussian noise and a fixed amplitude phase modulated sine wave. The autocorrelation function of the cascade response is obtained as a function of the signal-to-noise ratio x at the limiter input, the normalized autocorrelation function of the lowpass equivalent for the limiter input noise $r_v(\tau)$, and the phase modulation $\theta(t)$.

The cascade response $z(t)$ has the autocorrelation function $R_z(t_1, t_2)$

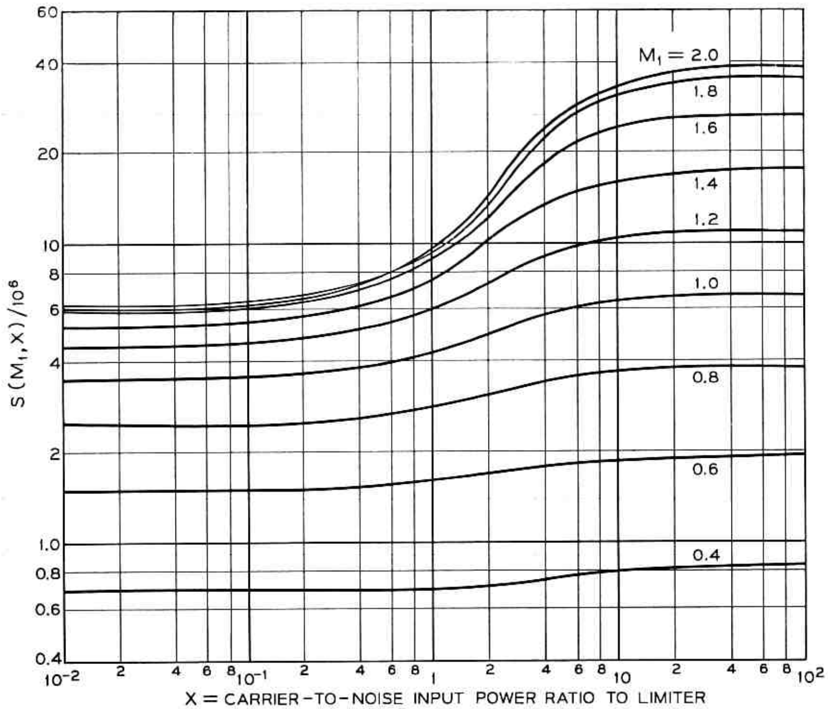


Fig. 4 — The unit bandwidth subcarrier filter output, S/N normalized by x where $10^{-2} \leqq x \leqq 10^2$ and M_1 is a parameter.

that can be time dependent. However, for some important cases of $\theta(t)$, $R_z(t_1, t_2) = R_z(\tau)$, and the cascade response has the average power spectrum $S_z(\omega) = F[R_z(\tau)]$ where F is the Fourier transform operation with respect to τ. The cases of $\theta(t)$ considered that yield $R_z(\tau)$ are the random biphase waveform $\theta = \pm |\theta|$, the single tone $\theta(t) = m_1 \sin(\omega_1 t + \xi)$, and the stationary gaussian process with autocorrelation function $K_\theta(\tau)$.

The dependence of $R_z(t_1, t_2)$ on the limiter input S/N appears in the h parameters. These parameters can be obtained in closed form as functions of the modified Bessel functions $I_0(x/2)$ and $I_1(x/2)$. The lower order h parameters encountered in the first few terms of the series for R_z are found, and recurrence relations are derived through which higher order h parameters can be derived easily.

For the modulation types that make R_z a function of τ alone, the power spectrum $S_z(\omega)$ is known for all values of the limiter input S/N x. Then the S/N can be derived in any frequency band at the output of the video filter in Fig. 1 as a function of any S/N into the limiter.

The performance of a subcarrier channel was considered where $\theta(t) = m_1 \sin [\omega_1 t + \lambda(t) + \xi]$. The subcarrier was assumed to be phase modulated by a narrowband low pass process $\lambda(t)$. The S/N at the output of the subcarrier filter was obtained by computation of the approximation of equation (32). For this example, a gaussian prelimiter bandpass filter was assumed. For this filter shape, $r_v^n(\tau)$ and its transform $S_{v,n}(\omega)$ are gaussian for all integers n. Some representative parameters from the Apollo unified S-band communication system[1] were assumed. These were

(*i*) A prelimiter noise equivalent bandwidth of 4 MHz.

(*ii*) A subcarrier frequency of 1.024 MHz.

(*iii*) A subcarrier noise equivalent bandwidth of 0.2 MHz.

(*iv*) An input S/N range of $0.01 \leqq x \leqq 100$.

(*v*) A set of modulation indices $m_1 = (0.2)k$, $k = 2, 3, 4, 5, 6, 7, 8, 9, 10$.

The results are given in Fig. 4. The differential between subcarrier filter output S/N at low and high values of x is a monotonically increasing function of m_1 for $0.4 \leqq m_1 \leqq 2.0$. The shapes of the curves are similar to that of the $(S/N)_0/(S/N)_i$ curve obtained by Davenport.[2]

VII. ACKNOWLEDGMENT

The author is indebted to R. L. Selden for his comments on this work and to S. O. Rice for his suggestions which gave a concise derivation of equation (14).

REFERENCES

1. National Aeronautics and Space Administration, "Proceedings of the Apollo Unified S-Band Technical Conference," Goddard Space Flight Center, NASA SP-87, July 14–15, 1965.
2. Davenport, W. B., Jr., "Signal-to-Noise Ratios in Bandpass Limiters," Journal of Applied Physics, *24,* No. 6 (June 1953), pp. 720–727.
3. Davenport, W. B., Jr., and Root, W. L., *An Introduction to The Theory of Random Signals and Noise,* New York: McGraw-Hill Book Company, Inc., 1958.
4. Papoulis, Athanasios, *Probability, Random Variables and Stochastic Processes,* New York: McGraw-Hill Book Company, Inc., 1965.
5. Magnus, W. and Oberhettinger, F., *Special Functions of Mathematical Physics,* New York: Chelsen, 1949.
6. Middleton, D., *An Introduction to Statistical Communication Theory,* New York: McGraw-Hill Book Company, Inc., 1960.

Multiplex *Touch-Tone*® Detection Using Time Speed-Up

By J. F. O'NEILL

(Manuscript received August 19, 1968)

A signal may be read from a storage medium faster than the rate that would correspond to real time reconstruction of the signal; this process has been named time compression or time speed-up. Cheap serial shift registers make time speed-up an attractive means to detect **Touch-Tone**® *calling (or other format) signals on a multiplicity of channels using a single detector.*

I. BACKGROUND

Time speed-up (TSU) of a signal consists of reading the signal from a store faster than the rate at which it was recorded. (This is generally faster than real time reconstruction of the signal, thus the name). I propose this process for multiplexing several voiceband channels in time, so that one multifrequency receiver can detect *Touch-Tone*® signaling on a multiplicity of channels.

Processing of a single signal using TSU configurations based on electric or acoustic delay lines (called DELTIC systems, for delay line time compression) has been done since the 1950's.[1,2] At Bell Laboratories, TSU is being investigated for multiplexing *Picturephone*® visual telephone channels on slightly nonlinear microwave radio systems.[3]

The inherent simplicity and versatility of a digital TSU signal processing system is enhanced by the availability of inexpensive serial shift registers based on the insulated gate field effect transistor. These registers typically store 64 bits, and are sufficiently fast to permit a single detection circuit to serve eight to 16 *Touch-Tone* voiceband signaling channels or hundreds of channels in a low frequency application, such as 20 Hz ringing detection.

The attractiveness of TSU multiplex tone detection is demonstrated by, and most of this article treats of, the *Touch-Tone* detection case.

If the system is realized as a digital multiplexer (concentrator) in tandem with an analog frequency detector, it will become apparent that the frequency detector can (but need not) be exactly the type of circuit now used, but with a scaling applied to all reactances. This scaling is to raise all spectral features by a constant factor which is the ratio of time compression. Thus, all the linear and nonlinear signal processing now used can be included in a TSU system and its performance would simulate that of present *Touch-Tone* receivers. By adding additional data smoothing, which could involve using the signal samples more than once, the present tolerance to digit simulation by speech can be exceeded.

II. SIXTY-FOUR CHANNEL TSU RINGING DETECTOR

An exploratory key telephone system must detect the presence of 20 Hz ringing on 64 central office lines. This detection could be performed on each channel, but the availability of 64 bit serial shift registers has made centralized TSU detection economically more attractive.

Figure 1 shows the TSU arrangement to be used in this exploratory system. A transducer V_i, $i = 1, 2 \ldots 64$ at each channel slices (limits) the ringing signal and presents a rectangular wave at logic level to the sampling gates S_{ai}. Binary data is sufficient to specify the input signal because it is basically a single tone; there are interfering tones from power line cross-coupling but are suppressed to a large extent by the larger 20 Hz signal and the limiting operation. (It will be apparent that a multitone format such as *Touch-Tone* signaling would not be well represented by binary coded signal samples.) The sampling gates load the long serial register SR1 with a sequence of samples V_x from all the channels. The order of the samples is the same as the order of the channels: $\ldots, 1, 2, \ldots 63, 64, 1, 2, \ldots$.

The register SR1 has taps every 64 bits, however, and at these m taps (including the input and output) the samples at any instant are all for the same channel, as shown by V_y. These samples can be processed in a high speed detector, and the result registered in either a common or per-channel answer depository. A digital detector, for instance, could examine the m samples in a time consisting of a few logic gate delays. Alternatively, the m samples can be placed in an independent register SR2 as shown in Fig. 1, from which they can be clocked into an analog frequency discriminator of any type, such as a two-pole resonator. With this system, the SR2 read-out clock is in-

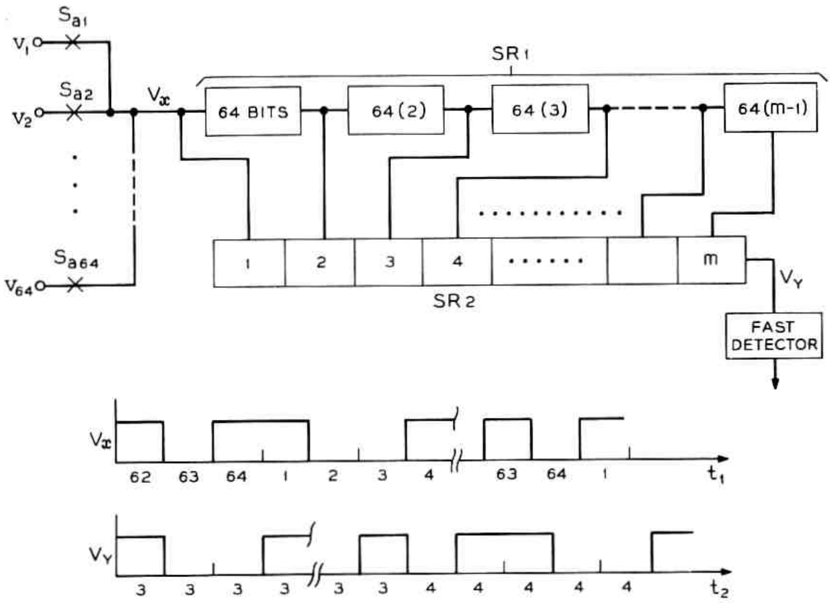


Fig. 1 — 64 Channel TSU single frequency detector.

herently a part of the detection process, since it controls the ratio of time speed-up; if the channels are sampled at rate f_s per second and the SR2 register is read at kf_s per second, then k is the ratio of time compression (and spectral expansion).

Modifications to the Fig. 1 TSU tone detector permit detection of a single tone of unspecified frequency. To do this, more detectors could be added at the output of SR2. The same result would be attained by using only one detector with various clock rates to read out SR2, and a return path from SR2 output to input, so that the samples for a particular channel could be processed repeatedly.

In the exploratory key telephone ringing detector the 64 channels are sampled at seven times per cycle of the input 20 Hz wave and a digital detector is used to examine the samples from one cycle ($m = 7$). The detector stores this tentative result in another serial shift register, and when enough 50 msec intervals appear to have ringing present, a RING output is delivered to the common controller. This detection operation is low Q, but this is by design, and is not dependent on either the TSU structure or the technology. An analog detector in this system would need $m \gg 7$ but would not require the added integration. As always, an appropriate trade-off between selectivity (high Q) and fast detection (low Q) must be made.

The effectiveness of the detector in suppressing frequencies other than 20 Hz is a function of the sampling rate, as well as the parameters of the detector. Waveform preservation is not necessary for detection, so the sampling theorem requirement (two samples per cycle at the highest frequency of interest) need not be met. As few samples as possible should be handled to conserve storage, but the lower limit is set by the signal duty cycle variation and the size and frequency of the interfering signals. The equivalence between periodic sampling and modulation permits intelligent selection of the sampling rate.

III. EIGHT CHANNEL TSU *Touch-Tone* DETECTOR

The TSU configuration of Fig. 1 could be adapted to multifrequency detection by means of a few additions. First, incorporate at the register SR1 input an analog to digital converter to code the signal samples sufficiently accurately to preserve the information content, say b bits per sample. Replace shift register SR1 by b parallel shift registers, one for each bit at the analog-digital output. Finally, add a digital to analog converter at the output of the b parallel read-out registers (SR2 in Fig. 1). No change in principle is involved; the added circuitry only preserves the signal amplitude through the TSU system. A delta coder with a (longer) single shift register could be used for the digitizing operation; the type of code is a detail.

However, the structure of Fig. 1 is not well suited to *Touch-Tone* detection. The serial registers are conveniently available in 64 bit and larger sizes. (Smaller sizes would be economically wasteful; adding taps increases the lead count perilously.) Only seldom is there a need to detect 64 *Touch-Tone* signals simultaneously, and reliability requirements would be excessively difficult, even if the need existed. By using the Fig. 2 TSU configuration, the 64 bit registers are used very efficiently.

In Fig. 2, each channel has a private b-register store. The channel $i (i = 1, 8)$ inputs are sampled in multiplex by switches S_{ai} and coded by a common analog-to-digital converter. The coded samples are steered by logic gates S_{bi} to the registers for channel i. Sometime between (or synchronized with) input samples, the registers are read at high speed into the digital-to-analog converter, which is assumed to be simple enough to build for each channel. Transmission gate S_{ci}

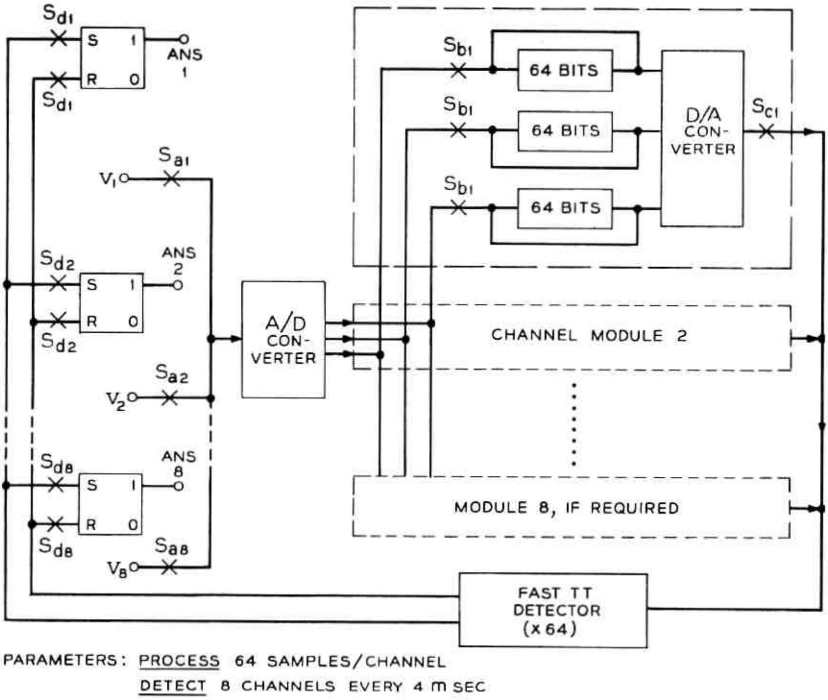


Fig. 2 — Eight channel TSU *Touch-Tone*® dialing detector.

simultaneously connects the (per channel) converter output to the detector input bus, so that the detector is time shared by all channel circuits. This detector can be a carbon copy of any of the standard receivers, but with all reactances scaled up in frequency by the time compression ratio. Or, it could be all-digital. In either case, the readout of the channel register bank must be sufficiently fast to permit the detector to answer and return to quiescence before the next channel is examined.

An important feature of the Fig. 2 parallel register TSU system is that the channel registers need be supplied only for as many channels as are actually required. The Fig. 1 serial system must be built entirely in order to operate at all.

An 8-channel *Touch-Tone* receiver using 3 bit ($b = 3$) coding has been built and operated by Mr. R. J. Violet of Bell Telephone Laboratories. In this demonstration system, the channel sampling is done at 4000 Hz with 64 samples being stored per channel. Each channel is examined every 4 ms, and the detection requires 0.5 ms. Half of this 0.5 ms is to allow the receiver to become quiescent. The time compression ratio is the sampling interval divided by the sample read time, or 64. The detector is thus constructed for input frequencies at 64 times the normal *Touch-Tone* frequencies. This simple demonstration system immediately registers the detected results through gates S_{di} in a per-channel flip-flop bank (shown in skeletal form). An attractive feature of TSU detection is that further processing, such as delay or data format conversions, can be made by common equipment. Thus, *Touch-Tone* signal to dial pulse translators for conversion of step-by-step switching machines could be very effectively built using TSU multiplexing.

If the input signal can vary considerably in amplitude, either a per-channel automatic gain control or more accurate sample coding would be required to preserve the signal waveform through coding and decoding. Also, a sampling rate higher than 4000 Hz and a larger number of samples per detection might be used in a production circuit. In compensation, a rate of more than 8 channels is within the speed capability of the circuitry; additional signal integration to improve the tolerance to digit simulation is easy to incorporate.

The economic advantages of large scale production can be gained through the use of 64 bit serial shift registers in many of the digital systems. Preliminary economic analysis indicates that the marginal cost of one *Touch-Tone* detector in a TSU multiplex system would be less than the equivalent single channel receiver; a cost crossover can be expected at about three channels. In comparison with multiplex receiving based on digital filtering, TSU offers easier maintenance, per-channel modularity, and the ability to incorporate future improvements in the detector circuitry.

REFERENCES

1. Anderson, V. C., "Delay Line Time Compressor," Patent No. 2,958,039, filed May 18, 1956.
2. Allen, W. B., "Series-Parallel Recirculation Time Compressor," Patent No. 3,274,341, filed December 17, 1962.
3. Heightley, J. D. and Perneski, A. J., "An 800 nsec Plated-Wire Store for a Time Compression Multiplex Transmission System," Int. Solid State Circuits Conf., Digest of Technical Papers, February 14, 1968, pp. 112, 113.

Data Transmission Error Probabilities in the Presence of Low-Frequency Removal and Noise

By B. R. SALTZBERG and M. K. SIMON

(Manuscript received July 19, 1968)

Upper bounds on error probability are derived for data transmission systems which are subjected to gaussian noise and to the removal of the low-frequency components of the signal. This error probability can be quite low for random data, even though the eye pattern is closed. Both standard format and partial response signaling are considered, as are binary and multilevel alphabets. Numerical results are given for a high-pass filter containing a single pole and for a cascade of several such identical filters.

I. INTRODUCTION

It is frequently desirable, or unavoidable, that the low-frequency components of a data signal be eliminated. This may occur through the use of capacitor or transformer coupling in the terminal equipment or in the baseband transmission facilities. Another instance results from the necessity of removing low-frequency baseband components before modulation in order to provide a spectral guard band in the vicinity of the carrier frequency.

Since dc is usually completely attenuated, no linear operation can correct for low-frequency removal. One commonly used approach uses nonlinear feedback to restore the low-frequency components.[1] Another solution to this problem involves dc-free signal formats.[2, 3]

We evaluate the penalty resulting from the removal of low-frequency components from a standard format data signal (Nyquist I shaping) and a partial response signaling format (multilevel extension of duobinary with precoding.)[4] Clearly, in both of these cases, the degradation is most severe when the transmitted data sequence contains long strings of identical digits. In fact, when the system bandwidth is less than the signaling rate, which is usual in data communication systems, the received signal will be zero. This follows from the fact that for a periodic input impulse train the lowest frequency components are at dc and the signaling frequency, both of which are filtered out. However, the degradation of a random signal can be quite small when the cutoff frequency of the offending high pass filter is far below the signaling rate.

We consider binary and multilevel data-transmission systems with signaling formats as above, degraded by a single-pole high-pass filter or a cascade of such filters. The systems are evaluated for error probability in the presence of additive gaussian noise. A previously derived error probability bound[5] is used, which takes the form of a gaussian distribution of the signal to noise ratio, in which the larger intersymbol interference components subtract from the signal amplitude and the smaller ones add to the noise power.[5] In general, the optimum splitting of intersymbol interference terms between signal amplitude and noise power cannot be determined analytically. We show that for intersymbol interference components, related by a single exponential damping factor, an optimum subdivision can be explicitly specified. Where the eye is open, the error probability bound is given directly in terms of the eye opening to rms noise ratio.

