THE BELL SYSTEM

TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 48 May–June 1969 Number 5

On the Identification of Linear Time-Invariant Systems from Input-Output Data

By B. GOPINATH

(Manuscript received December 17, 1968)

This paper presents a new method for computing the parameters which determine the differential equations governing a linear time-invariant system with multiple inputs and outputs. Unlike earlier approaches the method presented does not involve computation of the impulse response. One of the main advantages of this method is its easy generalization to the case when the given data is contaminated with noise.

The identification of multiple input-output linear systems has been a problem of considerable interest because of its importance in circuit and control system theory. In circuit theory the problem is that of synthesizing a linear time invariant circuit to exhibit a prescribed input-output behavior. In control theory, however, the problem arises out of a need to model a given linear system with a suitable set of differential equations, given its input-output behavior. References 1, 2, and 3 deal with the problem of determining the parameters of the differential equation model from the impulse response. To the best of the author's knowledge, there is no published method which determines the impulse response from a finite segment of input-output data in the case of systems with more than one input and output.

I. THE "STATE INVARIANT" DESCRIPTION

In most applications of identification techniques, one is given only a record of input sequences and a record of output sequences, rather than the impulse response function. In these cases it seems best to get an internal description of the system directly from these data; that is, avoid the intermediate step of synthesizing the impulse response. In many applications the structure of systems that are being identified remains the same, while values of parameters change. Therefore, it is convenient to work in a certain coordinate frame which is fixed for the given system. Most important of all, a method of arriving at the values of parameters directly from input-output data is easier to analyze than the method in which impulse response is synthesized, since the sensitivity of intermediate computations required to obtain the impulse response matrix need not be analyzed.

The problem is therefore formulated as follows. Let Σ be a linear system in discrete time modeled by equations (1) and (2):

$$x(s+1) = Fx(s) + Gu(s) \tag{1}$$

$$y(s) = Hx(s). \tag{2}$$

$x(s) \in E^n$ (the "n" dimensional Euclidean space) is the state of Σ at time s; similarly $u(s)$ and $y(s)$ are the m-dimensional input and the p-dimensional output of Σ. F, G, H are real constant matrices of appropriate dimensions. Σ is assumed to be completely reachable and completely observable (for details about these terms see Ref. 4), namely

$$\text{rank of } [G, FG, \cdots, F^{n-1}G] = n \tag{3}$$

and

$$\text{rank of } [H', F'H', \cdots, F'^{n-1}H'] = n \tag{4}$$

where prime (′) denotes the transpose. Given a sequence of inputs $u(s)$ and outputs $y(s)$ for $s = 1, 2, \cdots, N$ (where N is sufficiently large), find a system $\hat{\Sigma}$ of the same dimension as Σ namely n such that $\hat{\Sigma}$ simulates the input-output behavior of Σ.

Remark 1: It is clear that there are some sequences $u(s)$ which will not be sufficient to uniquely specify $\hat{\Sigma}$. Theorems, presented in Section II, give sufficient conditions for $u(s)$ and N which uniquely determine $\hat{\Sigma}$.

Remark 2: When $\hat{\Sigma}$ is uniquely determined it will be shown that the state of $\hat{\Sigma}$ is uniquely related to the state of Σ. In fact the $\hat{F}$, $\hat{G}$, and

$\hat{H}$ of $\hat{\Sigma}$ will be related to the F, G, and H of Σ by a nonsingular transformation such that $HF^iG = \hat{H}\hat{F}^i\hat{G}$ which implies that the impulse responses of Σ and $\hat{\Sigma}$ are identical. Notice that for any nonsingular T

$$\hat{H} = HT^{-1} \qquad \hat{F} = TFT^{-1} \qquad \hat{G} = TG$$

implies that

$$HF^iG = \hat{H}\hat{F}^i\hat{G}.$$

The main difficulty in obtaining a direct algorithm is in getting at the state $x(s)$ from output sequences when the parameters of the system are not known. When, for example, H in the equation below is identity, or equivalently the output itself is the state, it is easy to find an internal description from sequences of inputs and outputs. From writing this equation as

$$x(s+1) = [F \quad G]\begin{bmatrix} x(s) \\ u(s) \end{bmatrix}$$

$$y(s) = Hx(s) = x(s),$$

it follows that given enough observations one can solve for F and G from the above equation for most nontrivial input sequences (see Theorem 2). An easy way is to multiply both sides of this equation by $[x'(s) \; u'(s)]$ and sum from $s = 1$ to $s = N$ where N is the number of observations:

$$\sum_{s=1}^{N} \{x(s+1)[x'(s) \quad u'(s)]\} = [F \quad G] \sum_{s=1}^{N} \left\{\begin{bmatrix} x(s) \\ u(s) \end{bmatrix}[x'(s) \quad u'(s)]\right\}.$$

Whenever the matrix multiplying $[F \; G]$ in the above equation has an inverse, there exists a unique solution for F and G.

In the case when $y(s)$ is not the state itself but only a linear function of the state, the problem is much more complex and one has to select certain appropriate components of the output sequence for an external description in terms of the observables, namely $y(i)$ and $u(i)$. The selection of the right components can be done by introducing an operator to be called the selector matrix as defined below.

In describing the theory of the direct identification method, considerable use is made of the input-output description to be detailed below.

Definition: $\mathcal{S}$ will denote the set of $k \times l$ matrices ($k \leqq l$) with the following properties:

$$(i)\ S = \{s_{ij}\} \text{ where } s_{ij} = 0 \text{ or } 1. \tag{5}$$

$$(ii)\ \forall_i,\ s_{ij} = 1 \text{ for one and only one } j, \text{ say } j_i. \tag{6}$$

$$(iii)\ j_1 < j_2 < \cdots < j_k, \qquad j_i \leqq l, \quad i \leqq k. \tag{7}$$

Examples of matrices belonging to $\mathcal{S}$ are

$$[1]\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and so on.}$$

Any matrix $S \varepsilon \mathcal{S}$ will be referred to as a selector matrix, because S operating on a linear space E^l transforms it into a linear space E^k by mapping every vector $x \varepsilon E^l$ to a vector $y \varepsilon E^k$ by selecting the components $j_1, \cdots, j_k$ of $x \varepsilon E^l$.

The description presented is an "external" description in the sense that the dynamical equations are given in terms of quantities which can be observed from outside, that is, values of input and values of output.

Consider a completely reachable and completely observable discrete time system Σ represented as follows

$$x(s+1) = Fx(s) + Gu(s), \tag{8}$$

$$y(s) = Hx(s), \quad H: p \times n; \quad F: n \times n; \quad G: n \times m. \tag{9}$$

"Completely observable" implies*

$$\rho([F' : H']) = n. \tag{10}$$

$\rho(A)$ = rank of A. Therefore, $\exists$ an $S \varepsilon \mathcal{S}$, such that

$$S\begin{bmatrix} H \\ \vdots \\ HF^{n-1} \end{bmatrix} = T \text{ where } T \text{ is nonsingular;} \tag{11}$$

that is, T^{-1} exists. Without loss of generality it can be assumed from remark 2 that $T = I$ so far as the external description is concerned. Using equations (8) and (9) repeatedly, it follows that

$$\begin{aligned} y(s) &= Hx(s), \\ y(s+1) &= Hx(s+1) = HFx(s) + HGu(s) \\ y(s+n-1) &= HF^{n-1}x(s) + HF^{n-2}Gu(s) + \cdots + HGu(s+n-2). \end{aligned} \tag{12}$$

Let

* $[F' : H'] \triangleq [H', F'H', \cdots, F'^{(n-1)}H']$.