We also discuss the refinements of the generalized bound in the case of interysmbol interference from a single exponential signal tail, and then apply the results to Nyquist I shaped and partial response signaling formats respectively. Single poles and a cascade of identical poles are considered, and numerical results are given for practical data system parameters.

II. DERIVATION OF A SIMPLIFIED ERROR PROBABILITY BOUND FOR SINGLE EXPONENTIAL INTERSYMBOL INTERFERENCE

Reference 5 gives an upper bound for the probability of error in the reception of a random digital message perturbed by gaussian noise and intersymbol interference. This gives

$$P_e \leqq A \exp\left\{-\frac{[f_0 - (N-1)\sum_{k \epsilon K} f_k]}{\left[2\sigma_n^2 + \frac{N^2-1}{3}\sum_{k \notin K} f_k^2\right]}\right\} \tag{1}$$

which is subject to

$$\sum_{k \epsilon K} f_k < \frac{f_0}{N-1}$$

where

N is the number of levels of the input random message.
σ_n^2 is the variance of the additive noise.
$f(t)$ is the signaling waveform.

$\frac{1}{T}$ is the signaling rate.

$$f_k = \begin{cases} |\, f(kT) \,| & \text{for standard format signaling} \\ |\, f[(k - \frac{1}{2})T] \,| & \text{for } N \text{ level partial response signaling with precoding*} \end{cases}$$

and

$$A = \begin{cases} \dfrac{2(N-1)}{N} & \text{for standard format signaling} \\ \dfrac{2(N^2-1)}{N^2} & \text{for } N \text{ level partial response signaling with precoding} \end{cases}$$

We notice that the applicability of the error probability bound to partial response signaling formats was not discussed in the original paper but is presented here as a further extension of the result.[5]

The sets $k \varepsilon K$ and $k \not\varepsilon K$ include all members except $k = 0$. It is also shown in Ref. 5 that

$$\underset{\ell \varepsilon K}{f_\ell} > \underset{m \not\varepsilon K}{f_m} \tag{2}$$

Thus, if the signal sample set $\{f_k\}$ excluding $k = 0$ is rearranged in order of decreasing magnitude to form a set $\{g_k\}$, then the sums in equation (1) may be replaced by

$$\begin{aligned} \sum_{k \varepsilon K} f_k &= \sum_{k=1}^{M} g_k \\ \sum_{k \not\varepsilon K} f_k^2 &= \sum_{k=M+1}^{\infty} g_k^2 \,. \end{aligned} \tag{3}$$

For an arbitrary signaling waveform, $f(t)$, the optimum M [in the sense of minimizing the right side of equation (1)] must be determined by a trial comparison method as decribed in Ref. 5.

* In the partial response case, f_1 must be replaced by f_1—f_0 in both numerator and denominator summations of equation (1) since only the unintentional intersymbol interference should be included there.

For an exponential signal tail,

$$f_k = rf_{k-1}, \qquad 0 < r < 1; \quad k = 2, 3, \cdots \tag{4}$$

Thus, since $f(t)$ is already monotonically decreasing for all $t \geqq T$, the ordered sets $\{f_k\}$ and $\{g_k\}$ are identical in this case.

$$\begin{aligned} \sum_{k=1}^{M} f_k &= \frac{f_1[1 - r^M]}{1 - r} \\ \sum_{k=M+1}^{\infty} f_k^2 &= \frac{f_1^2 r^{2M}}{1 - r^2}. \end{aligned} \tag{5}$$

To minimize the right side of equation (1), it is sufficient to maximize

$$Q = \frac{[f_0 - (N-1) \sum_{k \in K} f_k]^2}{2\left[\sigma_n^2 + \dfrac{N^2-1}{3} \sum_{k \notin K} f_k^2\right]} = \frac{\left[f_0 - (N-1)\dfrac{f_1(1-r^M)}{1-r}\right]^2}{2\left[\sigma_n^2 + \dfrac{N^2-1}{3} f_1^2 \dfrac{r^{2M}}{1-r^2}\right]}. \tag{6}$$

Differentiating Q with respect to M gives

$$\begin{aligned} \frac{dQ}{dM} = {} & x \ln r \left[f_0 - \frac{(N-1)}{1-r} f_1(1-x)\right]\left[f_0 - \frac{(N-1)}{1-r} f_1\right]\left[\frac{f_1^2}{1-r^2}\right] \\ & \cdot \left[\frac{\sigma_n^2(N-1)}{\left(\dfrac{N^2-1}{3}\right)\left(\dfrac{f_1}{1+r}\right)\left(f_0 - \dfrac{(N-1)f_1}{1-r}\right)} - x\right]\left[\frac{N^2-1}{3}\right] \\ & \Big/ \left[\sigma_n^2 + \frac{N^2-1}{3} \frac{f_1^2}{1-r^2} x^2\right]^2 \end{aligned} \tag{7}$$

where

$$x = r^M \qquad (0 \leqq x \leqq 1)$$

and

$$\frac{(N-1)}{1-r} f_1(1-x) < f_0 . \tag{8}$$

Three separate cases must now be examined.

(*i*) If $f_0 - (N-1)f_1/(1-r) < 0$, then the eye is closed. From equation (7) it follows that $dQ/dM < 0$ for $0 < x \leqq 1$. Therefore the positive maximum of Q occurs at the boundary $x = 1$, so the optimum value of M is $M_{\text{opt}} = 0$.

(*ii*) If

$$\frac{\sigma_n^2(N-1)}{\left(\frac{N^2-1}{3}\right)\left(\frac{f_1}{1+r}\right)\left[f_0-\frac{(N-1)f_1}{1-r}\right]} > 1,$$

it is implicit that $f_0 - (N-1)f_1/(1-r) > 0$, and the eye is open. In this case it is again true that $dQ/dM < 0$ for $0 < x \leqq 1$, and $M_{\text{opt}} = 0$.

(*iii*) If

$$0 < \frac{\sigma_n^2(N-1)}{\left(\frac{N^2-1}{3}\right)\left(\frac{f_1}{1+r}\right)\left[f_0-\frac{(N-1)f_1}{(1-r)}\right]} < 1,$$

it is again implicit that $f_0 - (N-1)f_1/(1-r) > 0$, and the eye is open. In this case a positive maximum for Q occurs in the interval $0 < x < 1$. Solving for the point where $dQ/dM = 0$, we obtain

$$r^{M_{\text{opt}}} = \frac{\sigma_n^2(N-1)}{\left(\frac{N^2-1}{3}\right)\left(f_0-\frac{(N-1)f_1}{1-r}\right)\left(\frac{f_1}{1+r}\right)}. \tag{9}$$

Notice that condition (8) is automatically satisfied.

Since the solution for M_{opt} as given by equation (9) is not necessarily integer, the error probability bound as given by equation (1) must be modified in terms of the actual choice of an integer M. We will arbitrarily use the next higher integer. Letting $[M_{\text{opt}}]$ denote the next higher integer to M_{opt} and

$$z = 3\left(\frac{N-1}{N+1}\right)\left(\frac{1+r}{1-r}\right),$$

equation (6) may be expressed as:

$$Q = \frac{[S_p - I_{\max}(1-r^{[M_{\text{opt}}]})]^2}{2[\sigma_n^2 + I_{\max}^2 r^{2[M_{\text{opt}}]}/z]} \tag{10}$$

where

$$\begin{aligned} r^{[M_{\text{opt}}]} &= b\,\frac{\sigma_n^2 z}{I_{\max}(S_p - I_{\max})}\,; \\ b &= r^{[M_{\text{opt}}]-M_{\text{opt}}}, \qquad r < b < 1. \end{aligned} \tag{11}$$

$I_{\max} = (N-1)f_1/(1-r)$ denotes the maximum intersymbol interference and $S_p = f_0$ denotes the signal amplitude. Combining equations (10) and (11),

$$2Q = \frac{\left\{S_p - \left[\dfrac{I_{\max}(S_p - I_{\max}) - b\sigma_n^2 z}{S_p - I_{\max}}\right]\right\}^2}{\sigma_n^2 + \dfrac{b^2\sigma_n^4 z}{(S_p - I_{\max})^2}} = \frac{[(S_p - I_{\max})^2 + b\sigma_n^2 z]^2}{\sigma_n^2[(S_p - I_{\max})^2 + b^2\sigma_n^2 z]}.$$

Since $r < b < 1$,

$$2Q > \frac{[(S_p - I_{\max}) + b\sigma_n^2 z]^2}{\sigma_n^2[(S_p - I_{\max})^2 + b\sigma_n^2 z]} = \left(\frac{S_p - I_{\max}}{\sigma_n}\right)^2 + bz > \left(\frac{S_p - I_{\max}}{\sigma_n}\right)^2 + rz.$$

In terms of the error probability,

$$\begin{aligned} P_e &< A \exp[-Q] \\ &< A \exp\left\{-\frac{\left(\dfrac{S_p - I_{\max}}{\sigma_n}\right)^2 + 3r\left(\dfrac{N-1}{N+1}\right)\left(\dfrac{1+r}{1-r}\right)}{2}\right\}. \end{aligned} \tag{12}$$

For the situations where $M_{\text{opt}} = 0$ (that is, cases *i* and *ii*) equation (1) becomes

$$P_e < A \exp\left\{-\frac{S_p^2}{2\left[\sigma_n^2 + \dfrac{1}{3}\left(\dfrac{N+1}{N-1}\right)\left(\dfrac{1-r}{1+r}\right)I_{\max}^2\right]}\right\}. \tag{13}$$

III. ERROR PROBABILITY PERFORMANCE WITH A STANDARD FORMAT INPUT DATA SIGNAL

Figure 1 is a block diagram of the system considered. Although a baseband system is shown, a system using linear modulation and demodulation can readily be fit to this model. $P(\omega)$ is the basic shaping filter and it is assumed that the receiver is matched to this shaping filter. For simplicity, $P(\omega)$ is chosen to be real. The added noise is white gaussian. $H(\omega)$ is the narrow high-pass causal filter whose effects are considered. Since $H(\omega)$ is narrow, it makes little difference whether the noise is added ahead of, behind, or somewhere in the middle of this filter.

The source generates symbols randomly from an N-ary alphabet at a rate of $1/T$ symbols per second. The transmitted signal may be represented by

$$s(t) = \sum_{k=-\infty}^{\infty} a_k p(t - kT)$$

where the a_k's are independent, zero-mean random variables which take one of N equally spaced values with equal probability, and $p(t)$ is the impulse response of $P(\omega)$.

It is assumed that there is no distortion other than $H(\omega)$ and that $Q(\omega) = P^2(\omega)$ is a Nyquist shaped filter of bandwidth less than $1/T$, so that

$$q(kT) = 0, \quad \text{all} \quad k \neq 0. \tag{14}$$

If we let $P(0) = 1$, then

$$q(0) = \frac{1}{2\pi} \int Q(\omega)\, d\omega = T/P. \tag{15}$$

The power of the transmitted signal is

$$S = \frac{\langle a_k^2 \rangle_{\text{av}}}{2\pi T} \int P^2(\omega)\, d\omega = \frac{\sigma_a^2}{T^2} \tag{16}$$

where σ_a^2 is the variance of a_k.

The signal presented to the sampler may be written in the form

$$r(t) = \sum_{k=-\infty}^{\infty} a_k[q(t - kT) + e(t - kT)] + n(t)$$

where $e(t)$ is the error signal caused by the low frequency removal, $H(\omega)$. From equations (14) and (15),

$$r(mT) = a_m\left[\frac{1}{T} + e(0)\right] + \sum_{k \neq m} a_k e[(m - k)T] + n(mT). \tag{17}$$

The effect of the low frequency removal is both the reduction of the signal amplitude [since $e(0)$ is negative] and, more important, the introduction of intersymbol interference.

The Fourier transform of the error signal is

$$E(\omega) = Q(\omega)[H(\omega) - 1] \tag{18}$$

so that

$$e(t) = \int_{-\infty}^{\infty} q(t - x)h_{-1}(x)\, dx$$

where $h_{-1}(t)$ is the inverse Fourier transform of $[H(\omega) - 1]$.

In all cases of interest, $H(\omega) - 1$ is much narrower than $Q(\omega)$. The time function $h_{-1}(t)$ therefore is virtually constant over a time interval equal to the effective duration of $q(t)$. We may therefore approximate $q(t)$ by a delta function, whose area is unity since $Q(0) = 1$.

$$e(t) = \int_{-\infty}^{\infty} \delta(t - x)h_{-1}(x)\, dx.$$

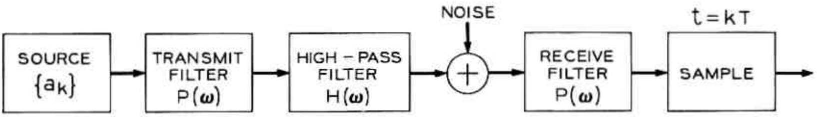


Fig. 1 — System block diagram.

If $H(\omega)$ is causal (which is the case we are interested in), then

$$e(t) = \begin{cases} 0 & t < 0 \\ \frac{1}{2}h_{-1}(0^+) & t = 0 \\ h_{-1}(t) & t > 0 \end{cases} \tag{19}$$

where $e(t)$ is the negative of the impulse response of a narrow causal low-pass filter. The generalized bound given in equation (1) can be applied to this case as:

$$P_e < \frac{2(N-1)}{N} \exp\left\{ - \frac{\left[\frac{1}{T} + e(0) - (N-1) \sum_{k \epsilon K} | e_k |\right]^2}{2\left[\sigma_n^2 + \frac{N^2-1}{3} \sum_{k \notin K} e_k^2\right]} \right\}. \tag{20}$$

The quantity σ_n^2 is the noise power at the sampler input and is also equal to the noise power at the receiver input, measured in a bandwidth equal to half the signaling rate. For the N-level system,

$$\sigma_a^2 = \frac{N^2 - 1}{3}.$$

In terms of the signal power, equation (16), equation (20) may be rewritten as

$$P_e < \frac{2(N-1)}{N} \exp\left\{ - \frac{[1 + g(0) - (N-1) \sum_{k \epsilon K} | g_k |]^2}{2\sigma_a^2\left[\frac{\sigma_n^2}{S} + \sum_{k \notin K} g_k^2\right]} \right\} \tag{21}$$

where $g(t)$ is the normalized error signal

$$\begin{aligned} g(t) &= Te(t) \\ G(\omega) &= T[H(\omega) - 1]. \end{aligned} \tag{22}$$

To apply the simplified bounds derived in equations (12) and (13), we must first specify the high pass filter, $H(\omega)$.

3.1 *Single Pole Filter*

A very common type of low frequency removal results from the use of a single capacitor or transformer. The transfer function is

$$H(s) = \frac{s\tau}{s\tau + 1}$$

where τ is the time constant of the low frequency removal circuit. Its corner frequency is then $1/(2\pi\tau)$. From equation (22), the normalized error signal is

$$G(s) = -\frac{T}{s\tau + 1} \tag{23}$$

and

$$g(t) = -\frac{T}{\tau} \exp\left(-\frac{t}{\tau}\right) u(t)$$

where $u(t)$ is the unit step function. Introducing the normalized quantity

$$a = \frac{T}{\tau}, \tag{24}$$

then

$$g(kT) = \begin{cases} 0 & k < 0 \\ -\frac{a}{2}, & k = 0. \\ -a \exp(-ka), & k > 0 \end{cases} \tag{25}$$

Letting

$$\begin{aligned} Tf_0 &= 1 + g(0) \\ Tf_k &= | g_k |, \qquad k = 1, 2, \cdots \end{aligned} \tag{26}$$

and

$$r = e^{-a},$$

the normalized eye opening becomes

$$Tf_0 - \frac{(N-1)Tf_1}{1-r} = 1 - \frac{a}{2} - \frac{(N-1)ae^{-a}}{1-e^{-a}} < 0. \tag{27}$$

Thus, $M_{\text{opt}} = 0$, and equation (13) becomes

$$P_e < \frac{2(N-1)}{N} \exp\left\{ - \frac{\left(1 - \dfrac{a}{2}\right)^2}{2\sigma_a^2\left[\dfrac{\sigma_n^2}{S} + \dfrac{a^2}{\exp(2a) - 1}\right]} \right\} \tag{28}$$

When $a \ll 1$, we may approximate equation (28) by

$$P_e < \frac{2(N-1)}{N} \exp\left[- \frac{1}{2\sigma_a^2\left(\dfrac{\sigma_n^2}{S} + \dfrac{a}{2}\right)} \right]. \tag{29}$$

The error bounds for binary, 4-level and 8-level systems are plotted in Figs. 2, 3, and 4, respectively, as a function of the signal to noise ratio, S/σ_n^2, and the normalized reciprocal time constant, a. The dashed curves are the exact values for no low-frequency removal.

$$\left(\text{that is,}\quad P_e = \frac{2(N-1)}{N} \operatorname{erfc}\left\{\frac{1}{T\sigma_n}\right\} = \frac{2(N-1)}{N} \operatorname{erfc} \frac{(S)^{\frac{1}{2}}}{\sigma_a\sigma_n}, \right.$$

$$\left. \text{where}\quad \operatorname{erfc}(x) = \frac{1}{(2\pi)^{\frac{1}{2}}} \int_x^\infty e^{-t^2/2}\, dt \right).$$

It is seen that, in the region of 10^{-5} error probability, these exact

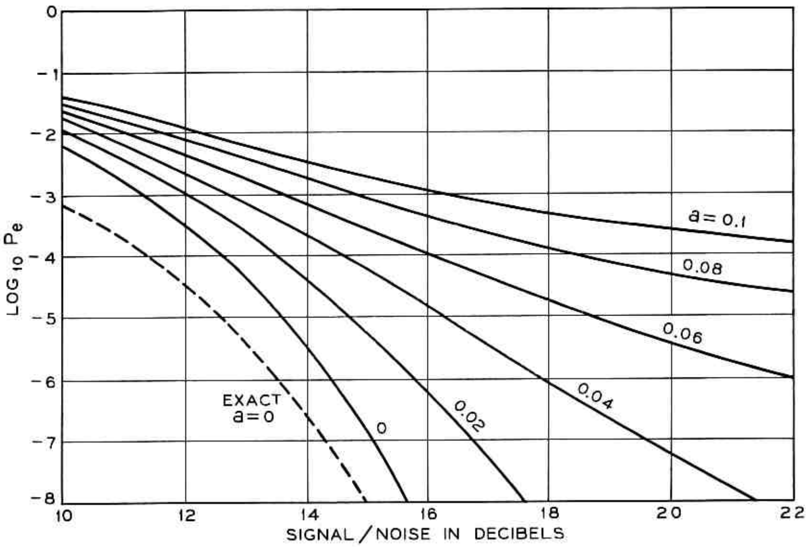


Fig. 2 — Upper bound of the error probability of a binary standard format system with a single-pole high-pass filter.

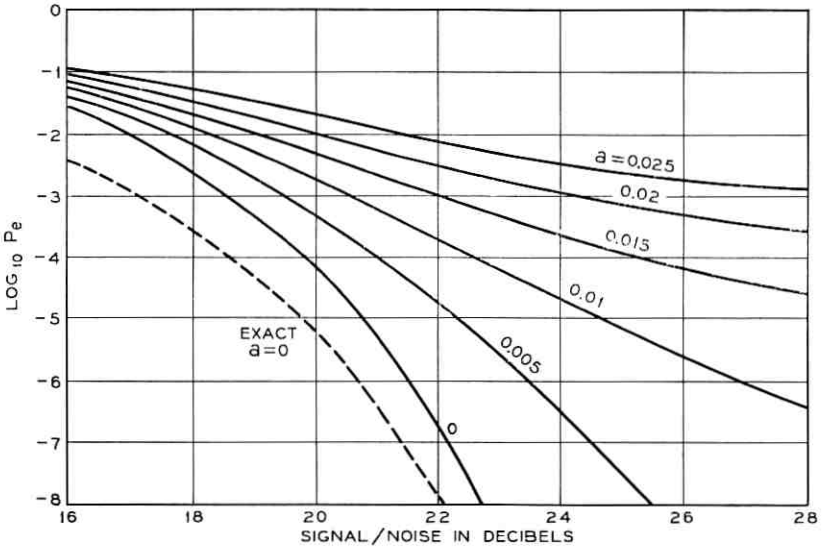


Fig. 3 — Upper bound of the error probability of a 4-level standard format system with a single-pole high-pass filter.

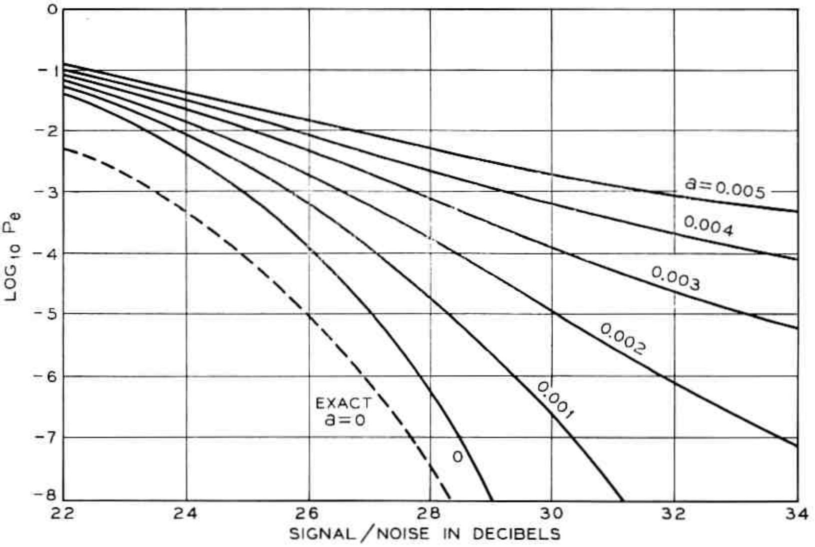


Fig. 4 — Upper bound of the error probability of an 8-level standard format system with a single-pole high-pass filter.

curves differ from the corresponding bounds by approximately a factor of 10 in error probability, or 1 decibel in signal to noise ratio.

In the binary case, it is seen that a simple high-pass filter with a time constant of 50 bit intervals introduces a degradation of only about 1 decibel in the region of 10^{-5} error probability. On the other hand, a time constant of 10 bit intervals leads to totally unacceptable performance. For the same amount of degradation and the same symbol rate, the 4- and 8-level systems must have high-pass time constants respectively 5 and 21 times that of the binary system.

3.2 *Cascaded Single Pole Filters*

In many cases, several single-pole high-pass filters are contained in the transmission path of the system. If n identical networks are used, then the overall high-pass transfer function is

$$H_n(s) = \left(\frac{s\tau}{s\tau + 1}\right)^n. \tag{30}$$

In many cases, a transfer function containing a large number of real poles of different values can be approximated by a transfer function of the form of equation (30).[6]

The Laplace transform of the error signal is

$$G_n(s) = T\left[\left(\frac{s\tau}{s\tau + 1}\right)^n - 1\right].$$

To find $g_n(t)$, we first evaluate

$$\mathcal{L}\left[\frac{1}{T}\exp\left(\frac{t}{\tau}\right)g_n(t)\right] = \frac{(s - 1/\tau)^n}{s^n} - 1 = \frac{1}{s^n}\sum_{k=1}^{n}\binom{n}{k}\left(-\frac{1}{\tau}\right)^k s^{n-k}$$

$$\frac{1}{T}\exp\left(\frac{t}{\tau}\right)g_n(t) = \sum_{k=1}^{n}\binom{n}{k}\left(-\frac{1}{\tau}\right)^k\frac{t^{k-1}}{(k-1)!}, \qquad t > 0$$

$$g_n(t) = -\frac{T}{\tau}\exp\left(-\frac{t}{\tau}\right)\sum_{k=0}^{n-1}\frac{1}{k!}\binom{n}{k+1}\left(-\frac{t}{\tau}\right)^k, \qquad t > 0.$$

At the sampling times,

$$g_n(mT) = \begin{cases} 0, & m < 0 \\ -\dfrac{na}{2}, & m = 0 \\ -a\exp(-ma)\displaystyle\sum_{k=0}^{n-1}\binom{n}{k+1}\frac{(-ma)^k}{k!}, & m > 0 \end{cases} \tag{31}$$

where again $a = T/\tau$.

This function may also be expressed in terms of the generalized Laguerre polynomial,[7]

$$L_n^{(-1)}(x) = \frac{1}{n}\sum_{k=1}^{n}\binom{n}{k}\frac{(-x)^k}{(k-1)!}$$
$$g_n(mT) = \frac{n}{m}\exp(-ma)L_n^{(-1)}(ma), \qquad m > 0. \tag{32}$$

It has been found empirically in several numerical examples that the best error probability bound was obtained when all intersymbol interference terms were added to the noise (that is, $M_{\text{opt}} = 0$). The resultant bound is therefore

$$P_e < \frac{2(N-1)}{N}\exp\left\{-\frac{\left(1-\frac{na}{2}\right)^2}{2\sigma_a^2\left[\frac{\sigma_n^2}{S}+\sum_{m=1}^{\infty}\left(\frac{n}{m}\exp(-ma)L_n^{(-1)}(ma)\right)^2\right]}\right\}. \tag{33}$$

An example of practical interest is the evaluation of the performance of a baseband binary 50,000 bits per second data set without dc restoration, operating over a transmission facility using transformer coupled repeaters. The transformers each have a corner frequency of 15 Hz, and therefore a time constant of

$$\tau = \frac{1}{2\pi \times 15} = 10.6 \text{ msec.}$$

so that

$$a = \frac{2 \times 10^{-5}}{0.0106} = 0.00188.$$

The results of Fig. 2 indicates a degradation of only about 0.1 decibel when a single transformer is introduced. However, several transformers are usually present in actual systems. The error signals, $g_n(t)$, and error probability bounds have been computed for both 14 and 28 transformers. The error signals for these two cases are shown in Fig. 5. Remember that one millisecond is equal to 50 bit intervals.

Figure 6 shows the error probability bounds for these situations; 28 transformers lead to unacceptable performance while 14 transformers introduce a degradation of 3 decibels at 10^{-5} error rate. It is significant that n transformers produce more degradation than a single transformer with a corner frequency n times greater. Also, under the assumptions of this paper, all of the above results apply independently of the roll-off characteristic of $Q(\omega)$, as long as it is a member of the Nyquist I class.

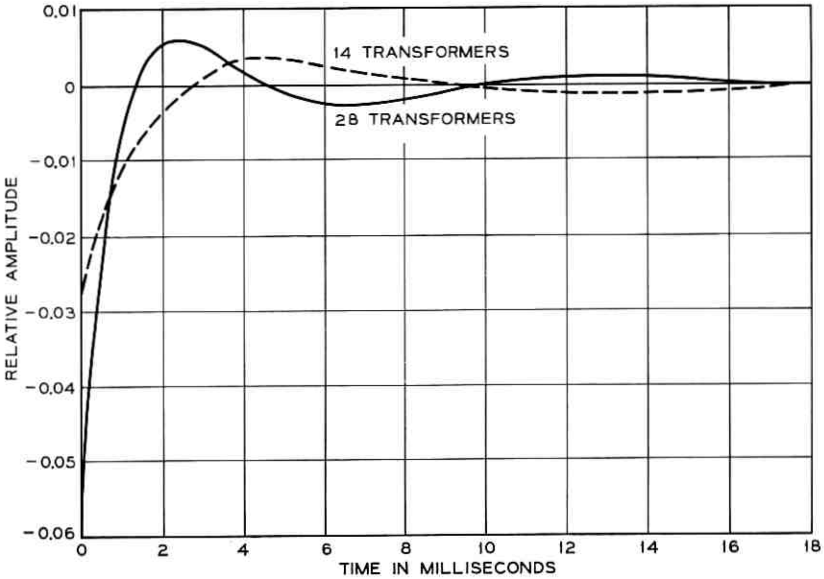


Fig. 5 — Errors signal for a cascade of transformers with 15 Hz corner frequencies.

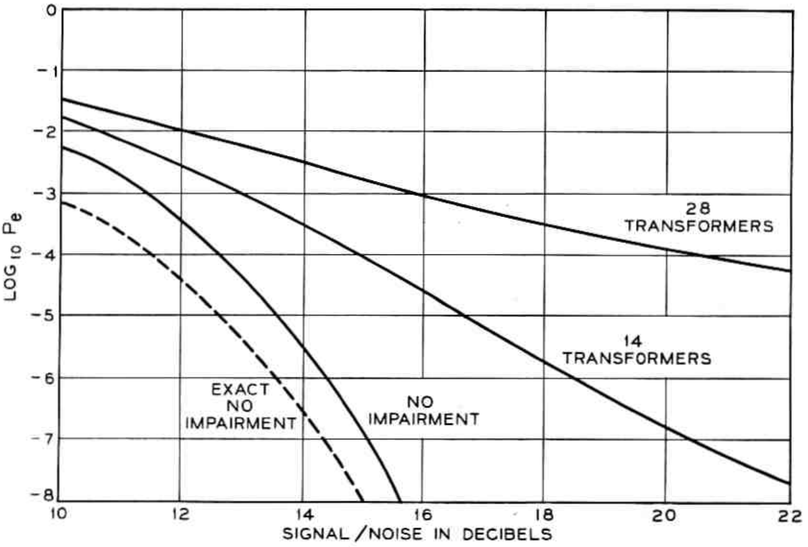


Fig. 6 — Upper bound of the error probability of a 50,000 bit/sec binary system with a cascade of transformers with 15 Hz corner frequencies.

IV. ERROR PROBABILITY PERFORMANCE WITH AN N-LEVEL EXTENSION OF A DUOBINARY INPUT DATA SIGNAL

The system model considered here is identical to Fig. 1 except that (*i*) a precoder which converts the input N-level sequence $\{a_k\}$ to another N-level sequence $\{b_k\}$ according to the relation

$$b_n = [a_{n-1} - b_{n-1}](\mathrm{mod}\ N) \tag{34}$$

is inserted between the source and the transmitting filter, $P(\omega)$, and (*ii*) a decoder follows the sampler which decodes the received levels modulo N to recover the original symbols a_n. The important point for our application is that by including precoding at the transmitter, no knowledge of any symbol or sample other than the received sample, r_k, is involved in deciding a_k.

Instead of the Nyquist shaping characteristic, the cosine filter is used for the composite signal shaping characteristic, $Q(\omega) = P^2(\omega)$, that is,

$$Q(\omega) = \cos \frac{T}{2}\, \omega, \qquad | \omega | \leqq \frac{\pi}{T}. \tag{35}$$

The system impulse response is given by

$$q(t) = \frac{2}{\pi T}\left[\frac{\cos \pi t/T}{1 - 4t^2/T^2}\right],$$

so its values at the sampling instant are

$$q[(k - \tfrac{1}{2})T] = \begin{cases} \dfrac{1}{2T} & k = 0, 1 \\ 0 & k \neq 0, 1. \end{cases} \tag{36}$$

The power of the transmitted signal is

$$S = \frac{\langle b_k^2 \rangle_{\mathrm{av}}}{2\pi T} \int_{-\pi/T}^{\pi/T} Q(\omega)\, d\omega = \frac{2\sigma_b^2}{\pi T^2} \tag{37}$$

where σ_b^2 is the variance of b_k. If the input symbols a_k are equally likely and independent, then so are the precoded symbols b_k. Thus, $\sigma_b^2 = \sigma_a^2$. The sampler input waveform, $r(t)$, may be expressed as

$$r(t) = \sum_{k=-\infty}^{\infty} b_k[q(t - kT) + e(t - kT)] + n(t) \tag{38}$$

where once again $e(t)$ is the degradation caused by the low frequency removal, $H(\omega)$. Substituting equation (36) in equation (38),

$$r[(m - \tfrac{1}{2})T] = r_m = (b_m + b_{m-1})\frac{1}{2T} + b_m e\left(-\frac{T}{2}\right) + \sum_{k \neq m} b_k e[(m - k - \tfrac{1}{2})T] + n[(m - \tfrac{1}{2})T]. \tag{39}$$

If $H(\omega)$ is causal as before, then $e(-T/2)$ will be zero. Making the same assumptions as in the standard signal format case, we arrive at an expression for error probability analogous to equation (21)

$$P_e < 2\left(\frac{N^2 - 1}{N^2}\right) \exp\left\{-\frac{[\frac{1}{2} - (N - 1)\sum_{k \epsilon K} | g_k |]^2}{2\sigma_a^2\left[\frac{2\sigma_n^2}{\pi S} + \sum_{k \epsilon K} g_k^2\right]}\right\}. \tag{40}$$

Here we consider only the single pole high-pass filter for $H(\omega)$. The result for a cascade of n identical poles follows immediately.

4.1 *Single Pole Filter*

We start by examining the normalized eye. Letting

$$\begin{aligned} Tf_0 &= 1/2 \\ Tf_k &= | g_k |; \qquad k = 1, 2, \cdots \end{aligned} \tag{41}$$

and $r = e^{-a}$

$$Tf_0 - \frac{(N - 1)f_1}{1 - r} = \frac{1}{2} - \frac{(N - 1)ae^{-a}}{1 - e^{-a}} < 0. \tag{42}$$

Thus, $M_{\text{opt}} = 0$ and equation (13) becomes for $a \ll 1$

$$P_e < 2\left(\frac{N^2 - 1}{N^2}\right) \exp\left\{-\frac{1/4}{2\sigma_a^2\left(\frac{2\sigma_n^2}{\pi S} + \frac{a}{2}\right)}\right\}. \tag{43}$$

Figures 7, 8, and 9 illustrate the behavior of the error probability bounds versus S/σ_n^2 for binary, 4-level and 8-level partial response signals with the normalized reciprocal time constant, a, as a parameter. The dotted curves give the exact values of P_e for the case $a = 0$

$$\left[\text{that is, } P_e = 2\left(\frac{N^2 - 1}{N^2}\right) \operatorname{erfc}\left(\frac{(S\pi/8)^{\frac{1}{2}}}{\sigma_a \sigma_n}\right)\right].$$

We once again observe that in the neighborhood of 10^{-5} error probability, the exact curves for $a = 0$ differ from the corresponding bounds by approximately a factor of 10 in error probability, or 1 decibel in signal to noise ratio.

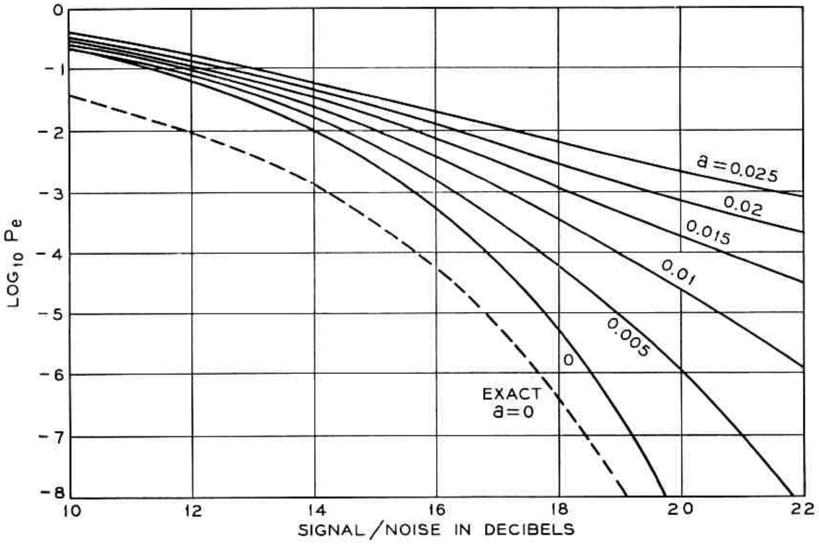


Fig. 7 — Upper bound of the error probability of a binary partial response system with a single-pole high-pass filter.

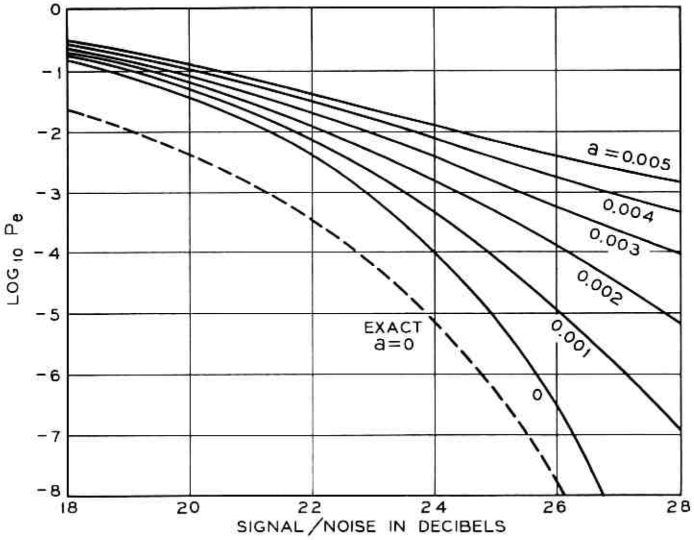


Fig. 8 — Upper bound of the error probability of a 4-level partial response system with a single-pole high-pass filter.

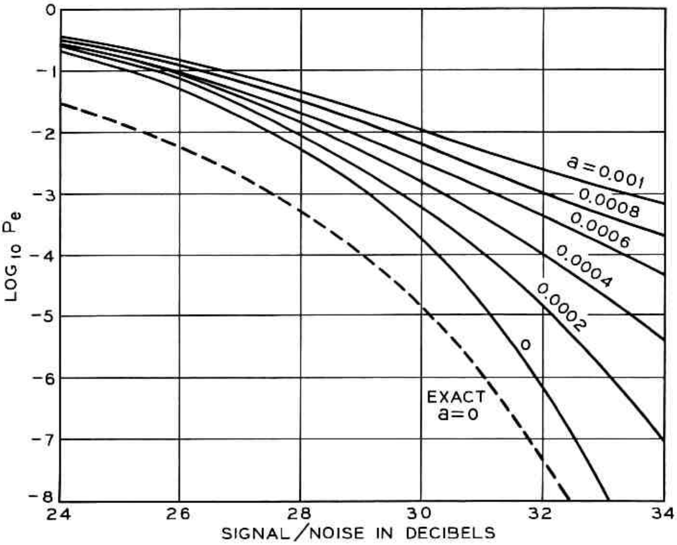


Fig. 9 — Upper bound of the error probability of an 8-level partial response system with a single-pole high-pass filter.

However, to achieve a S/N degradation of only 1 decibel in the region of 10^{-5} error probability with a simple high-pass filter, the time constant must be about four times that needed for the standard format signal. The above statement is true for the binary, 4-level, and 8-level cases. This more stringent requirement on the location of the low frequency cutoff may be viewed as a tradeoff for the saving in bandwidth associated with partial response signaling.

V. CONCLUSIONS

Although a high-pass filter will always close the eye pattern of *i* a standard format data signal (Nyquist I shaping) or *iii* a multilevel partial response signal (duobinary format), the error probability may still be quite low for random data provided that the high-pass filter is sufficiently narrow. This effect permits the use of capacitor or transformer coupling in the data terminals or transmission facilities. Multilevel systems require a longer time constant for these networks than do binary systems for the same performance.

Upper bounds of error probability have been given for binary, 4-level, and 8-level systems with gaussian noise and a single-pole high-pass filter (exponential time response). A binary system with a standard format input signal is degraded by only about 1 decibel by a simple high-pass filter whose time constant is 50 bit intervals. Four-level and 8-level systems require time constants of 250 and 1000 baud intervals, respectively, for the same performance.

A data system whose input is a binary, 4-level, or 8-level partial response signal must have a low frequency cutoff which is two octaves lower in order to achieve the same performance as a standard format system.

The error signal for a multiple-order pole is an exponential multiplied by a generalized Laguerre polynomial. The performance of a system with an nth order pole high-pass filter is worse than one with a single pole n times as large.

REFERENCES

1. Becker, F. K., Davey, J. R., and Saltzberg, B. R., "An AM Vestigial Sideband Data Set Using Synchronous Detection," AIEE Trans. Commun. and Elec., *81,* No. 60 (May 1962), pp. 97–101.
2. Aaron, M. R., "PCM Transmission in the Exchange Plant," B.S.T.J., *41,* No. 1 (January 1962), pp. 99–141.
3. Kretzmer, E. R., "Generalization of a Technique for Binary Data Communication," IEEE Trans. Commun. Tech., *COM-14,* No. 1 (February 1966), pp. 67–68.
4. Gerrish, A. M. and Howson, R. D., "Multilevel Partial-Response Signaling," IEEE Int. Conf. Commun. Digest, 1967, p. 186.
5. Saltzberg, B. R., "Intersymbol Interference Error Bounds with Application to Ideal Bandlimited Signaling," IEEE Trans. Inform. Theor., *IT-14,* No. 4 (July 1968), pp. 563–568.
6. Papoulis, A., *The Fourier Integral and Its Applications,* New York: McGraw-Hill, 1962, pp. 234–237.
7. Sansone, G., *Orthogonal Functions,* New York: Interscience Publishers, 1959, Chap. 4.

Computer-Aided Circuit Design by Singular Imbedding

By E. B. KOZEMCHAK and M. A. MURRAY-LASSO

(Manuscript received August 5, 1968)

We give a new and powerful method for the direct solution of circuit design problems. The method begins with a prespecified topology and some or all elements undetermined in value. The designer imposes on the circuit any desired set of node-pair voltages, branch currents, or driving point and transfer immittances. Values of circuit elements that satisfy the constraints are directly calculated. This direct method of solution avoids the usual iterative analysis-optimization schemes, reducing computer times by up to three orders of magnitude.