$$\bar{y}'(s) \triangleq [y'(s) \quad y'(s+1) \cdots y'(s+n-1)], \tag{13}$$

$$\bar{u}'(s) \triangleq [u'(s) \quad u'(s+1) \cdots u'(s+n-1)]. \tag{14}$$

Then, writing equation (12) in vector form, and also using equations (13) and (14), it follows that

$$\bar{y}(s) = \begin{bmatrix} H \\ HF \\ \vdots \\ HF^{n-1} \end{bmatrix} x(s) + \begin{bmatrix} 0 & 0 & & 0 \\ HG & 0 & & 0 \\ \vdots & \vdots & & \vdots \\ HF^{n-2}G & HF^{n-1}G & \cdots & 0 \end{bmatrix} \bar{u}(s). \tag{15}$$

Let[†]

$$\begin{bmatrix} 0 & 0 & 0 & & 0 & 0 & 0 \\ HG & 0 & 0 & & 0 & 0 & 0 \\ HFG & HG & 0 & & 0 & 0 & 0 \\ \cdot & HFG & HG & & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot \\ HF^{n-3}G & HF^{n-4}G & HF^{n-5}G & \cdots & HG & 0 & 0 \\ HF^{n-2}G & HF^{n-3}G & HF^{n-4}G & \cdots & HFG & HG & 0 \end{bmatrix} \triangleq R_1 ; \tag{16}$$

then multiplying both sides of equation (15) by S, using the comments given below equation (11), it follows that

$$S\bar{y}(s) = x(s) + SR_1\bar{u}(s). \tag{17}$$

Once again, using equation (9),

$$x(s+1) = Fx(s) + Gu(s),$$

which because of equation (17), with s replaced by $s+1$, reduces to

$$x(s+1) = S\bar{y}(s+1) - SR_1\bar{u}(s+1); \tag{18}$$

substituting equation (9) for $x(s)$ in equation (17) gives

$$S\bar{y}(s+1) = F(S\bar{y}(s) - SR_1\bar{u}(s)) + S \begin{bmatrix} H \\ HF \\ \vdots \\ HF^{n-1} \end{bmatrix} Gu(s) + SR_1\bar{u}(s+1). \tag{19}$$

[†] The last column of zeroes in R_1 is added so that $\bar{y}$ and $\bar{u}$ may be consistently defined.

Since it has been shown that [see equation (11)]

$$S\begin{bmatrix} H \\ HF \\ \vdots \\ HF^{n-1} \end{bmatrix} = I, \tag{20}$$

it follows that

$$S\bar{y}(s+1) = FS\bar{y}(s) + R\bar{u}(s), \tag{21}$$

where‡

$$R \triangleq -FSR_1 + S\begin{bmatrix} HG & 0 & 0 & 0 & 0 \\ HFG & HG & 0 & 0 & 0 \\ HF^2G & HFG & 0 & 0 & 0 \\ \cdot & \cdot & 0 & \cdot & \cdot \\ HF^{n-1}G & HF^{n-2}G & \cdot & \cdot & HG \end{bmatrix}. \tag{22}$$

Equation (21) gives a relation between the input sequence $u(i)$ and the output sequence $y(i)$ which does not involve the state. It is an external description in the sense that the variables in equation (21), namely $u(i)$ and $y(i)$, can be measured externally. From equation (22) it follows that if R is partitioned as

$$[R_0 \quad R_1 \cdots R_{n-1}], \quad R_i : \quad n \times m \quad \forall i, \tag{23}$$

then

$$R_{n-1} = S\begin{bmatrix} 0 \\ \vdots \\ HG \end{bmatrix}. \tag{24}$$

It is obvious how one obtains the columns of the second product in equation (22). To obtain the contribution from $-FSR_1$, notice from equation (16) that S times the second column from the end of R_1 is, from equation (24), merely R_{n-1}. Therefore, the second column of FSR_1 from the end is simply FR_{n-1} and therefore

‡ In adding $SR_1\bar{u}(s + 1)$ to the second term in equation (20), the last column of R_1 may be dropped because it is all zeroes.

$$R_{n-2} = -FR_{n-1} + S\begin{bmatrix} 0 \\ \vdots \\ HG \\ HFG \end{bmatrix} \tag{25}$$