A linear set of design equations is formulated by choosing undetermined element currents and node voltages as the variables. Singular elements are introduced to impose the desired constraints. Inequality as well as equality constraints are permitted. Element values are determined from the solution of these equations. In this paper we emphasize our method of solution in relation to dc networks.

I. INTRODUCTION

The most significant advances made in computer-aided circuit design have been in analysis programs. The designer can now choose from among several general purpose programs that program which most nearly suits his particular needs. In designing a circuit to meet a given set of requirements, the usual approach has been to use analysis programs in some optimization scheme. Through an iterative process, carried out by the machine, the man, or a man-machine interaction, a final design is reached. The approach presented here provides a direct solution, and does not rely on such iterative schemes.

The method is most fertile in the area of active network design, where one often wishes to choose element values in a specified topology in order to meet some set of requirements. The method has been applied to a number of design problems of current interest including biasing direct coupled transistor circuits; designing transistor amplifiers for specified midband gain, input, and output impedances; and simultaneously realizing several specified impedance or admittance parameters of a network.

In the design of electronic circuitry, one usually wishes to imbed passive elements into a network containing active devices, and to determine the required passive element values. Therefore, this paper deals with the determination of element values in a prespecified topology for which a given performance is required. Two new elements, a voltage forcing element (VFE) and current forcing element (CFE), are introduced in order to constrain network voltages and currents. These elements may be realized with independent voltage and current sources, and the nullator, a somewhat "pathological" element used in theoretical network studies.

The method of singular imbedding places the VFE's and CFE's in a network to constrain the desired variables. The terminal voltage-current behavior of the variable elements is not specified. Instead, the constraints imposed upon the network by the VFE's and CFE's are used to determine allowed voltage-current relations for the variable elements. The formulation remains linear in these variables. The last step involves determining the element values through Ohm's law once the allowed voltage-current relations are known.

By appending the original set of equations with a set of inequality constraints, it is possible to restrict the range of element values in the solution. For example, realizations employing only element values between specified lower and upper bounds are possible. For simplicity, only the case of linear dc networks are illustrated. Extensions of the method to ac and nonlinear design are considered elsewhere.

II. A NEW APPROACH

To understand the philosophy of this new approach to design, consider the train of events in realizing a set of requirements with electronic circuitry. Since the choice of topology is better handled by the man than the computer, we will assume some specified topology in which some or all of the element values are to be chosen to meet the given criteria. For example, in designing transistor circuitry it is necessary to choose some resistance values to properly bias the transistors. Similarly, one must often choose element values to give a desired voltage gain, driving point impedance, transfer impedance, or similar network function.

The invariant feature in all of these problems is that a set of network currents and voltages, or their ratios, has been constrained. The design problem is to find any set of element values consistent with these constraints. If the problem is posed with sufficient freedom, many sets of element values may exist consistent with the imposed constraints. Conversely, if the problem is posed with insufficient freedom, inconsistent equations arise and there is no solution.

If one can find a general method of imposing these network constraints, and can simultaneously monitor the voltage-current relations these constraints force at the terminals of the variable elements, then indeed a direct solution to many computer-aided design problems will have been found.

Before proceeding, however, consider a very simple example of how one might presently handle the design problem and the difficulties that would ensue. Suppose in the network of Fig. 1, one wishes to choose G_1 and G_2 such that V' is constrained to be 0.1 volt. A set of nodal equations may be written:

$$\begin{bmatrix} 1 + G_1 & -G_1 \\ -G_1 & G_1 + G_2 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{1}$$

The first step involves a transformation of coordinates so that the desired quantities appear explicitly in the equations. In general, this will necessitate using hybrid parameters. For this case, the following transformation might be used:

$$\begin{bmatrix} V_1^{\dagger} \\ V_2^{\dagger} \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}. \tag{2}$$

Inverting the relation, we have

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} V_1^{\dagger} \\ V_2^{\dagger} \end{bmatrix} \tag{3}$$

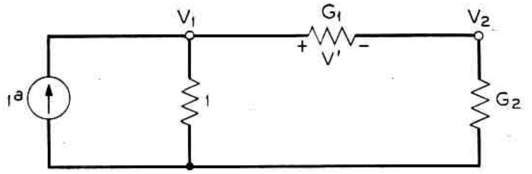


Fig. 1 — Simple design problem.

and the current-law equations become

$$\begin{bmatrix} 1 + G_1 & -G_1 \\ -G_1 & G_1 + G_2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} V_1^{\dagger} \\ V_2^{\dagger} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{4}$$

$$\begin{bmatrix} 1 + G_1 & 1 \\ -G_1 & G_2 \end{bmatrix} \begin{bmatrix} V_1^{\dagger} \\ V_2^{\dagger} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{5}$$

Substituting the constant $V' = V^{\dagger}_{1} = 0.1$, the set of equations becomes

$$\begin{aligned} 0.1(1 + G_1) + V_2^{\dagger} &= 1 \\ -0.1G_1 + G_2V_2^{\dagger} &= 0. \end{aligned} \tag{6}$$

Thus, even if one is successful in finding a transformation to a basis that includes the variables that are constrained, the result is usually a set of nonlinear equations in the network elements and voltage variables. Solving this set of nonlinear equations for the unknown voltages and element values is extremely difficult. A method of handling this difficulty has been suggested, involving the use of optimizing techniques to vary element values until the network variables take on their desired values—in this case $V' = 0.1$ volt.[1] While this is a useful approach, it has several disadvantages. First, it is time consuming since many iterations are required for convergence. Second, local minima, or lack of sufficient numerical accuracy, may prevent convergence to a correct solution. Finally, although an infinity of sets (G_1, G_2) exist to satisfy the given constraints, the optimization yields only one of these sets.

With these difficulties in mind, let us repeat the philosophy of design presented here. We first determine how the requirements constrain network currents and voltages. We then force these currents and voltages to take on the desired values. Finally, we determine the effect of such constraints upon the voltage-current relations at the terminals of variable elements. These $v - i$ relations then determine the values of the variable elements.

III. NETWORK CONSTRAINTS

The common feature of all network synthesis problems is that they require some specified relation between some voltages and currents in the network. For example, synthesis of a given driving point impedance constrains the ratio of a port voltage to the current at that port. Synthesis of a transfer impedance constrains the ratio of a port voltage to the current at a different port. A specified voltage or current gain constrains the ratio of two port voltages or port currents, respectively. Indeed the synthesis of entire network matrices is a combination of such constraints. Similarly, the static design problem in electronic circuits involves fixing certain branch currents and branch voltages. For example, one usually wishes to bias a transistor for a given collector current and collector-emitter voltage. Resistance values are chosen consistent with these constraints.

It is essential to demonstrate a method for constraining voltages and currents in a network. The required constraints are shown in Fig. 2. We introduce two new elements, a current forcing element, CFE(I_o), and a voltage forcing element, VFE(V_o), which will be realized with more conventional elements shortly. We want the CFE(I_o) to be such that it constrains the current through branch j to be I_o, without otherwise affecting the behavior of the network. We want the VFE(V_o) to be such that it constrains the voltage across branch j to be V_o without otherwise affecting the behavior of the network.

In discussing the properties of the CFE and VFE, we use the concept of admissible or allowed pairs of voltage and current variables (v, i).[2] The set of voltage-current pairs that a system N allows can be used to completely describe that system.[3] For example, let the system under consideration, N_R, consist of a single resistor of value R. Then the system is completely described by its allowed terminal voltage and current pairs; namely, $(Ri, i) \varepsilon N_R$. Similarly, a capacitance of value C, denoted N_C, is completely described by its allowed pairs $(v, d(Cv)/dt) \varepsilon N_C$.

We now define the CFE(I_o) and VFE(V_o) in terms of their allowed pairs.

Current forcing element (I_o):

$$(0, I_o) \varepsilon N_{\mathrm{CFE}(I_o)} . \tag{7}$$

Here we postulate an element which allows no voltage drop across its terminals, and passes only a specified current I_o.

Next, we postulate an element which allows only a fixed voltage V_o to exist at its terminals, and passes no current.

Voltage forcing element (V_o):

$$(V_o , 0) \varepsilon N_{\mathrm{VFE}(V_o)} . \tag{8}$$

Figure 2 makes clear the use of these elements in constraining network variables. In Fig. 2a, the current in branch j is forced to be

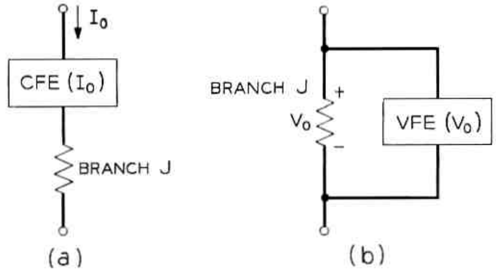


Fig. 2 — Network constraints. (a) Branch J current constrained by current forcing element (CFE); (b) Branch J voltage constrained by voltage forcing element (VFE).

I_o by inserting a CFE in series. Since the CFE(I_o) allows no voltage to exist across its terminals, its presence affects Kirchhoff's current and voltage laws only to the extent that branch j current is constrained to be I_o. Notice that this would not be the case had we inserted a current source in series with branch j. The current source would allow some voltage to exist between its terminals which would have been included in Kirchhoff's voltage law equations. Thus, a current source of value I_o would not only constrain branch j current to be I_o, but would also introduce a new degree of freedom, namely, the voltage across the current source.

Similar reasoning can be applied to Fig. 2b. Here a VFE(V_o) is applied across branch j to constrain that voltage to be V_o. Since the VFE(V_o) passes no current, Kirchhoff's laws are affected only to the extent that branch j voltage is now constrained to be V_o. The network cannot respond with a new degree of freedom, as it could if a voltage source were placed across branch j and thus allowed to introduce a new current variable in Kirchhoff's current law equations. It should be noted that the VFE(V_o), can be placed between any two nodes to constrain the voltage between those nodes; it need not be placed across a branch.

By using current sources and voltage sources in conjunction with VFE's and CFE's, current-voltage ratios may be constrained. For example, in Fig. 3a,

$$\frac{V_1}{I_1} = \frac{V_o}{I} = \frac{ZI}{I} = Z. \tag{9}$$

In Fig. 3b,

$$\frac{I_1}{V_1} = \frac{I_o}{V} = \frac{YV}{V} = Y. \tag{10}$$

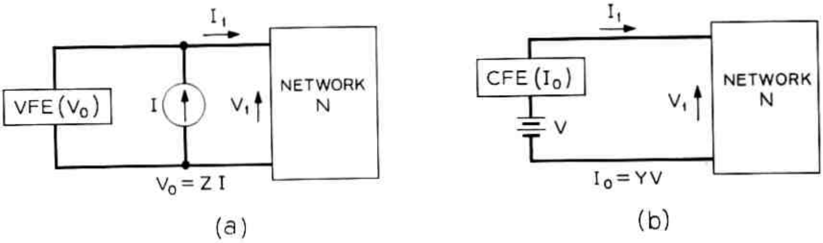


Fig. 3 — Methods of constraining current-voltage ratios. (a) Impedance forcing element [IFE(Z)]; (b) admittance forcing element [AFE(Y)].

Thus we are constraining the network N to have, in the first case, a driving point impedance Z, and in the second case, a driving point admittance Y. The configurations used to constrain impedances or admittances will be denoted impedance forcing elements, IFE(Z), and admittance forcing elements, AFE(Y). Notice that IFE's and AFE's are composed of CFE's, VFE's, and independent sources. They are useful in constraining a network to have a desired driving point impedance or admittance.

We already mentioned that VFE's and CFE's could be realized in terms of existing elements. The necessary elements are the ideal current source, the ideal voltage source, and the nullator, a somewhat "pathological" network element introduced by Tellegen.[4] Returning to the allowed pair concept, the nullator is defined to be a two-terminal element for which the only allowed voltage-current pair is (0, 0). It can be looked upon as a simultaneous open and short circuit, since it allows only zero voltage at its terminals and passes no current.

From its definition, one could not hope to physically realize and isolate such a device. However it's characteristics may be observed at the input to an operational amplifier imbedded in a feedback network, where the input is at a virtual ground (short circuit) and yet passes no current (open circuit). The nullator is represented schematically in Fig. 4.

By appropriate connections of voltage sources, current sources, and nullators, the VFE's and CFE's may be realized as in Fig. 5. Remembering that the nullator passes zero current and has zero voltage across its terminals, the equivalents of Fig. 5 becomes clear. In

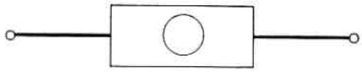


Fig. 4 — Schematic representation of nullator.

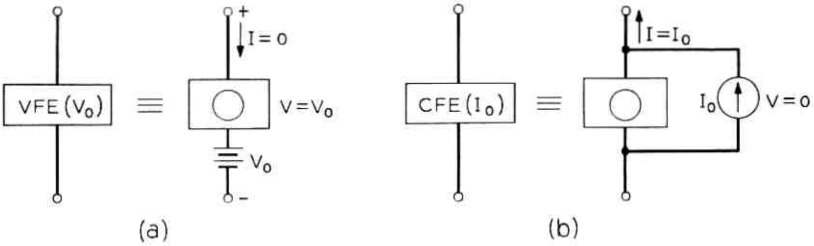


Fig. 5 — Equivalent circuits for VFE and CFE using nullators.

Fig. 5a, the terminal voltage must be V_o, and since no current exists in the element the combination voltage source and nullator is by definition a VFE(V_o). In Fig. 5b, a current I_o exists at the terminals but no voltage drop exists across the terminals. Thus by definition, the combination current source and nullator is a CFE(I_o).

IV. ADDING FREEDOM TO THE NETWORK

In the previous section, we placed constraints on the network that would generally lead to a set of inconsistent equations if all the elements were also specified. However, if some network elements are variable, we can determine how the constraints affect the voltage-current relations at the variable element terminals, and then choose variable elements in such a way as to be consistent with these $v - i$ relations.

We propose two methods of characterizing the variable elements. First, since the element is variable, we can ascribe no functional relation between the voltage and current of that branch. This is handled in writing the nodal equations for the network by explicitly adding the currents through variable elements into the equations, rather than first transforming them into voltage variables through a functional relation of the form

$$i_b = Y_b v_b \tag{11}$$

where the b implies the variable refers to some branch. The nodal equations are of the form

$$[\mathbf{Y}_f]\mathbf{V}] = \mathbf{I}_S] + [\mathbf{C}]\mathbf{I}], \tag{12}$$

where

$\mathbf{V}]$ is an n-vector of node voltages.

$\mathbf{I}_S]$ is an n-vector of forcing currents at each node.

$[\mathbf{Y}_f]$ is the $n \times n$ nodal admittance matrix of the fixed portion of the network.

$\mathbf{I}]$ is an r-vector of unknown currents through variable elements (r is the number of variable elements).

$[\mathbf{C}]$ is the $n \times r$ node cutset matrix for the graph of variable elements.

$\mathbf{I}]$ and $\mathbf{V}]$ are both vectors of network variables, and may be combined by matrix partitioning as

$$[-\mathbf{C} \mid \mathbf{Y}_f]\left[\frac{I}{V}\right] = I_s]. \tag{13}$$

Equation (13) describes a network in which some element values can be chosen to meet the given constraints. In the remainder of this paper, we combine the added degrees of freedom given by the variable elements in equation (13) with the constraints imposed by the CFE's and VFE's. All networks, satisfying the VFE and CFE constraints and the specified topology, with be generated.

A simple example will help clarify these concepts. Figure 6 is the network of Fig. 1, with the 1-ohm resistor replaced by a known resistance of R ohms. Currents I_1 and I_2 are those carried by the variable conductances G_1 and G_2, respectively. The set of nodal equations is

$$\begin{bmatrix} 1/R & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} I_1 \\ I_2 \end{bmatrix}. \tag{14}$$

Rearranging into the form of equation (13),

$$\left[\begin{array}{rr:rr} 1 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 \end{array}\right]\begin{bmatrix} I_1 \\ I_2 \\ V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{15}$$

From this example, the method of generating equation (13) should become clear.

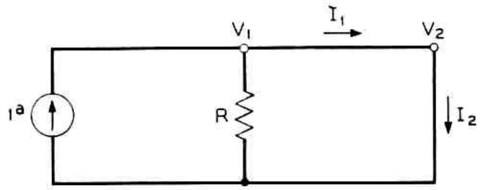


Fig. 6 — Simple design problem.

The second approach useful in dealing with variable elements in a network is the introduction of another pathological element, the norator, shown in Fig. 7, also introduced by Tellegen.[4] The norator is a two-terminal element with allowed pairs (v, i), with v and i independent and arbitrary. Thus, any voltage and current may appear across its terminals simultaneously, which is the property that we desire of variable elements. We do not wish to force any functional relation between the voltage across and the current through variable elements. We wish only to observe constraints that may be imposed on the $v - i$ relations by the VFE's and CFE's. The norator allows the network the extra degree of freedom taken away by the introduction of nullators.

V. FORMULATION OF NETWORK EQUATIONS

Since the introduction of nullators and norators into a network will generally introduce singularities into the corresponding equations, we call the approach we are considering the method of singular imbedding. It has been demonstrated that the design problem can be reduced to the appropriate imbedding of nullators, norators, and independent voltage and current sources. Let us now examine the effect of such imbedding on the network equilibrium equations. Since a nodal admittance formulation is used, it is important to determine the effect of nullators and norators on the admittance matrix.

Independent voltage sources may be conveniently incorporated into an admittance formulation. If a series impedance exists with the voltage source, application of Norton's Theorem is sufficient. If no series impedance exists, the introduction of positive and negative impedances is necessary in transforming the voltage source to an independent current source (see Fig. 8).

The effect of nullators and norators upon the admittance matrix of a network has been considered by A. C. Davies.[5] Let us write the nodal equations for the network with all nullators and norators removed. The equations are of the form

$$[\mathbf{Y}_o]\mathbf{V}] = \mathbf{I}_S] \tag{16}$$

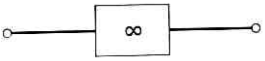


Fig. 7 — Schematic representation of norator.

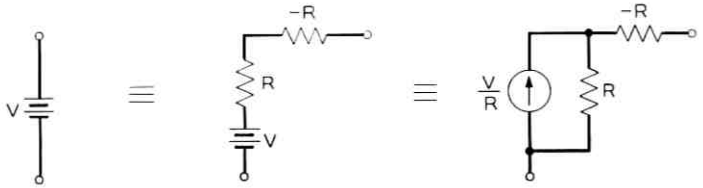


Fig. 8 — Equivalent circuit for ideal voltage source.

where

- $[\mathbf{Y}_o]$ is the admittance matrix of the network with nullators and norators removed
- $\mathbf{V}]$ is the vector of node voltages with respect to ground
- $\mathbf{I}_S]$ is the vector of currents injected into each node.

Suppose now that a nullator is connected between nodes i and j. Since the nullator passes only zero current, the current law equations at those nodes are not affected. However, since there is zero voltage across the nullator, V_i and V_j are now constrained to be equal. Call this new value V_{ij}. Clearly, one degree of freedom has been removed from the network response. In addition to the matrix equation (16), one equation of the form

$$V_i = V_j \tag{17}$$

is added for each nullator imbedded in the network. Thus, if k nullators are imbedded, k additional constraint equations are added.

Two viewpoints can be taken here. First, the original set of equations, equation (16), has been appended by a set of the form

$$[\mathbf{B}]\mathbf{V}] = \mathbf{0}] \tag{18}$$

where

- $\mathbf{V}]$ is the n-vector of node voltages.
- $[\mathbf{B}]$ is a $k \times n$ matrix of -1, 0, 1 entries expressing the set of constraints of equation (17) for the k nullators.

The final set of equations becomes

$$\left[\frac{\mathbf{Y}_o}{\mathbf{B}}\right] \mathbf{V}] = \begin{matrix} \mathbf{I}_S \\ \\ \mathbf{0} \end{matrix}\Bigg]. \tag{19}$$

A second approach to the problem was suggested by Davies.[5] In the nodal equations below

$$\begin{bmatrix} y_{11} & \cdots & y_{1i} & \cdots & y_{ij} & \cdots & y_{1n} \\ y_{21} & & y_{2i} & & y_{2j} & & \\ \vdots & & & & & & \\ y_{n1} & \cdots & y_{ni} & \cdots & y_{nj} & \cdots & y_{nn} \end{bmatrix} \left. \begin{matrix} V_1 \\ \vdots \\ V_i \\ \vdots \\ V_j \\ \vdots \\ V_n \end{matrix} \right] = I_S] \tag{20}$$

the addition of a nullator between nodes i and j makes $V_i = V_j = V_{ij}$. The ith and jth column of the Y matrix are both multiplied by V_{ij}, thus they may be added and the equations written as

$$\begin{bmatrix} y_{11} & \cdots & (y_{1i} + y_{1j}) & \cdots & y_{1n} \\ y_{21} & & (y_{2i} + y_{2j}) & & y_{2n} \\ \vdots & & & & \\ y_{n1} & & (y_{ni} + y_{nj}) & \cdots & y_{nn} \end{bmatrix} \left. \begin{matrix} V_1 \\ \vdots \\ V_{ij} \\ \vdots \\ V_n \end{matrix} \right] = I_S]. \tag{21}$$

The addition of k independent nullators (no nullator loops) causes k additions of columns of Y and reduces the dimension of $V]$ by k. We denote the reduced set of equations by

$$[\mathbf{Y}_o']_{n\times(n-k)}\ V']_{(n-k)\times 1} = I_S']_{n\times 1}. \tag{22}$$

In either interpretation, we observe that the resulting set of equations is no longer square. In the first interpretation, we are increasing the dimensionality of the vector space that the column vectors must span, without adding new basis vectors to span that space. In general, the equations will be inconsistent. In the second interpretation, we are keeping the dimension of the space fixed, but reducing the number of vectors available to form a basis and the space may no longer be spanned. Again inconsistencies will generally arise. In either interpretation, the inconsistencies are to be expected since nullators (VFE's or CFE's) have been introduced to constrain network variables.

Let us now examine the way in which variable elements (additional degrees of freedom) remove these inconsistencies. Again two points of view may be taken. One provides us with new basis vectors to span the space of possible injected current vectors $\mathbf{I}_S]$, the second reduces the dimensionality of the space of $\mathbf{I}_S]$ in order that the existing number of basis vectors might again span the space.

Section III gives the essence of the first interpretation with the important result, equation (13). Observe that imbedding variable elements in a network provides an additional set of column vectors, namely, those of $[-\mathbf{C}]$, that may be used as basis vectors in spanning the space of possible $\mathbf{I}_S]$. Thus, if one has complete freedom in selecting variable elements, a set of column vectors, the columns of $[-\mathbf{C}]$ can always be found to assure that the space of all possible $\mathbf{I}_S]$ will be spanned, regardless of how the nullators reduce the space of the column vector of the $\mathbf{Y}$ matrix. This concept, which involves growing new elements to satisfy imposed constraints, will be the subject of future study.

A second approach in handling the freedom introduced by variable elements is to replace each variable element by a norator, as suggested in Section III. The method of Davies may then be employed to analyze the network containing norators.[5] Again assume that the admittance matrix $\mathbf{Y}_o$ of the network without nullators and norators is available. Thus,

$$[\mathbf{Y}_o]\mathbf{V}] = \mathbf{I}_S]. \tag{23}$$

Now suppose that a norator is connected between nodes h and k, and that the reference direction for the arbitrary norator current I_o is from h to k. The current-law equations for nodes h and k will be of the form

$$I_{Sh} - I_o = \sum_i Y_{hi}v_i \tag{24}$$

$$I_{Sk} + I_o = \sum_i Y_{ki}v_i. \tag{25}$$

Since I_o is arbitrary, and is not needed to solve for the node voltages, adding the two equations gives

$$I_{Sh} + I_{Sk} = \sum_i (Y_{hi} + Y_{ki})v_i. \tag{26}$$

This corresponds to the addition of rows h and k of the nodal equations of the network without norators. Thus for a network containing n nodes and r norators, only $n - 1 - r$ independent equations can be written.

Observe in Fig. 9 that the effect of connecting the norator between nodes h and k is to replace the nodal equations for nodes h and k

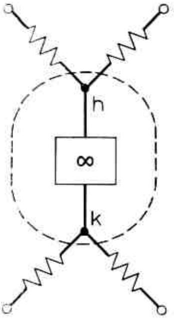


Fig. 9 — Effect of connecting norator between two nodes.

by a single current law equation for the ambit (broken line) surrounding both nodes h and k. Thus any functional relation between the current and voltage of branch j is removed, as is desired for a variable element.

To summarize thus far, the following manipulations may be performed on the network current law equations to deal with VFE's, CFE's and variable elements. To include network constraints, first imbed the CFE's and VFE's. Write the Y matrix with nullators removed. Then reduce the matrix by adding appropriate columns. This may be stated compactly by a matrix transformation as[5]

$$\mathbf{I}_S] = [\mathbf{Y}_0][\mathbf{U}_c]\mathbf{V}] \tag{27}$$

where $[\mathbf{U}_c]$ is a matrix obtained from the unit matrix by adding columns corresponding to nodes between which nullators are connected. Since the transformation $[\mathbf{U}_c]$ is singular, not all components of $\mathbf{V}]$ are determined. The undetermined ones are found from the relation

$$[\mathbf{B}]\mathbf{V}] = \mathbf{0}]. \tag{28}$$

To include variable elements, either

(*i*) Augment the $\mathbf{Y}$ matrix of the fixed portion of the network with the node cutset matrix of the graph of the variable elements to get

$$[-\mathbf{C} \mid \mathbf{Y}_o] \left.\frac{\mathbf{I}}{\mathbf{V}}\right] = \mathbf{I}_S] \tag{29}$$

or

(*ii*) Add the current law equation corresponding to nodes to which a nullator is connected. This is compactly stated by a matrix transformation as[5]

$$[\mathbf{U}_r]\mathbf{I}_S] = [\mathbf{U}_r][\mathbf{Y}_o]\mathbf{V}] \tag{30}$$

where $[\mathbf{U}_r]$ is a matrix obtained from the unit matrix by adding rows corresponding to nodes between which norators are connected. The vector of currents through variable resistors is then formed by equation (29).

VI. SOLVING THE NETWORK EQUATIONS

We now wish to solve the set of equations after imbedding CFE's, VFE's, and variable elements. We assume equation (29) to be our starting point. A similar formulation may be made using equation (30) as the starting point. CFE's and VFE's are imbedded, variable elements are specified, and nullators are removed to generate the set of equations

$$[-\mathbf{C} \;\vdots\; \mathbf{Y}_o]\left.\frac{\mathbf{I}}{\mathbf{V}}\right] = \mathbf{I}_S]. \tag{31}$$

Addition of nullators to the network adds the set of equations

$$\mathbf{B}\mathbf{V}] = \mathbf{0}] \tag{32}$$

and, from equation (27), the corresponding transformation $[\mathbf{U}_c]$ on the admittance matrix. Thus the final set of equations becomes

$$\left[\begin{array}{c|c} -\mathbf{C} & \mathbf{Y}_f\mathbf{U}_c \\ \hline \mathbf{0} & \mathbf{B} \end{array}\right]\left.\frac{\mathbf{I}}{\mathbf{V}}\right] = \left.\frac{\mathbf{I}_S}{\mathbf{0}}\right]. \tag{33}$$

As seen in the previous section, the transformation $[\mathbf{U}_c]$ (which adds columns of $\mathbf{Y}_f$) is consistent with the set of equations $[\mathbf{B}]\mathbf{V}] = 0$. Thus the second matrix equation in equation (33) will always have a solution, provided the first one does. It remains only to solve

$$[-\mathbf{C} \;\vdots\; \mathbf{Y}_f\mathbf{U}_c]\left.\frac{\mathbf{I}}{\mathbf{V}}\right] = \mathbf{I}_S]. \tag{34}$$

in order to determine the proper element values. Let

$$\left.\frac{\mathbf{I}}{\mathbf{V}}\right]_{(r+n-k)\times 1} = \mathbf{x}].$$

By using the Gauss–Jordan method one can bring these equations into the form

$$\left[\begin{array}{c|c} \mathbf{U} & \mathbf{Q} \\ \hline 0 & 0 \end{array}\right]\left.\frac{\mathbf{X}_1}{\mathbf{X}_2}\right] = \left.\frac{\mathbf{I}_{S1}}{\mathbf{I}_{S2}}\right] \tag{35}$$

where

$\mathbf{X}_1|\mathbf{X}_2]$ is a vector of node voltages and currents through variable resistors,

$[\mathbf{U}]$ is the unit matrix,

$[\mathbf{Q}]$, $\mathbf{I}_{S1}]$, $\mathbf{I}_{S2}]$ are the resulting submatrices after transformation.

If $\mathbf{I}_{S2}] = 0$ (the equations are consistent), the first equation can be solved for $\mathbf{X}_1]$ in terms of $\mathbf{X}_2]$.

$$\mathbf{X}_1 = \mathbf{I}_{S1}] - [\mathbf{Q}]\mathbf{X}_2]. \tag{36}$$

The case $\mathbf{I}_{S2}] \neq 0$ implies that there are no values of variable elements consistent with the imposed constraints. For $\mathbf{I}_{S2}] = \mathbf{0}]$, equation (36) generates all solutions to the problem. Some network variables $\mathbf{X}_2]$ can be chosen arbitrarily and the remaining variables $\mathbf{X}_1]$ determined. At each setting of $\mathbf{X}_2]$ the variable elements can be determined since all node voltages and currents through variable elements are known.

Thus

$$Z_i = \frac{V_{i1} - V_{i2}}{I_i} \qquad \text{for} \qquad i = 1, r \tag{37}$$

where $i1$ and $i2$ are connection nodes of the ith variable element. By allowing the free variables $\mathbf{X}_2]$ to take on a continuum of values, all solutions to the problem are determined directly.

Returning to the example already discussed (Fig. b), let us apply the method of singular imbedding. The circuit is redrawn in Fig. 10 with the introduction of a VFE to constrain the voltage between nodes 1 and 2 to be 0.1* volt. With the nullator removed, a set of nodal equations is written in the form of equation (13)

$$\begin{bmatrix} 1 & 0 & 1/R & 0 & 0 \\ -1 & 1 & 0 & 1 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ \hline V_1 \\ V_2 \\ V_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -0.1 \\ 0.1 \end{bmatrix}. \tag{38a}$$

The introduction of a nullator between nodes 1 and 3 results in the addition of the corresponding columns and the equality $V_1 = V_3 =$

* Since the nullator passes zero current, the series battery in the VFE model may have a nonzero resistance and still maintain the proper terminal voltage. Thus the introduction of positive and negative resistances are unnecessary here.

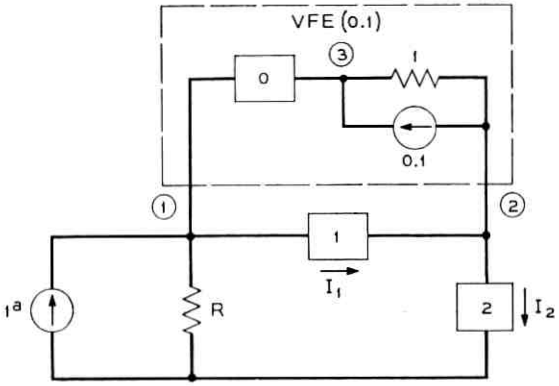


Fig. 10 — Network after singular imbedding.

V_{13}. Thus,

$$\begin{bmatrix} 1 & 0 & 1/R & 0 \\ -1 & 1 & -1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ V_{13} \\ V_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -0.1 \\ 0.1 \end{bmatrix}. \tag{38b}$$

With $R = 1$ ohm for ease of visualization, elementary row operations yield

$$\left[\begin{array}{ccc:c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{array}\right] \left[\begin{array}{c} I_1 \\ I_2 \\ V_{13} \\ \hdashline V_2 \end{array}\right] = \begin{bmatrix} 0.9 \\ 0.9 \\ 0.1 \end{bmatrix}. \tag{39}$$

Thus,

$$\begin{bmatrix} I_1 \\ I_2 \\ V_{13} \end{bmatrix} = \begin{bmatrix} 0.9 \\ 0.9 \\ 0.1 \end{bmatrix} - V_2 \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}. \tag{40}$$

It is clear that V_2 can take on arbitrary values while maintaining the constraints. We will demonstrate this for two particular values of V_2. For $V_2 = 0$

$$\begin{bmatrix} I_1 \\ I_2 \\ V_{13} \\ V_2 \end{bmatrix} = \begin{bmatrix} 0.9 \\ 0.9 \\ 0.1 \\ 0 \end{bmatrix},$$

$$R_1 = \frac{V_1 - V_2}{I_1} = \frac{1}{9},$$

$$R_2 = \frac{V_2}{I_1} = 0.$$

It is easily verified that these values, when substituting into the circuit of Fig. 1, result in $V' = V_1 - V_2 = 0.1$ volt.

Similarly for $V_2 = 0.6$ volt

$$\begin{bmatrix} I_1 \\ I_2 \\ V_{13} \\ V_2 \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.7 \\ 0.6 \end{bmatrix}$$

and $R_1 = 1/3$, $R_2 = 2$.

Again it is easily verified that $V' = V_1 - V_2 = 0.1$ volt. With this simple example in mind, let us consider the solution of more complicated networks by computer.

VII. COMPUTER SOLUTION

A program has been written to solve the design problem for resistive networks. The program performs the following operations

(*i*) Accepts input of circuit description in conversational mode. The circuit may contain resistors (both fixed and variable), VFE's CFE's batteries, independent current sources, and current controlled current sources.

(*ii*) Generates $\mathbf{C}$, $\mathbf{Y}_t$, and $\mathbf{I}_S$ matrices for the network.

(*iii*) Reduces equations to triangular form by a Gaussian reduction which pivots around largest elements in array.

(*iv*) Those variables not in the basis after gaussian elimination are passed to the right side and stepped through specified range. Resistance values are printed for each setting of the free variables. Each set of resistance values will satisfy the given constraints.

Four examples demonstrate the flexibility of the method. Suppose in the circuit of Fig. 11 one wishes to choose R_1 and R_2 to provide

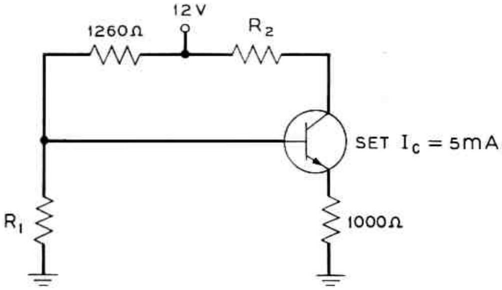


Fig. 11 — Transistor design problem.

a collector current of 5 mA. A CFE of value 0.005 is placed in series with the collector and the circuit of Fig. 12 is fed into the program as in Table I. After the program sets up the equations and performs the gaussian elimination, it prints, that the voltage at node 3 is free. It can be arbitrarily chosen to generate sets of solutions.

This free voltage is then, at the user's request, stepped from 7 volts to 10 volts in 1 volt increments. Combinations of R_1 and R_2 which provide a collector current of 5 mA are printed in Table I. To verify these results the program DCANAL[7] was used to determine the transistor collector current for the fifth set of resistor values. As the table shows, the collector current is 5 mA.

A second example involves simultaneously constraining $I_c = 5$ mA and VCE = 5 volts. As Fig. 13 shows, R_1, R_2, and R_3 are variable. The network with a VFE and CFE imbedded is shown in Fig. 14, and the results given in Table II. Verification of the first set of resistance values is given. Observe that $I_c = 5$ mA and VCE = 5 volts.

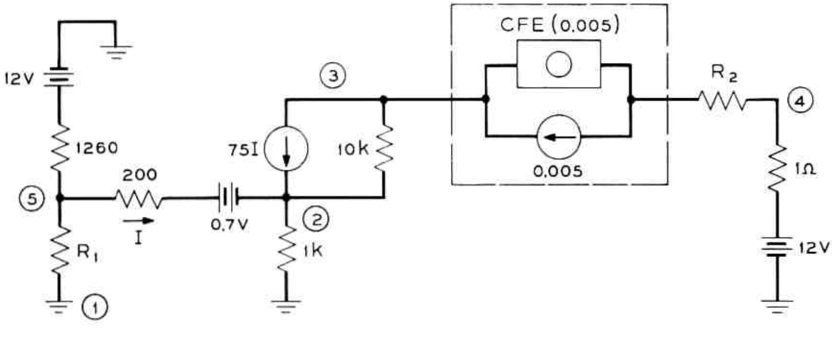


Fig. 12 — Network after transistor modelling and singular imbedding.

Table I—Printout of the Session to Solve the Circuit of Figure 11

```
 TYPE NO. OF BRANCHES,NODES,CONTROLLED SOURCES,BATTERIES,CURRENT SOURCES
A=7 5 1 3 Ø
TYPE BRANCH RESISTANCES
B=1. 1. 1260. 200. 1.E3 1.E4 1.
TYPE FOR EACH BRANCH: INITIAL NODE,FINAL NODE,BATTERY NO.
C=1 5 1   4 3 1   1 5 2   5 2 3   1 2 1   3 2 1   1 4 2
TYPE VALUES OF BATTERIES
D=Ø. 12. -.7
TYPE FOR EACH CONTROLLED SOURCE: BRANCH NO. AND CONTROLLING BRANCH NO.
E=6 4
TYPE VALUES OF BETAS
F=75.

OPTION COMMANDS=DESIGN R

TYPE NO. VAR. RESISTANCES,NO. VOLTAGE CONSTRAINTS, AND NO CURRENT CONSTRAINTS
I=2 Ø 1
TYPE BRANCH NO. OF VARIABLE RESISTANCES
J=1 2
TYPE BRANCH CURRENTS BEING CONSTRAINED
M=6
TYPE VALUE OF EACH CURRENT BEING CONSTRAINED
N=.ØØ5

THE FOLLOWING NODE VOLTAGES ARE FREE
       3

ENTER LOWER LIMIT AND INCREMENT FOR EACH FREE VARIABLE AND NO. OF SETTINGS
O=7. 1. 4

THE FREE VARIABLE= 7.
R( 1)=1.185Ø722E+Ø3
R( 2)=9.9ØØØØØ5E+Ø2

THE FREE VARIABLE= 8.
R( 1)=1.1841104E+Ø3
R( 2)=7.99ØØØØ3E+Ø2

THE FREE VARIABLE= 9.
R( 1)=1.1831496E+Ø3
R( 2)=5.99ØØØØØØE+Ø2

THE FREE VARIABLE= 1Ø.
R( 1)=1.1821898E+Ø3
R( 2)=3.9899998E+Ø2

DESIGN COMMAND=KEEP
ENTER VALUES OF FREE VARIABLES FOR DESIRED SET
=1Ø.

OPTION COMMAND=TRAN ALL

         VCE        IC
TRANS #
  1   4.93985Ø2  4.9999999E-Ø3
```

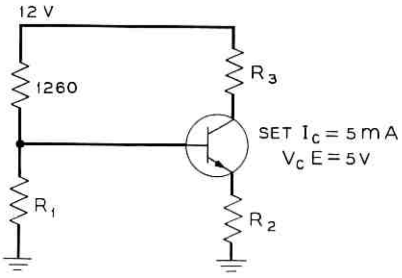


Fig. 13 — Transistor design problem.

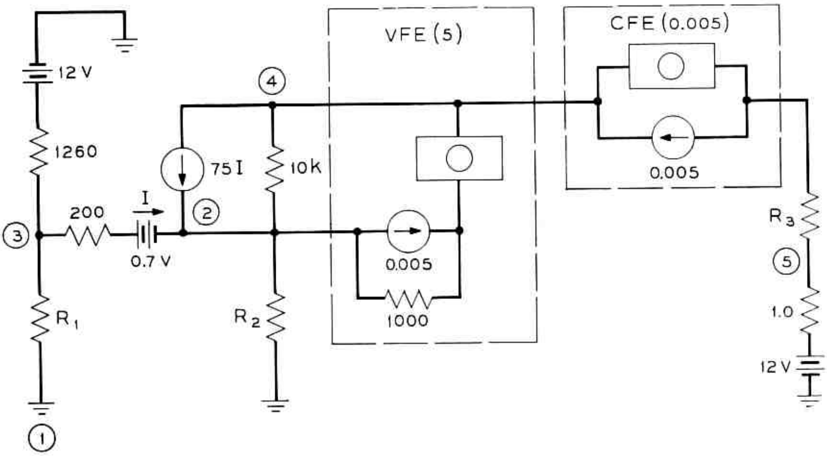


Fig. 14 — Network after transistor modelling and singular imbedding.

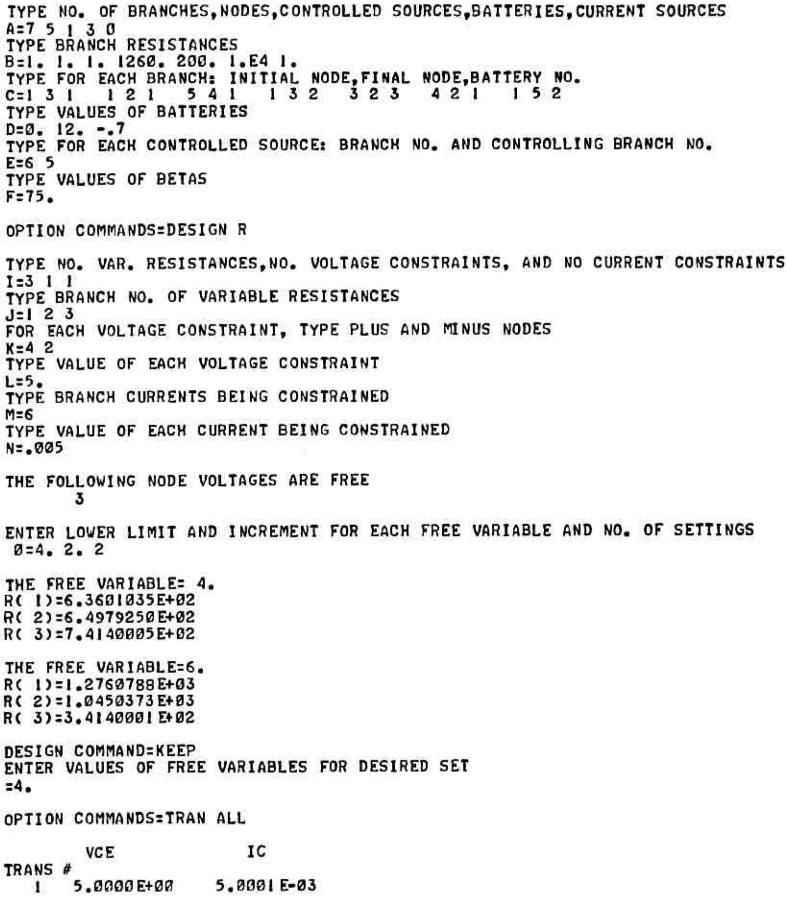
A third example involves the rather complex three transistor circuit illustrated in Fig. 15. The imbedding of VFE's and CFE's to constrain collector emitter voltages to 5 volts, and collector currents to 10 mA is shown.