Now notice that $R_{n-2} + FR_{n-1}$ is S times the third column from the end of R_1. Therefore the third column from the end of R is

$$R_{n-3} = -FR_{n-2} - F^2R_{n-1} + S\begin{bmatrix} 0 \\ \vdots \\ HG \\ HFG \\ HF^2G \end{bmatrix}.$$

Continuing in the same way,

$$R_{n-4} = -FR_{n-3} - F^2R_{n-2} - F^3R_{n-1} + S\begin{bmatrix} 0 \\ \cdot \\ 0 \\ HG \\ HFG \\ HF^2G \\ HF^3G \end{bmatrix}$$

and finally

$$S\begin{bmatrix} HG \\ HFG \\ \vdots \\ HF^{n-1}G \end{bmatrix} = R_0 + FR_1 + \cdots + F^{n-1}R_{n-1} . \tag{26}$$

Now, since it was possible to choose a basis such that

$$S\begin{bmatrix} HG \\ \vdots \\ HF^{n-1}G \end{bmatrix} = IG,$$

one has

$$G = R_0 + FR_1 + \cdots F^{n-1}R_{n-1} . \tag{27}$$

Equation (21) may be written in the form

$$S\bar{y}(s + 1) = [F \quad R]\begin{bmatrix} S\bar{y}(s) \\ \bar{u}(s) \end{bmatrix}$$

and may in principle be solved for F and R. Thus from equations (27) it is clear that if the values for $u(i)$ and $y(i)$ were given and S were known, one could also solve for one set of values for F; and since in most cases H is full rank, H can be assumed to be $[I\ 0]$.

II. THE MINIMAL REPRESENTATION AND THE DIRECT ALGORITHM.

It was shown in Section I that, corresponding to every internal description of Σ which is completely controllable and completely observable, there is a description in the form of equation (21). In this section we show that from the knowledge of the values of $u(i)$, $i = 1, \cdots, N$, and $y(i)$, $i = 1, \cdots, N$, one can get the internal description of Σ under very general conditions on $u(i)$. Central to the discussion are a few results which are presented in the form of theorems for the sake of clarity and precision.

Given $u(i)$, $i = 1, \cdots, N$, the inputs to a system Σ of dimension n which is completely observable and completely reachable, and the corresponding outputs $y(i)$, $i = 1, \cdots, N$, the following propositions hold true:

Note 1: It will be assumed in the following that the column dimension k of the selector matrix is always a multiple of p; further if $k = rp$, then

$$(r - 1)p \leqq j_l \leqq rp.$$

It is obvious that there is no loss of generality involved in this assumption. (l is the row dimension of S.)

Note 2: In the definition of $\bar{y}(s)$ and $\bar{u}(s)$ in equations (13) and (14), the n should be replaced by r defined in Note 1 above.

Theorem 1: Let S *be* l × k (= rp); *then*

$$\rho\left\{ S \begin{bmatrix} H \\ HF \\ \vdots \\ HF^{r-1} \end{bmatrix} \right\} < 1 \tag{28}$$

implies that

$$\sum_{s=1}^{N-r+1} \begin{bmatrix} \mathrm{S}\bar{\mathrm{y}}(\mathrm{s}) \\ \bar{\mathrm{u}}(\mathrm{s}) \end{bmatrix} [\bar{\mathrm{y}}'(\mathrm{s})\mathrm{S}' \quad \bar{\mathrm{u}}'(\mathrm{s})] \text{ is a singular matrix} \tag{29}$$

for every sequence u(i), i = *1, 2,* $\cdots$, N.