Table III illustrates the results of a computer solution to the problem by the method of singular imbedding. Observe that currents through variable resistors 10 and 14 can be arbitrarily chosen and sets of resistors R_{10} through R_{18} generated. Four such sets are presented in Table III. Observe the results of an analysis indicating one such set properly biases the network. Table IV presents the results of an optimization program, based on pattern search,[1] to bias the network, for which forty-eight exploratory moves and 105 pattern moves were required. Each exploratory move involves between eight and 16 circuit analyses. Each pattern move involves an average of four analyses. Thus, approximately 1000 matrix inversions are required. Since each inversion involves $(n^3)/3$ operations, the number of operations to generate a single bias network $\approx$243,000.

Singular imbedding increases the number of nodes from 9 to 15. However, only one matrix inversion is required to generate a solution. Thus the number of operations $\cong n^3/3 \cong 1125$.

Singular imbedding increases the efficiency in finding a solution to this problem by a factor of approximately 200. What is even more important is the ease with which equivalent networks are generated. Each equivalent network is generated by a matrix multiplication of

TABLE II—PRINTOUT OF THE SESSION TO SOLVE THE CIRCUIT OF FIGURE 13



```
TYPE NO. OF BRANCHES,NODES,CONTROLLED SOURCES,BATTERIES,CURRENT SOURCES
A=7 5 1 3 Ø
TYPE BRANCH RESISTANCES
B=1. 1. 1. 126Ø. 2ØØ. 1.E4 1.
TYPE FOR EACH BRANCH: INITIAL NODE,FINAL NODE,BATTERY NO.
C=1 3 1   1 2 1   5 4 1   1 3 2   3 2 3   4 2 1   1 5 2
TYPE VALUES OF BATTERIES
D=Ø. 12. -.7
TYPE FOR EACH CONTROLLED SOURCE: BRANCH NO. AND CONTROLLING BRANCH NO.
E=6 5
TYPE VALUES OF BETAS
F=75.

OPTION COMMANDS=DESIGN R

TYPE NO. VAR. RESISTANCES,NO. VOLTAGE CONSTRAINTS, AND NO CURRENT CONSTRAINTS
I=3 1 1
TYPE BRANCH NO. OF VARIABLE RESISTANCES
J=1 2 3
FOR EACH VOLTAGE CONSTRAINT, TYPE PLUS AND MINUS NODES
K=4 2
TYPE VALUE OF EACH VOLTAGE CONSTRAINT
L=5.
TYPE BRANCH CURRENTS BEING CONSTRAINED
M=6
TYPE VALUE OF EACH CURRENT BEING CONSTRAINED
N=.ØØ5

THE FOLLOWING NODE VOLTAGES ARE FREE
       3

ENTER LOWER LIMIT AND INCREMENT FOR EACH FREE VARIABLE AND NO. OF SETTINGS
 Ø=4. 2. 2

THE FREE VARIABLE= 4.
R( 1)=6.36Ø1Ø35E+Ø2
R( 2)=6.497925ØE+Ø2
R( 3)=7.414ØØØ5E+Ø2

THE FREE VARIABLE=6.
R( 1)=1.276Ø788E+Ø3
R( 2)=1.Ø45Ø373E+Ø3
R( 3)=3.414ØØØ1E+Ø2

DESIGN COMMAND=KEEP
ENTER VALUES OF FREE VARIABLES FOR DESIRED SET
=4.

OPTION COMMANDS=TRAN ALL

         VCE               IC
TRANS #
   1   5.ØØØØE+ØØ    5.ØØØ1E-Ø3
```

the vector of free variables, which is stepped through a specified range, and the matrix of vectors not taken into the basis after triangulation. In this case the matrix is 19 × 2 and the vector of free variables is 2 × 1. Each multiplication involves 2 × 19 = 38 operations. This means that up to 14,000 equivalent networks can be generated with the same number of operations needed to give one solution by optimization techniques.

The value of singular imbedding is apparent here. Only one equation need be solved, and from it, all solutions are generated.

As a fourth example, a network was designed for a specified z_{11}

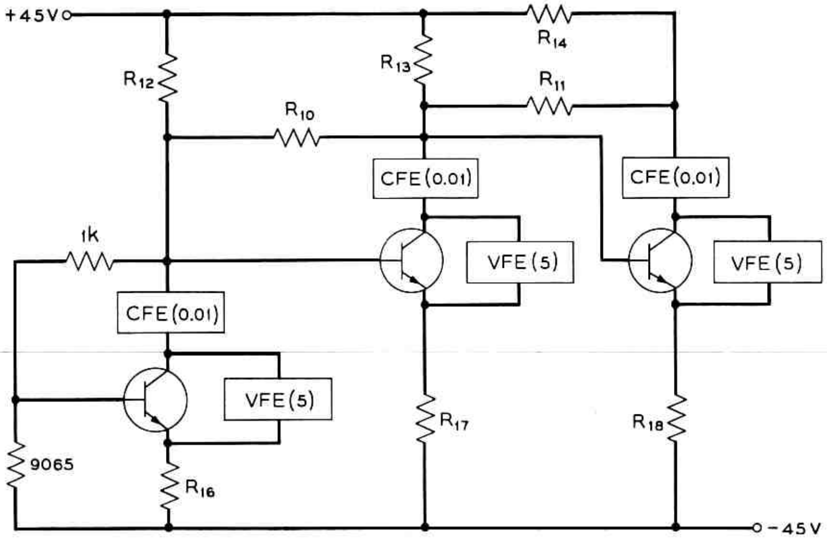


Fig. 15 — Three transistor network with VFE's and CFE's imbedded for desired biasing.

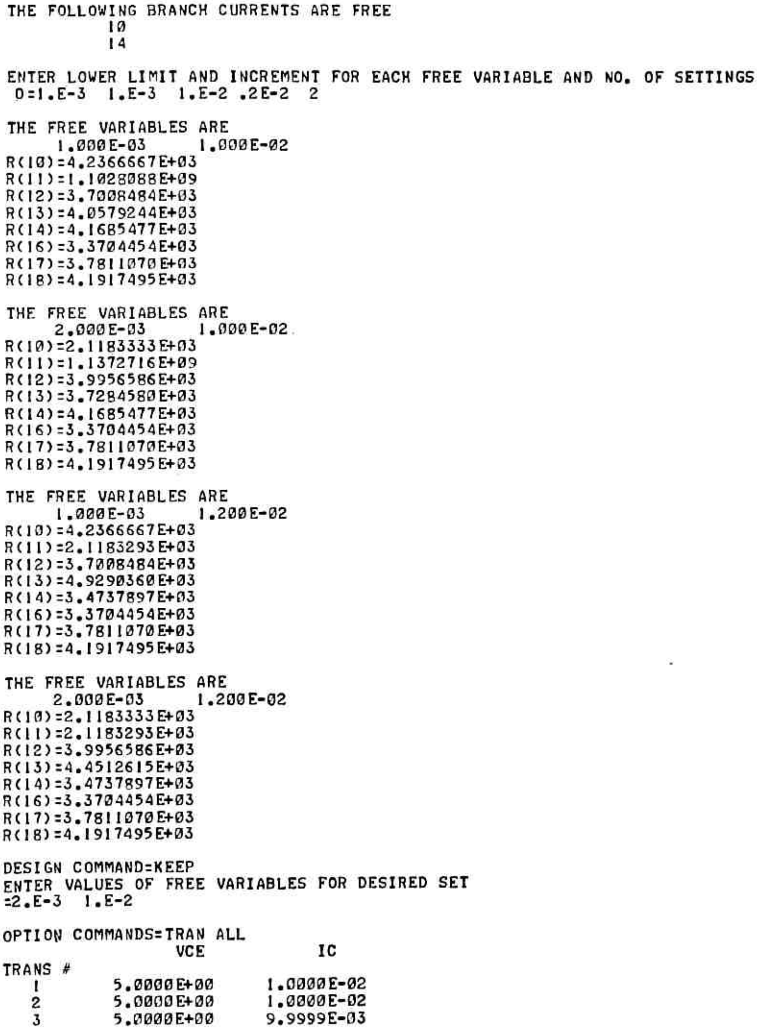
and z_{21} simultaneously. The circuit is given in Fig. 16. R_1, R_2, and R_3 are to be selected to give $z_{11} = 2/3$ and $z_{21} = 1/3$. After proper imbedding of VFE's and CFE's the network of Fig. 17 results. Table V gives the results of a computer run to design the circuit. The third set, $R_1 = R_2 = R_3 = 2$ is shown to give the desired z-parameters through the $Y - \Delta$ transformation of Fig. 18.

VIII. RESISTOR CONSTRAINTS

In many design problems it is desirable to constrain the values that the parameters take to lie within certain limits. For example, in biasing a transistor network, although solutions in which some resistors are negative are mathematically correct, in practice such networks are unacceptable.

If the designer has a good feeling for the circuit he is working with, his choice of the free variables resulting from gaussian elimination with maximum pivoting will usually yield resistors with positive values. There are, however, instances involving multiple feedback paths where intuition cannot always be relied upon. In these instances it is possible that the values given by the designer to the free variables yield negative resistances. Furthermore, it may be

TABLE III—PRINTOUT OF THE SESSION TO SOLVE THE CIRCUIT OF FIGURE 15



```
THE FOLLOWING BRANCH CURRENTS ARE FREE
          1Ø
          14

ENTER LOWER LIMIT AND INCREMENT FOR EACH FREE VARIABLE AND NO. OF SETTINGS
 0=1.E-3  1.E-3  1.E-2 .2E-2  2

THE FREE VARIABLES ARE
    1.ØØØE-Ø3      1.ØØØE-Ø2
R(1Ø)=4.2366667E+Ø3
R(11)=1.1Ø28Ø88E+Ø9
R(12)=3.7ØØ8484E+Ø3
R(13)=4.Ø579244E+Ø3
R(14)=4.1685477E+Ø3
R(16)=3.37Ø4454E+Ø3
R(17)=3.78111Ø7ØE+Ø3
R(18)=4.1917495E+Ø3

THE FREE VARIABLES ARE
    2.ØØØE-Ø3      1.ØØØE-Ø2
R(1Ø)=2.1183333E+Ø3
R(11)=1.1372716E+Ø9
R(12)=3.9956586E+Ø3
R(13)=3.728458ØE+Ø3
R(14)=4.1685477E+Ø3
R(16)=3.37Ø4454E+Ø3
R(17)=3.78111Ø7ØE+Ø3
R(18)=4.1917495E+Ø3

THE FREE VARIABLES ARE
    1.ØØØE-Ø3      1.2ØØE-Ø2
R(1Ø)=4.2366667E+Ø3
R(11)=2.1183293E+Ø3
R(12)=3.7ØØ8484E+Ø3
R(13)=4.929Ø36ØE+Ø3
R(14)=3.4737897E+Ø3
R(16)=3.37Ø4454E+Ø3
R(17)=3.78111Ø7ØE+Ø3
R(18)=4.1917495E+Ø3

THE FREE VARIABLES ARE
    2.ØØØE-Ø3      1.2ØØE-Ø2
R(1Ø)=2.1183333E+Ø3
R(11)=2.1183293E+Ø3
R(12)=3.9956586E+Ø3
R(13)=4.4512615E+Ø3
R(14)=3.4737897E+Ø3
R(16)=3.37Ø4454E+Ø3
R(17)=3.78111Ø7ØE+Ø3
R(18)=4.1917495E+Ø3

DESIGN COMMAND=KEEP
ENTER VALUES OF FREE VARIABLES FOR DESIRED SET
=2.E-3  1.E-2

OPTION COMMANDS=TRAN ALL
                  VCE           IC
TRANS #
   1        5.ØØØØE+ØØ    1.ØØØØE-Ø2
   2        5.ØØØØE+ØØ    1.ØØØØE-Ø2
   3        5.ØØØØE+ØØ    9.9999E-Ø3
```

difficult to explore the space of the free variables looking for regions where all the resistors are positive.

One possibility for finding positive resistor regions is to use an optimization technique in which, considering the free variables as adjustable parameters, the sum of the absolute magnitudes of the negative resistors is reduced to a minimum. If there exist solutions with all resistors positive, the minimum (zero) hopefully would be found automatically by the optimization routine.

This optimization is more efficient than solving the problem by exploring a space in which all the variable resistors are parameters to be adjusted.[1]

Although the method given has been tried with success, a superior method having several advantages over the one proposed is explained in Section IX. The method avoids some of the most important problems associated with nonlinear programming.

Some of these problems are:

(*i*) The routine may get trapped in local minima.

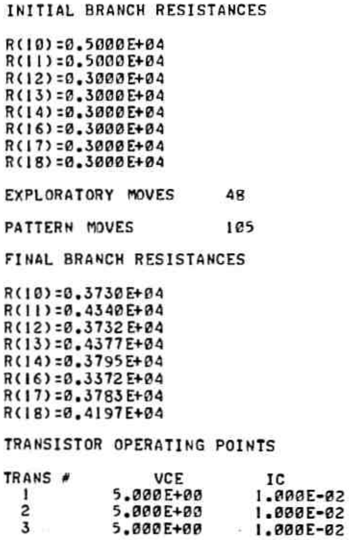
(*ii*) Depending on the shapes of the surfaces involved and on the methods used the convergence towards the minimum may be very slow.

(*iii*) If the optimization is with constraints the nonlinear constraints are usually difficult to handle.

If it were possible to reduce the problem to a linear programming problem, the following would have been gained:

(*i*) If the problem has a finite minimum it will be achieved in a

TABLE IV—PRINTOUT OF OPTIMIZATION PROGRAM



```
INITIAL BRANCH RESISTANCES

R(10)=0.5000E+04
R(11)=0.5000E+04
R(12)=0.3000E+04
R(13)=0.3000E+04
R(14)=0.3000E+04
R(16)=0.3000E+04
R(17)=0.3000E+04
R(18)=0.3000E+04

EXPLORATORY MOVES      48

PATTERN MOVES          105

FINAL BRANCH RESISTANCES

R(10)=0.3730E+04
R(11)=0.4340E+04
R(12)=0.3732E+04
R(13)=0.4377E+04
R(14)=0.3795E+04
R(16)=0.3372E+04
R(17)=0.3783E+04
R(18)=0.4197E+04

TRANSISTOR OPERATING POINTS

TRANS #         VCE           IC
  1          5.000E+00     1.000E-02
  2          5.000E+00     1.000E-02
  3          5.000E+00     1.000E-02
```

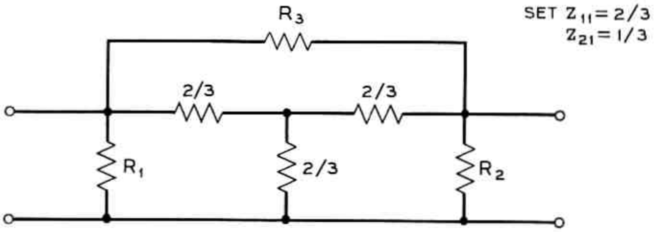


Fig. 16 — Z-parameter design problem.

finite number of steps. No local minima which are not also global minima exist.

(*ii*) Algorithms exist which converge to the minimum efficiently.

(*iii*) The linear constraints generally complicate the problem only moderately.

In Section IX the problem of biasing transistor networks is reduced to a linear programming problem.

IX. APPLIED LINEAR PROGRAMMING

Let us start by assuming a network in which the designer knows the correct signs of the node voltages with respect to the datum and the direction of the currents in the variable resistors. Generally the former is an easy task since it only involves knowing the nodes with the lowest potential. If this node is chosen as the datum, all the node voltages will be positive. Knowing the correct direction of the current through the variable resistors requires a better understanding of the circuit operation. Furthermore, there may be solutions in which the current through some resistors may flow in either direction. For this reason this requirement will eventually be relaxed.

Linear programming requires the right side vector of equation (33)

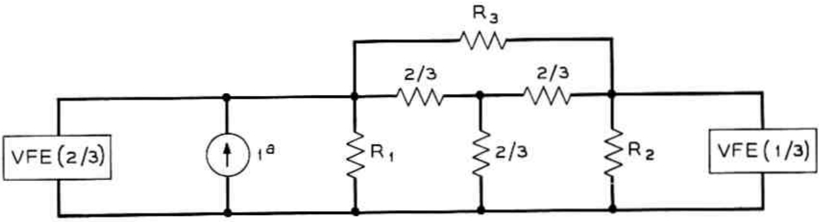
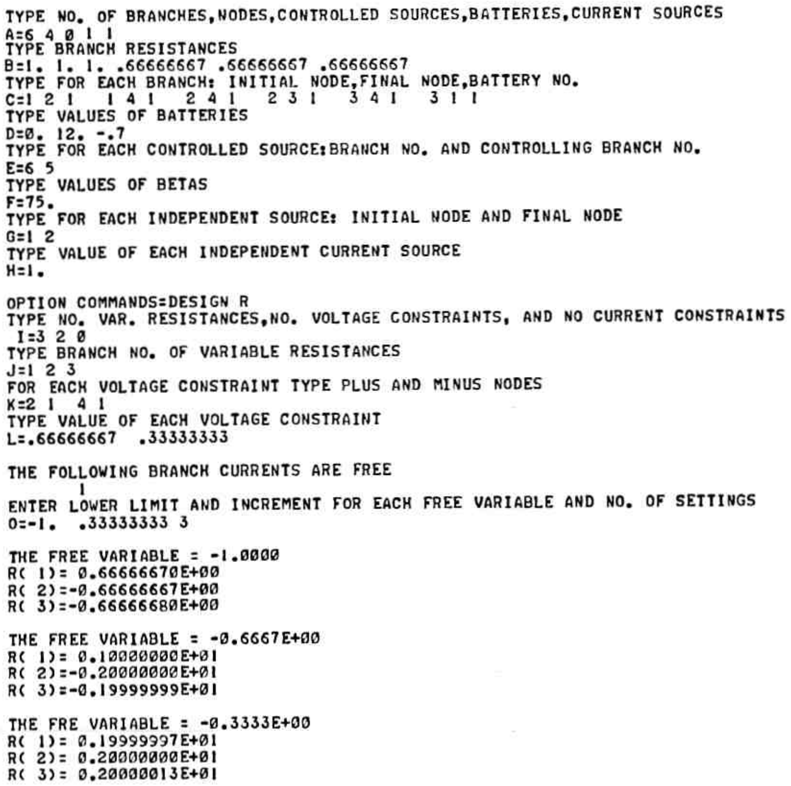
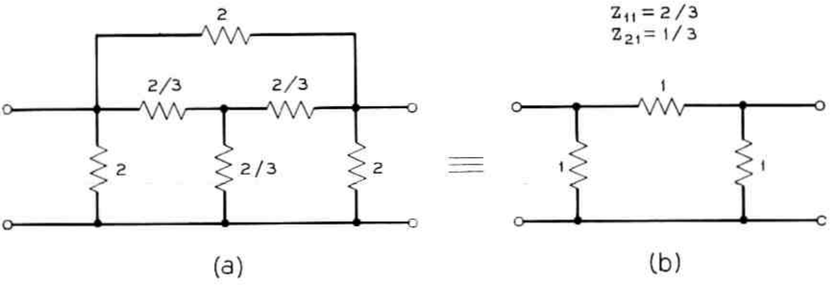