Proof: Multiplying equation (10) on the left by S and replacing n by r we have,

$$S\bar{y}(s) = S \begin{bmatrix} H \\ HF \\ \vdots \\ HF^{r-1} \end{bmatrix} x(s) + SR_1\bar{u}(s). \tag{30}$$

Because of equation (28) $\exists$ a vector, $z \neq 0$, and in E^l such that

$$z'S \begin{bmatrix} H \\ HF \\ \vdots \\ HF^{r-1} \end{bmatrix} = 0. \tag{31}$$

Therefore, multiplying equation (30) on the left by z' gives

$$z'S\bar{y}(s) = z'SR_1\bar{u}(s). \tag{32}$$

Therefore,

$$[z' \quad -z'SR_1] \begin{bmatrix} S\bar{y}(s) \\ \bar{u}(s) \end{bmatrix} = 0 \qquad \forall s \leqq N - r + 1 \tag{33}$$

which implies that

$$\sum_{s=1}^{N-r+1} \begin{bmatrix} \bar{S}\bar{y}(s) \\ \bar{u}(s) \end{bmatrix} [\bar{y}'(s)S' \quad \bar{u}'(s)] \text{ is singular.} \qquad \text{QED}$$

Theorem 2: If Σ *is completely observable and completely reachable, the matrices* F, G, *and* H *are* n × n, n × m, *and* p × n *respectively; then* $\exists$ *an* S: n × np *such that*

$$\mathrm{T} \triangleq \sum_{s=1}^{n+nm} \begin{bmatrix} \mathrm{S}\bar{\mathrm{y}}(\mathrm{s}) \\ \bar{\mathrm{u}}(\mathrm{s}) \end{bmatrix} [\bar{\mathrm{y}}'(\mathrm{s})\mathrm{S}' \quad \bar{\mathrm{u}}(\mathrm{s})] > 0 \text{ almost surely} \tag{34}$$

where u(i) *are random variables having a joint nonlattice distribution.**

Proof: The first step in the proof consists of establishing Lemma 1.

Lemma 1: T > 0 *if and only if*

$$\sum_{s=1}^{n+nm} \begin{bmatrix} x(s) \\ \bar{u}(s) \end{bmatrix} [x'(s) \quad \bar{u}(s)] > 0. \tag{35}$$

Proof of Lemma 1: If $T \ngtr 0$, $\exists$ z' such that $z' = [z_1' \quad z_2'] \neq 0$, and

$$z_1' S \bar{y}(s) + z_2' \bar{u}(s) = 0 \quad \forall s. \tag{36}$$

Since Σ is completely observable, multiplying equation (17)

$$S\bar{y}(s) = x(s) + SR_1 \bar{u}(s) \tag{37}$$

on the left by z_1' one obtains

$$z_1' S \bar{y}(s) = z_1' x(s) + z_1' S R_1 \bar{u}(s). \tag{38}$$

Combining equations (36) and (38),

$$z_1' x(s) + (z_1' S R_1 - z_2') \bar{u}(s) = 0 \qquad \forall s, \tag{39}$$

and

$$[z_1' \, , z_1' S R_1 - z_2'] \neq 0;$$

for if $[z_1' \, , z_1' S R_1 - z_2'] = 0$, then $[z_1' \quad z_2'] = 0$, which contradicts $z \neq 0$. Therefore,

$$\sum_{s=1}^{n+nm} \begin{bmatrix} x(s) \\ \bar{u}(s) \end{bmatrix} [x'(s) \quad \bar{u}'(s)] \ngtr 0. \tag{40}$$

Now suppose $T > 0$. Let

$$\sum_{s=1}^{n+nm} \begin{bmatrix} x(s) \\ u(s) \end{bmatrix} [x'(s) \quad \bar{u}(s)] \ngtr 0.$$

Then $\exists$ a $z' = [z_1' \quad z_2'] \neq 0$ such that

$$z_1' x(s) + z_2' \bar{u}(s) = 0 \quad \forall s. \tag{41}$$

Again multiplying equation (37) by z_1' and using equation (41), it follows that

$$z_1' S \bar{y}(s) = -z_2' \bar{u}(s) + z_1' S R_1 \bar{u}(s) \tag{42}$$

* A nonlattice distribution is one in which no nonzero probability mass is concentrated on a surface less than the dimension of the random variable.

$$z_1' S\bar{y}(s) + (z_2' - z_1' SR_1) \quad \bar{u}(s) = 0 \qquad \forall \mathrm{s}. \tag{43}$$

Once again $[z_1', (z_2' - z_1' SR_1)] \neq 0$ since $z \neq 0$, which contradicts $T > 0$. The proof of Lemma 2 will now complete the proof of Theorem 2.