Fig. 17 — Network after singular imbedding

TABLE V—PRINTOUT OF THE SESSION TO SOLVE THE CIRCUIT OF FIGURE 16



```
TYPE NO. OF BRANCHES,NODES,CONTROLLED SOURCES,BATTERIES,CURRENT SOURCES
A=6 4 Ø 1 1
TYPE BRANCH RESISTANCES
B=1. 1. 1. .66666667 .66666667 .66666667
TYPE FOR EACH BRANCH: INITIAL NODE,FINAL NODE,BATTERY NO.
C=1 2 1   1 4 1   2 4 1   2 3 1   3 4 1   3 1 1
TYPE VALUES OF BATTERIES
D=Ø. 12. -.7
TYPE FOR EACH CONTROLLED SOURCE:BRANCH NO. AND CONTROLLING BRANCH NO.
E=6 5
TYPE VALUES OF BETAS
F=75.
TYPE FOR EACH INDEPENDENT SOURCE: INITIAL NODE AND FINAL NODE
G=1 2
TYPE VALUE OF EACH INDEPENDENT CURRENT SOURCE
H=1.

OPTION COMMANDS=DESIGN R
TYPE NO. VAR. RESISTANCES,NO. VOLTAGE CONSTRAINTS, AND NO CURRENT CONSTRAINTS
 I=3 2 Ø
TYPE BRANCH NO. OF VARIABLE RESISTANCES
J=1 2 3
FOR EACH VOLTAGE CONSTRAINT TYPE PLUS AND MINUS NODES
K=2 1  4 1
TYPE VALUE OF EACH VOLTAGE CONSTRAINT
L=.66666667  .33333333

THE FOLLOWING BRANCH CURRENTS ARE FREE
           1
ENTER LOWER LIMIT AND INCREMENT FOR EACH FREE VARIABLE AND NO. OF SETTINGS
O=-1.  .33333333 3

THE FREE VARIABLE = -1.ØØØØ
R( 1)= Ø.6666667ØE+ØØ
R( 2)=-Ø.66666667E+ØØ
R( 3)=-Ø.6666668ØE+ØØ

THE FREE VARIABLE = -Ø.6667E+ØØ
R( 1)= Ø.1ØØØØØØØE+Ø1
R( 2)=-Ø.2ØØØØØØØE+Ø1
R( 3)=-Ø.19999999E+Ø1

THE FRE VARIABLE = -Ø.3333E+ØØ
R( 1)= Ø.19999997E+Ø1
R( 2)= Ø.2ØØØØØØØE+Ø1
R( 3)= Ø.2ØØØØØ13E+Ø1
```



(a)

(b)

Fig. 18 — Verification of computer solution.

to have positive entries. This may be achieved by multiplying by -1 all those rows in equation (33) which have a negative entry in the right side vector and thus obtain the set of equations

$$\left[\begin{array}{c|c}\mathbf{H}_1 & \mathbf{H}_2\\ \hline \mathbf{0} & \mathbf{H}_3\end{array}\right]\left.\begin{array}{c}\mathbf{I}\\ \hline \mathbf{V}\end{array}\right] = \left.\begin{array}{c}\mathbf{f}\\ \hline \mathbf{0}\end{array}\right] \tag{41}$$

where

$$\mathbf{H} = \left[\begin{array}{c|c}\mathbf{H}_1 & \mathbf{H}_2\\ \hline \mathbf{0} & \mathbf{H}_3\end{array}\right]$$

is obtained from

$$\left[\begin{array}{c|c}-\mathbf{C} & \mathbf{Y}_f\mathbf{U}_c\\ \hline \mathbf{0} & \mathbf{B}\end{array}\right] \quad \text{and} \quad \left.\begin{array}{c}\mathbf{f}\\ \hline \mathbf{0}\end{array}\right] \quad \text{from} \quad \left.\begin{array}{c}\mathbf{I}_S\\ \hline \mathbf{0}\end{array}\right]$$

by possibly multiplying some rows by -1.

To force all the branch voltages to be positive let us add the constraint

$$-[\mathbf{C}]\mathbf{V}] \geqq \mathbf{0}] \tag{42}$$

where $[\mathbf{C}]$ is the matrix appearing in equation (12).

Equation (41) and inequality (42) together with the condition

$$\left.\begin{array}{c}\mathbf{I}\\ \hline \mathbf{V}\end{array}\right] \geqq \mathbf{0}] \tag{43}$$

can be looked upon as a linear programming problem in which it is desired to find the value of a positive vector satisfying a set of linear equalities and inequalities and which minimizes the linear function where

$$z = \underline{\mathbf{D}}\left.\begin{array}{c}\mathbf{I}\\ \hline \mathbf{V}\end{array}\right] \tag{44}$$

$$\underline{\mathbf{D}} = [0, 0, \cdots, 0].$$

Since the minimization of the constant zero is of no interest, all that is required is to obtain the feasible solutions of the linear programming problem.[6]

Once the feasible solutions are obtained, the fact that the solution satisfies equation (41) guarantees that the circuit is properly biased while the positivity condition on the vectors $\underline{\mathbf{I}|\mathbf{V}}$ and $-[\mathbf{C}]\mathbf{V}]$ guarantee that all the variable resistors are positive, since both the currents and voltages across them are positive.

To obtain the feasible solutions phase I of the two phase simplex method may be used.[6]

Phase I of the simplex method finds the basic positive solutions of the system of equations

$$\left[\begin{array}{c|c|c} \mathbf{H}_1 & \mathbf{H}_2 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{H}_3 & \mathbf{0} \\ \hline \mathbf{0} & -\mathbf{C} & -\mathbf{U} \end{array}\right] \left[\begin{array}{c} \mathbf{I} \\ \hline \mathbf{V} \\ \hline \mathbf{w} \end{array}\right] = \left[\begin{array}{c} \mathbf{f} \\ \hline \mathbf{0} \\ \hline \mathbf{0} \end{array}\right] \tag{45}$$

where the vector **w**] (which is constrained to be positive) is a slack vector and **U** is a unit matrix.

By denoting with **A** the matrix on the left of equation (45), with **x** the column on the left, and with **b** the column on the right side, equation (45) may be written

$$\mathbf{Ax} = \mathbf{b}. \tag{46}$$

Let the dimensions be: **A**, $m \times n$; **x**, $n \times 1$; **b**, $m \times 1$. Let **A** and [**A** | **b**] have rank r. This implies equation (46) is compatible. (The case in which this is not true is of no interest since in such case no solution—whether positive or not—exists.)

Phase I of the simplex method finds positive solutions of equation (46) for r of the variables x_i, $i = 1, 2, \cdots, r$ setting the rest of the x_j, $j = r + 1, \cdots, n$ to zero.[*] Each one of this set is a basic feasible solution. There may be several such sets for a given problem. The totality of nonnegative solutions of equation (46) is the convex hull of the basic solutions. By extending the simplex algorithm so that once a basic feasible solution is found the other basic feasible solutions are also searched for, it is possible to obtain all basic feasible solutions.

Suppose $x^1, x^2, \cdots, x^P$ are basic feasible solutions. Then any vector x satisfying

$$\left.\begin{array}{ll} & x = \lambda_1 x^1 + \lambda_2 x^2 + \cdots + \lambda_P x^P \\ \text{with} & \lambda_1, \lambda_2, \cdots, \lambda_P \geqq 0 \\ \text{and} & \lambda_1 + \lambda_2 + \cdots + \lambda_P = 1 \end{array}\right\} \tag{47}$$

is also a feasible solution.

[*] In case no nonnegative solutions to equation (46) exist, the simplex algorithm is able to detect it.

If $x^1, x^2, \cdots, x^P$ is the set of all basic feasible solutions, then all the solutions of equation (47) constitute the complete set of feasible solutions.

X. RESISTORS WITH UPPER AND LOWER BOUNDS

In the previous discussion the appearance of nonnegative resistors was precluded by adding inequality (42). Often it is desirable to impose lower and upper bounds for the resistors because the technology used to realize them requires it. For example, if tantalum thin film resistors are used it is desirable to restrict them to lie between 10 and 10^5 ohms.

Let the kth variable resistor be connected from node i to node j. The value of R_k is given by

$$R_k = \frac{V_i - V_j}{I_k}.$$

If it is desired to have this resistor lie within 10 and 10^5 ohms the following conditions are imposed

$$\frac{V_i - V_j}{I_k} \geqq 10, \qquad I_k \neq 0$$

$$\frac{V_i - V_j}{I_k} \leqq 10^5, \qquad I_k \neq 0$$

which may be rewritten (recall I_k is nonnegative)

$$\left.\begin{aligned} V_i - V_j - 10I_k &\geqq 0 \\ V_i - V_j - 10^5 I_k &\leqq 0 \end{aligned}\right\}. \tag{48}$$

If instead of equation (42) inequalities similar to equation (48) are written for all variable resistors, the resulting circuits will have all variable resistors within specified upper and lower bounds (except for the possibility $I_k = 0$, which implies an open circuit, in which case the resistor disappears altogether).

The problem of biasing of transistor networks with positive resistors is equivalent to solving

$$\left[\begin{array}{c|c} \mathbf{H}_1 & \mathbf{H}_2 \\ \hline \mathbf{0} & \mathbf{H}_3 \end{array}\right] \left[\begin{array}{c} \mathbf{I} \\ \hline \mathbf{V} \end{array}\right] = \left[\begin{array}{c} \mathbf{f} \\ \hline \mathbf{0} \end{array}\right]$$

$$[\mathbf{D}_1 \mid -\mathbf{C}] \left[\begin{array}{c} \mathbf{I} \\ \hline \mathbf{V} \end{array}\right] \geqq \left[\begin{array}{c} \mathbf{0} \\ \hline \mathbf{0} \end{array}\right] \tag{49}$$

$$[\mathbf{D}_2 \mid -\mathbf{C}] \left[\begin{array}{c}\mathbf{I}\\ \hline \mathbf{V}\end{array}\right] \leqq \left[\begin{array}{c}\mathbf{0}\\ \hline \mathbf{0}\end{array}\right]$$

$$\left[\begin{array}{c}\mathbf{I}\\ \hline \mathbf{V}\end{array}\right] \geqq \left[\begin{array}{c}\mathbf{0}\\ \hline \mathbf{0}\end{array}\right]$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal matrices whose diagonal elements contain the minima and maxima for the variable resistors. By adding positive slack vectors $\mathbf{w}_1$ and $\mathbf{w}_2$, equation (49) is equivalent to

$$\left[\begin{array}{c|c|c|c}\mathbf{H}_1 & \mathbf{H}_2 & \mathbf{0} & \mathbf{0}\\ \hline \mathbf{0} & \mathbf{H}_3 & \mathbf{0} & \mathbf{0}\\ \hline \mathbf{D}_1 & -\mathbf{C} & -\mathbf{U} & \mathbf{0}\\ \hline \mathbf{D}_2 & -\mathbf{C} & \mathbf{0} & \mathbf{U}\end{array}\right] \left[\begin{array}{c}\mathbf{I}\\ \hline \mathbf{V}\\ \hline \mathbf{w}_1\\ \hline \mathbf{w}_2\end{array}\right] = \left[\begin{array}{c}\mathbf{f}\\ \hline \mathbf{0}\\ \hline \mathbf{0}\\ \hline \mathbf{0}\end{array}\right] \tag{50}$$

where $\mathbf{U}$ is a unit matrix and the vector on the left is restricted to be nonnegative.

XI. RELAXING SIGN CONDITIONS

So far it has been assumed that the direction of the current flow in variable resistors is known beforehand. This condition may not hold for some cases and hence it is desirable to relax it.

When a variable in a linear programming problem is not required to be positive it is customary to write it as the difference of two positive quantities. Thus if I_k and $V_j - V_k = V$ are not required to be positive one may write

$$I_k = I_{k'} - I_{k''}$$

$$V_l = V_{l'} - V_{l''}$$

where $I_{k'}$, $I_{k''}$, $V_{l'}$, $V_{l''} \geqq 0$.

A current I_k of variable sign may be restricted to have a magnitude no less than $I_{ok} \geqq 0$ by imposing the pair of conditions*

$$I_k \geqq I_{ok} \quad \text{or} \quad -I_k \geqq I_{ok} \, . \tag{51}$$

Likewise a branch voltage V_l of variable sign across a resistor may be

* The constraint set on the currents is not convex, therefore it is necessary to solve the problem twice, once with each inequality, and take the union of the two solutions. If n variable resistors may have currents flowing in either direction, the solution will be the union of the solutions of 2^n problems in which all the combinations of the inequalities are used.

restricted to have a magnitude no greater than $V_{ok} \geqq 0$ by imposing the pair of conditions

$$V_l \leqq V_{ok} \quad \text{and} \quad -V_l \leqq V_{ok} . \tag{52}$$

If V_l is the voltage across the kth resistor and I_k its current, then inequalities (51) and (52) insure that the magnitude of the kth resistor satisfies

$$|R_k| \leqq \frac{V_{ok}}{I_{ok}}.^{*} \tag{53}$$

The resistor R_k may be negative or positive. However, if each variable resistor is made of two resistors in series one of value V_{ok}/I_{ok} and the second to be determined by the computer subject to equation (53), the series combination of the two resistors will never be negative. This constitutes a technique for guaranteeing positive variable resistors without previous knowledge of the directions of current flows.**

The method described can also handle circuits with variable resistors whose values lie within upper and lower limits. If $R_{k_{\min}}$ and $R_{k_{\max}}$ are the minimum and maximum values allowed for the kth variable resistor, the fixed series resistor should be

$$R_{kf} = R_{k_{\min}} + V_{ok}/I_{ok} \tag{54}$$

with V_{ok}/I_{ok} chosen such that

$$V_{ok}/I_{ok} = R_{k_{\max}} - R_{k_{\min}} . \tag{55}$$

The value of $R_{k_{\min}}$ may be zero. Thus, a resistor may disappear as a short circuit. If instead of bounding the value of a resistance from above, the value of an admittance is bounded, a dual method may be used to guarantee positive resistors.

Instead of equations (51) and (52) the following restrictions are imposed

$$I_k \leqq I_{ok} \quad \text{and} \quad -I_k \leqq I_{ok} , \tag{56}$$

$$V_l \geqq V_{ok} \quad \text{or} \quad -V_l \geqq V_{ok} .\dagger \tag{57}$$

* Both V_{ok} and I_{ok} are variables in the linear program which will be determined by the simplex algorithm. The ratio is constrained by a linear inequality $I_{ok} \mid R_k \mid - V_{ok} \leq 0$, where $\mid R_k \mid$ is given.

** Another approach is to reverse the reference direction of the current and voltage drop across each variable resistor and apply the methods of the previous section. If n variable resistors may have currents flowing in either direction it is necessary to consider 2^n possibilities.

† See footnote to equation 51.

These guarantee that

$$|G_k| \leqq \frac{I_{ok}}{V_{ok}} \tag{58}$$

where $G_k = 1/R_k$. G_k may be positive or negative if each variable resistor is made of two resistors in parallel, one of admittance I_{ok}/V_{ok} and the second to be determined by the computer subject to equation (58). However, the parallel combination of the resistors will never be negative.

The dual method can also handle circuits with variable resistors whose admittance lies within upper and lower limits $G_{k_{\max}}$ and $G_{k_{\min}}$. The value of $G_{k_{\min}}$ may be zero. Thus a resistor may disappear as an open circuit.

XII. CHOOSING TOPOLOGY BY COMPUTER

As already pointed out, Phase I of the simplex method obtains the basic feasible solutions of a set of linear equations. The set of equations may come from a set of equalities and inequalities to which slack variables have been added. Usually the number of variables (including slack variables) is greater than the number of equations and the system is redundant. If r is the rank of the system and n is the number of variables (including slack variables), at least n-r variables are set to zero in obtaining a basic feasible solution. Some of the variables set to zero may be node voltages or variable resistor currents. If a node voltage is set to zero, the corresponding node is grounded. If a variable resistor current is set to zero, the corresponding resistor disappears as an open circuit. If a slack variable is set to zero, the inequality constraints are met with equalities.

For example, for equation (50) if the kth entry of w, is zero, the kth resistor acquires its minimum allowed value.

One way of viewing equation (50) is to consider the columns of the matrix on the left as elements of a vector space and the entries of the column multiplying the matrix as those positive coefficients which synthesize the column on the right in the form of a linear combination of the columns of the matrix. A final tableau of Phase I of the simplex method will contain a number of independent unit columns (with all entries zero except one) equal to the rank of the matrix on the left side of equation (50). The unit columns are obtained by the special gaussian reduction provided by the simplex algorithm. Each column corresponds to a variable in the column multiplying the matrix of equation (50). Those variables whose corresponding columns are not unit columns are set to zero.

If a set of columns corresponding to the currents through a set of variable resistors are linearly dependent, one or more of the currents will be set to zero. This implies the disappearance of a resistor as an open circuit. The choice of which resistors disappear is automatically determined with the aid of the simplex algorithm, so that the nonzero currents acquire positive values (if such a choice exists). If two columns of the matrix of equation (50), corresponding to currents through variable resistors, are linearly dependent it means that Kirchhoff's voltage and current law may be satisfied with one of the currents zero, making one of the resistors unnecessary.

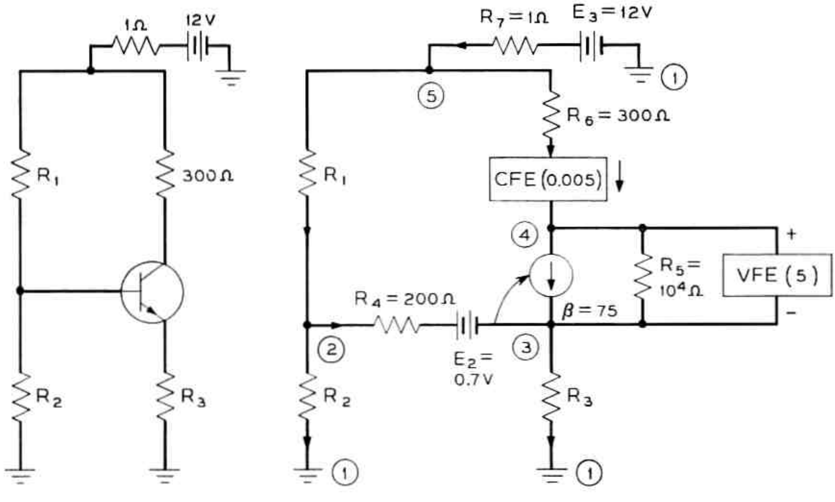
The above argument provides a method for letting a computer program choose the topology and resistor values of a dc network in which certain voltages and currents are imposed by CFE's and VFE's. One connects an excess of resistors between different nodes (including additional internal nodes if desired). By using a linear programming formulation some node voltages and variable resistor currents are set to zero by the computer program, thus determining a set of "linearly independent positive resistors" that satisfy all the circuit equations.

XIII. EXAMPLES

Consider the circuit of Fig. 19(a). The equivalent circuit is shown in Fig. 19(b) with a VFE and CFE in place. As indicated on Fig. 19(b) it is desired to impose on the transistor a collector current of 5 mA and a collector-emitter voltage of 5 volts. The resistors marked R_1, R_2 and R_3 are variable.

The nodal equations for the circuit after the effect of the nullators introduced by the VFE's and CFE's are taken into consideration are, in matrix form

$$\begin{bmatrix} -1. & 1. & 0. & 0.005 & -0.005 & 0. & -1 \times 10^{-10} \\ 0. & 0. & 1. & -0.38 & 0.3811 & -0.0011 & 0. \\ 0. & 0. & 0. & 0. & 0. & 0.00333 & 0. \\ 1. & 0. & 0. & -1. \times 10^{-10} & 0. & 0. & 1. \\ 0. & 0. & 0. & 0. & -0.001 & 0.001 & 0. \\ 0. & 0. & 0. & 0.375 & -0.3751 & 0.0001 & 0. \end{bmatrix}$$



(a) CIRCUIT (b) MODEL

Fig. 19 — Circuit biased with constrained singular imbedding; (a) circuit, (b) model.

$$\cdot \begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ V_2 \\ V_3 \\ V_4 \\ V_5 \end{bmatrix} = \begin{bmatrix} 0.0035 \\ -0.27 \\ 0.035 \\ 12. \\ 0.005 \\ 0.2675 \end{bmatrix} \cdot \qquad (59a)$$

By multiplying the second row by −1, the entry −0.27 in the right side vector is made positive. (As indicated above, linear programming assumes the right side vector is nonnegative). Notice that since the matrix in equation (59a) is 6 × 7, we therefore generally expect a one parameter infinity of solutions. If the system of equations were solved using the simplex method (with arbitrary cost coefficients), solutions in which all the variables acquire non-negative values may be obtained. Resistors R_2 and R_3, which are grounded, will automatically be positive. However the voltage differences across ungrounded resistors may turn out to be negative, yielding negative resistances. To assure a non-negative voltage difference across R, the following additional constraint will be imposed

$$V_5 - V_2 \geqq 0$$

which may also be written*

$$V_2 - V_5 \leqq 0. \tag{59b}$$

There are two basic feasible solutions to this problem:

$$\begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ V_2 \\ V_3 \\ V_4 \\ V_5 \end{bmatrix} = \begin{bmatrix} 5.788 & & 5.9966425 \times 10^{-5} \\ 5.78794 & & 0. \\ 5.060073 \times 10^{-3} & & 5.060073 \times 10^{-3} \\ 6.212 & , & 6.212 \\ 5.5 & & 5.5 \\ 10.5 & & 10.5 \\ 6.212 & & 11.99994 \end{bmatrix}. \tag{60}$$

The first basic solution yields the following set of resistors

$$\begin{aligned} R_1 &= \frac{V_5 - V_2}{I_1} = \frac{6.212 - 6.212}{5.788} = 0 \text{ ohms} \\ R_2 &= \frac{V_2}{I_2} = \frac{6.212}{5.78794} = 1.073266 \text{ ohms} \\ R_3 &= \frac{V_3}{I_3} = \frac{5.5}{5.060054 \times 10^{-3}} = 1086.941 \text{ ohms.} \end{aligned} \tag{61}$$

R_1 is a short circuit.

The second basic solution yields the set

$$R_1 = 97008.514, \qquad R_2 = \infty, \qquad R_3 = 1086.941.$$

R_2 is an open circuit. Notice also that R_3 is the same for both solutions. This is expected since the voltage of node 3 is virtually fixed by the requirements.

The totality of the solutions with non-negative voltage differences across the variable resistors may be written, according to equation (47)

$$x = \lambda x^1 + (1 - \lambda)x^2$$

where x^1 and x^2 are the basic feasible solutions of equation (60), and $0 \leqq \lambda \leqq 1$.

* When the right side of an inequality is zero, it is preferable to write it as a $\leqq$ inequality because the corresponding slack variable may be used as an artificial variable with savings on the size of the matrix to be manipulated.

Choosing $\lambda = \frac{1}{2}$ yields

$$x = \begin{bmatrix} 2.89402 \\ 2.89397 \\ 5.060073 \times 10^{-3} \\ 6.212 \\ 5.5 \\ 10.5 \\ 9.10597 \end{bmatrix}$$

which yields the set of resistors

$$R_1 = 0.997906, \qquad R_2 = 2.146532, \qquad R_3 = 1086941.$$

A continuous set of equivalent circuits, which achieve the requirements exactly and which have positive resistances, is obtained by varying λ between 0 to 1.

Suppose now that further considerations require that R_1 lie between 1000 and 2000 ohms. By replacing (59b) by

$$\frac{V_5 - V_2}{I_1} \geqq 1000 \qquad \text{and} \qquad \frac{V_5 - V_2}{I_1} \leqq 2000$$

which may be written

$$\begin{aligned} 1000\, I_1 - V_5 + V_2 &\leqq 0 \\ -2000\, I_1 + V_5 - V_2 &\leqq 0 \end{aligned} \tag{62}$$

the resistor R_1 is forced to remain between 1000 and 2000 ohms.

When the new problem is solved the basic feasible solutions are

$$x^1 = \begin{bmatrix} 5.782218 \times 10^{-3} \\ 5.722218 \times 10^{-3} \\ 5.060073 \times 10^{-3} \\ 6.212 \\ 5.5 \\ 10.5 \\ 11.99422 \end{bmatrix}, \qquad x^2 = \begin{bmatrix} 2.892554 \times 10^{-3} \\ 2.832554 \times 10^{-3} \\ 5.060073 \times 10^{-3} \\ 6.212 \\ 5.5 \\ 10.5 \\ 11.99711 \end{bmatrix}.$$

The basic feasible solutions yield the following sets of resistors

$$x^1 : R_1 = 1000., \qquad R_2 = 1085.593, \qquad R_3 = 1086.941$$
$$x^2 : R_1 = 2000., \qquad R_2 = 2193.074, \qquad R_3 = 1086.941.$$

Notice that R_1 acquired its allowable extreme values in each basic feasible solution.

Other sets of resistances may be obtained by convex combinations of the two basic feasible solutions.

As an example in which the topology of a circuit is determined by the computer, consider the circuit of Fig. 16 in which R_1, R_2, and R_3 are to be selected to give $z_{11} = \frac{2}{3}$ and $z_{21} = \frac{1}{3}$. The example was previously solved without linear programming techniques. Several solutions appear in Table V. By maximizing the negatives of the currents in the resistors, those currents which may be set to zero by taking them out of the basis for a basic feasible solution will be converted into open circuits. After the effect of the nullators introduced by the VFE's is accounted for, the matrix corresponding to the circuit of Fig. 16 is 5 × 6. We therefore expect a one parameter infinity of solutions and two basic feasible solutions which are

$$x^1 = \begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 5.9652404 \times 10^{-3} \\ 5.965238 \times 10^{-3} \\ 44.096881 \\ 0.3333333 \\ 0.3333333 \end{bmatrix}, \qquad x^2 = \begin{bmatrix} 0.28174743 \\ 4.5937138 \times 10^{-3} \\ 0.0 \\ 43.602828 \\ 0.33639577 \\ 0.3333333 \end{bmatrix}.$$

Notice that V_4 remains constant for both basic feasible solutions. This is expected since a VFE is connected from node 4 to node 1 (datum). The resistances corresponding to the basic feasible solutions are

$$x^1 : R_1 = \infty \qquad R_2 = 55.879273, \qquad R_3 = 7336.429$$
$$x^2 : R_1 = 154.4759, \qquad R_2 = 72.5629, \qquad R_3 = \infty .$$

In both basic feasible solutions one of the resistances disappeared as an open circuit. This indicates that given R_4, R_5 and R_6 with the values indicated in Fig. 16 the circuit is achievable with two topologies, each containing 5 resistors.

Let us now make R_6 a variable resistor. The nodal matrix after the elimination of the nullators is now 5 × 7. Thus we expect a two parameter infinity of solutions and at least 3 basic feasible solutions. The following sets of resistors correspond to basic feasible solutions

(i) $R_1 = \infty, \quad R_2 = \infty, \quad R_3 = 1.33333, \quad R_6 = 0.16666667$

(ii) $R_1 = \infty, \quad R_2 = 0.333333, \quad R_3 = 0.4444456, \quad R_6 = \infty$

(iii) $R_1 = 0.88888898, \quad R_2 = 1.333333, \quad R_3 = \infty, \quad R_6 = \infty$

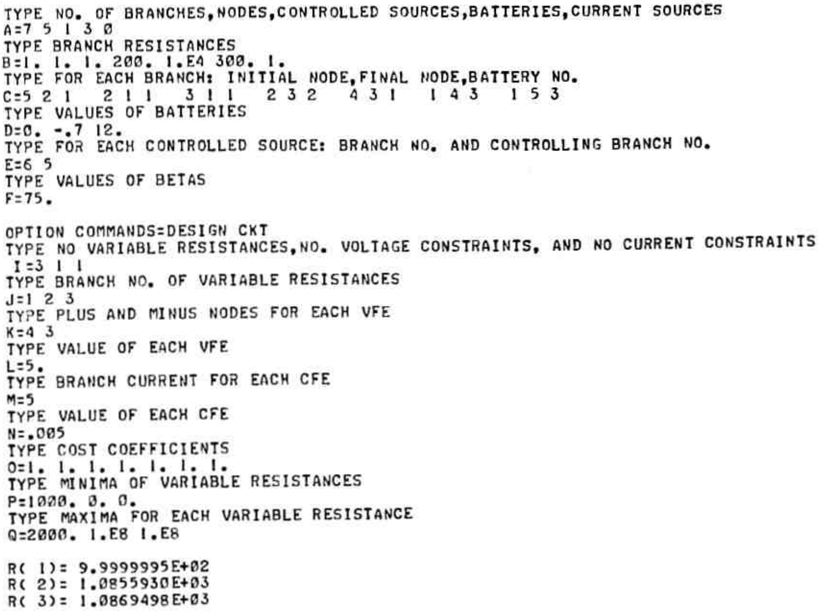
(iv) $R_1 = 1.3333336, \quad R_2 = \infty, \quad R_3 = \infty, \quad R_6 = 0.66666651.$

These sets provide four different topologies with which given two of the resistors (R_4 and R_5) a resistive network having $z_{11} = 2/3$, $z_{21} = 1/3$ may be realized.

The example illustrates how using the methods of this paper can solve the problem of realizing portions of a resistive matrix with certain elements prespecified. The prespecified elements need not be resistors but may also include controlled sources, gyrators, ideal transformers, and so on.

The methods discussed have been implemented on a time-shared

Table VI—Printout of the Session to Solve the Circuit of Figure 19



```
TYPE NO. OF BRANCHES,NODES,CONTROLLED SOURCES,BATTERIES,CURRENT SOURCES
A=7 5 1 3 Ø
TYPE BRANCH RESISTANCES
B=1. 1. 1. 2ØØ. 1.E4 3ØØ. 1.
TYPE FOR EACH BRANCH: INITIAL NODE,FINAL NODE,BATTERY NO.
C=5 2 1   2 1 1   3 1 1   2 3 2   4 3 1   1 4 3   1 5 3
TYPE VALUES OF BATTERIES
D=0. -.7 12.
TYPE FOR EACH CONTROLLED SOURCE: BRANCH NO. AND CONTROLLING BRANCH NO.
E=6 5
TYPE VALUES OF BETAS
F=75.

OPTION COMMANDS=DESIGN CKT
TYPE NO VARIABLE RESISTANCES,NO. VOLTAGE CONSTRAINTS, AND NO CURRENT CONSTRAINTS
 I=3 1 1
TYPE BRANCH NO. OF VARIABLE RESISTANCES
J=1 2 3
TYPE PLUS AND MINUS NODES FOR EACH VFE
K=4 3
TYPE VALUE OF EACH VFE
L=5.
TYPE BRANCH CURRENT FOR EACH CFE
M=5
TYPE VALUE OF EACH CFE
N=.0Ø5
TYPE COST COEFFICIENTS
O=1. 1. 1. 1. 1. 1. 1.
TYPE MINIMA OF VARIABLE RESISTANCES
P=1ØØØ. Ø. 0.
TYPE MAXIMA FOR EACH VARIABLE RESISTANCE
Q=2ØØ0. 1.E8 1.E8

R( 1)= 9.9999995E+Ø2
R( 2)= 1.Ø85593ØE+Ø3
R( 3)= 1.Ø869498E+Ø3
```

computer system. The program is conversational. A portion of a session in which a basic solution corresponding to the circuit of Fig. 19 with R_1 constrained betwen 1000 and 2000 ohms appears in Table VI.

XIV. CONCLUSIONS

The method of singular imbedding has been shown to be efficient for solving the following problem: Given a circuit with a prespecified topology, some of whose elements are prespecified, find the values of the unspecified elements which will yield desired node-pair voltages or branch currents. The unspecified element values may be restricted to lie within given upper and lower bounds.

By letting the upper and lower bounds become infinite and zero, the problem of finding the topology for the circuit may be also solved.

The method has been implemented on a time-shared computer, and several examples, including some practical transistor circuits, are given.

The usual approaches to the problems of this paper have been iterative analysis-optimization schemes. Singular imbedding requires, for a three transistor amplifier, three orders of magnitude less computation time. This makes the method appealing for time-shared applications.

Two new singular network elements, the voltage forcing element and the current forcing element, constrain node-pair voltages and branch currents without otherwise affecting the circuit. Elements of unspecified value are modeled by branches carrying unknown currents.

With the aid of these elements, the problem of design is reduced to one of analyzing a circuit containing unknown current sources and nullators. If there are more free elements than requirements, the solution space may be a linear manifold. By allowing the free circuit variables to take on a set of discrete values, sets of exact solutions to the design problem may be generated economically.

When the unspecified elements are required to lie within upper and lower bounds, the problem is one of analysis with linear inequality constraints. This may be solved efficiently using linear programming techniques.

Among the practical problems solved by singular imbedding are biasing a direct coupled transistor amplifier, designing midband gain and driving point impedance, synthesizing networks for several given admittance parameters, and determining circuit topology.

Areas being investigated include using singular imbedding in the synthesis of resistance networks (the synthesis of a single column of a specified resistance matrix has been illustrated). Synthesis of an entire resistance matrix results from the intersection in resistance space of the solution spaces for each column of the matrix. Similarly, by considering the intersection of solutions spaces for both a small signal design and a biasing design, the method may be extended to designing transistor circuits for desired small signal design and bias points simultaneously.

Although only fixed value CFE's and VFE's were used in this paper, CFE's and VFE's which may take any value within a given range may also be used. For example, a branch current may be forced to be greater than 1 mA and less than 10 mA. These elements are also useful in insuring that models for devices stay within their valid limits. For example, a transistor can be constrained to remain in the active region, for which the linear model used is valid.

For simplicity, only the case of linear dc networks has been illustrated in this paper. However, the method has usefulness in ac design, combined ac and dc design, and non-linear design. These topics will be covered elsewhere.

REFERENCES

1. Murray-Lasso, M. A. and Baker, W. D., "Computer Design of Multistage Transistor Bias Circuits," *Proc. Fifth Annual Allerton Conf.*, Monticello, Illinois, October 1967.
2. McMillan, B., "Introduction to Formal Realizability Theory," B.S.T.J., *31*, No. **2** (March 1952), pp. 217–279.
3. Newcomb, R. W., *Linear Multiport Synthesis*, New York: McGraw-Hill, 1966.
4. Tellegen, B. D. H., "La Recherche pour une Seŕie Complète d'éléments de Circuits Ideaux Non Linéaires," Rendiconti Seminario Mathematico e Fisico, Milano, *25* (1953–54), pp. 134–144.
5. Davies, A. C., "Matrix Analysis of Networks Containing Nullators and Norators," *Electronics Letters*, *2*, No. 2 (February 1966), pp. 48–49.
6. Dantzig, B. B., *Linear Programming and Extensions*, Princeton, N. J., Princeton University Press, 1963.
7. Murray-Lasso, M. A. and Kasper, F. J., "On-Line Circuit Analysis and Optimization with Commercially Available Time-Shared Computer Systems," Proceedings of the Design Automation Workshop, Washington, D. C., July 1968.

Contributors to This Issue

Allen Gersho, B.S., 1960, M.I.T.; M.S., 1961, Ph.D., 1963, Cornell University; Bell Telephone Laboratories, 1963—. During the 1966–67 academic year he was Assistant Professor of Electrical Engineering at the City College of the City University of New York. He has been engaged in research problems related to time-varying and nonlinear signal processing, synchronization of remote clocks, synthesis of distributed networks, and adaptive equalization for digital communications.

Harry Heffes, B.E.E., 1962, City College of New York; M.E.E., 1964; and Ph.D., 1968, New York University; Bell Telephone Laboratories, 1962—. Mr. Heffes has been involved in the study of linear stochastic control problems. This includes work which has been done in determining the sensitivity of filter performance to the assumed models and the approximation of complex control systems by simpler systems. Member, IEEE, Tau Beta Pi, Eta Kappa Nu.

Edward B. Kozemchak, B.S.E.E., 1966, University of Pennsylvania; M.S.E.E., 1967, Stanford University; Bell Telephone Laboratories, 1966—. Mr. Kozemchak has been engaged in developing operational amplifiers for monolithic integrated circuits, transversal filters for radar systems, and active filters. He is working in computer-aided development of design algorithms for electronic circuits. Member, Tau Beta Pi, Eta Kappa Nu.

Tien Pei Lee, B.S.E.E., 1957, National Taiwan University, Taiwan, China; M.S.E.E., 1959, Ohio State University; Ph.D., 1963, Stanford University; Bell Telephone Laboratories, 1963—. Mr. Lee participated in the research and development of solid-state microwave diodes and photodiodes. He is working on millimeter wave devices, optical modulators, mixers, and GaAs injection lasers. Members, Sigma Xi, IEEE.

Marco A. Murray-Lasso, National Mechanical and Electrical Engineer, 1960, University of Mexico; M.S.E.E., 1962, and Sc.D., 1965, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1965—. A former Associate Professor at the University of Mexico, Dr. Murray-Lasso has been engaged in communication component research, missile control systems, data reduction, and computer-aided design. Member, IEEE, Association of University Professors of Mexico, Association of Mechanical and Electrical Engineers of Mexico, Sigma Xi.

JOHN F. O'NEILL, B.S.E.E., 1959, St. Louis University; M.E.E., 1961, New York University; Ph.D., 1967, New York University; Bell Telephone Laboratories, 1959—. Mr. O'Neill has done exploration and development in digital data transmission and is now exploring common control customer telephone systems.

TADIKONDA N. RAO, B.Sc. (Physics), 1952, Madras University, India; B. Tech. (E.E.), 1957, Indian Institute of Technology, Kharagpur, India; M.S. (E.E.), 1962, University of California, Berkeley; Ph.D. (E.E.), 1967, Stanford University; Bell Telephone Laboratories, 1967—. Since joining Bell Laboratories Mr. Rao has been concerned with active and passive network theory and its applications to silicon and tantalum integrated circuits. Member, IEEE, AAAS, Sigma Xi.

STEPHEN O. RICE, B.S., 1929, D.Sc. (Hon.), 1961, Oregon State College; Graduate Studies, California Inst. of Tech., 1929–30 and 1934–35; Bell Telephone Laboratories, 1930—. In his first years at the Laboratories, Mr. Rice was concerned with nonlinear circuit theory, especially with methods of computing modulation products. Since 1935 he has served as a consultant on mathematical problems and in investigation of telephone transmission theory, including noise theory, and applications of electromagnetic theory. He was a Gordon McKay Visiting Lecturer in Applied Physics at Harvard University for the Spring, 1968, term. Fellow, IEEE.

G. H. ROBERTSON, B.Sc., 1943, University of Glasgow; after three years in the Royal Navy as an Air Radio Officer he returned to the University of Glasgow for two years and obtained a Post Graduate Certificate in Natural Philosophy; Bell Telephone Laboratories, 1948—. Until 1958 Mr. Robertson was engaged in electronics research and a variety of electron tube development projects. Since 1958 he has been working on signal propagation and processing studies in the Underwater Research and Systems Departments. Associate member, IEEE; member, AAAS.

BURTON R. SALTZBERG, B.E.E., 1954, New York University; M.S., 1955, University of Wisconsin; Eng. Sc.D., 1964, New York University; Bell Telephone Laboratories, 1957—. Mr. Saltzberg has been engaged in developing and analyzing data transmission systems. He supervises a group responsible for developing data sets for use over the telephone network. Member, IEEE, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

Irwin W. Sandberg, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. Mr. Sandberg has been concerned with analysis of military systems, synthesis and analysis of active and time-varying networks, studies of properties of nonlinear systems, and some problems in communication theory and numerical analysis. He is head of the Systems Theory Research Department. Member, IEEE, Eta Kappa Nu, Sigma Xi, Tau Beta Pi.

Philip E. Sarachik, A.B., 1953; B.S., 1954; M.S., 1955; Ph.D., 1958, Columbia University. Mr. Sarachik is a professor of electrical engineering at New York University. His main interests are in control theory with emphasis on optimal and adaptive systems. Member, Phi Beta Kappa, Tau Beta Pi, Eta Kappa Nu, Sigma Xi, Society for Industrial and Applied Mathematics.

Marvin K. Simon, B.E.E., 1960, City College of New York; M.S.E.E., 1961, Princeton University; Ph.D., 1966, New York University; Bell Telephone Laboratories, 1961–1963, 1966–1968. Mr. Simon's early work at Bell Telephone Laboratories dealt with station apparatus development including *Touch-Tone*® dialing telephone and *Picturephone*® visual telephone circuit design. Recently he has been engaged in theoretical studies of digital transmission systems. He is currently at Jet Propulsion Laboratory, Pasadena, California. Member, IEEE, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

R. D. Standley, B.S., 1957, University of Illinois; M.S., 1960, Rutgers University; Ph.D., 1966, Illinois Institute of Technology; U. S. Army Research and Development Laboratory, Fort Monmouth, N. J., 1957–1960; IIT Research Institute, Chicago, 1960–1966; Bell Telephone Laboratories, 1966—. At Fort Monmouth, Mr. Standley was project engineer on various microwave component development programs. His work at IITRI included microwave and antenna research. At Bell Telephone Laboratories he has been concerned with millimeter-wave component research. He is investigating millimeter-wave impact ionization avalanche transit time diode devices, integrated circuits, and optical modulators. Member, IEEE, Sigma Tau, Sigma Xi.

ALAN N. WILLSON, JR., B.E.E., 1961, Georgia Institute of Technology; M.S.E.E., 1965, Ph.D., 1967, Syracuse University; International Business Machines Corporation, 1961–1964; Bell Telephone Laboratories, 1967—. Mr. Willson is interested in network and systems theory. Member, IEEE, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

AARON D. WYNER, B.S., 1960, Queens College; B.S.E.E., 1960, M.S., 1961, and Ph.D., 1963, all from Columbia University; Bell Telephone Laboratories, 1963—. Mr. Wyner has been doing research in various aspects of information theory. He has also been Adjunct Associate Professor of Electrical Engineering at Columbia University and Chairman of the Metropolitan New York Chapter of the IEEE Information Theory Group. Member, IEEE, SIAM, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.

WOODSON, D. WYNN, B.E.E., 1959, M.S.E.E., 1961, Ph.D., 1965, Georgia Institute of Technology; Assistant Professor of electrical engineering at Georgia Tech, 1965–1966; Bellcomm, Inc., 1966—. Mr. Wynn has been concerned with statistical communication problems related to the manned space program. His current interest is in the design of communications systems for possible deep space missions. Member, IEEE, Tau Beta Pi, Eta Kappa Nu, Pi Mu Epsilon, Sigma Xi.