Lemma 2: If (i) Σ *is completely controllable,* (ii) u(i) *are random variables with a joint nonlattice distribution, then the* (n + nm) × (n + nm) *matrix*

$$\begin{bmatrix} \mathrm{x}(1) & \cdots & \mathrm{x}(\mathrm{n} + \mathrm{nm}) \\ \bar{\mathrm{u}}(1) & \cdots & \bar{\mathrm{u}}(\mathrm{n} + \mathrm{nm}) \end{bmatrix} \tag{44}$$

is almost surely nonsingular.

Proof: From Lemma A.2 in Appendix A of Ref. 5 it follows that if

$$\begin{gathered} z(s+1) = F_1 z(s) + G_1 u(s), \\ \text{with} \quad F_1 (n + nm) \times (n + nm), \end{gathered} \tag{45}$$

then $[z(1), \cdots, z(n + nm)]$ is nonsingular with probability one, if F_1, G_1 is completely controllable. Further, from equations (8) and the definition of u, it is clear that

$$\begin{bmatrix} x(s+1) \\ u(s+1) \\ \cdot \\ \cdot \\ u(s+n) \end{bmatrix} = \begin{bmatrix} F & G & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & I & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & \cdot & I \\ 0 & 0 & \cdot & \cdot & \cdot & \cdot & 0 \end{bmatrix} \begin{bmatrix} x(s) \\ u(s) \\ \cdot \\ \cdot \\ u(s+n-1) \end{bmatrix} + \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ I \end{bmatrix} . \; u\,(s+n) \tag{46}$$

Therefore, identifying F_1 and G_1 as

$$\begin{bmatrix} F & G & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & I & \cdot & \cdot & \cdot & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdot & \cdot & \cdot & \cdot & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 \\ \cdot \\ \vdots \\ I \end{bmatrix}$$

respectively, equation (44) follows since it can easily be shown that $[F \quad G]$ controllable implies that $[F_1 \quad G_1]$ is completely controllable. Lemma 2 implies that the matrix in equation (40) is positive definite since in general A nonsingular → $A^T A > 0$, which implies equation (34) by (i).

III. THE COMPUTATIONAL METHOD

The main part of the algorithm, as would be expected from the discussion in Section II, is to determine the right selector matrix. Once this has been done it is easy to solve for the parameters. In order to utilize certain properties of the matrix $[F' : H]$, a class of matrices $\bar{\mathcal{S}} \subset \mathcal{S}$ is detailed below since $\bar{\mathcal{S}} \subset \mathcal{S}$, the number of different selector matrices one has to try, is smaller than $\mathcal{S}$.

Definition: $\bar{\mathcal{S}}$ is the set of matrices $S \varepsilon \mathcal{S}$ such that S is $l \times k$; then

(*i*) k is an integral multiple of p and further, if $k = rp$, then

$$(r - 1)p < j_l \leqq rp$$

where j_l is as defined in equation (5).

(*ii*)
$$j_i - j_{i-1} \leqq 2p.$$

Observe that by (i) there always exists an $S \varepsilon \bar{\mathcal{S}}$ such that equation (11) holds, since as can easily be proved, if

$$\rho([H', F'H', \cdots, F'^{s}H']) = \rho([H', F'H', \cdots, F'^{(s+1)}H']) = q$$

then

$$\rho([H', F' H', F'^{(s+j)}H']) = q \qquad j = 0, 1, 2, \cdots ;$$

so that in spite of condition (ii) in the above definition, there exist an $S \varepsilon \bar{\mathcal{S}}$ such that equation (11) holds.

(*iii*) The formulas (21) and (27) are still valid for any $S \varepsilon \bar{\mathcal{S}}$ satisfying equation (11), with n replaced everywhere by r defined in condition (i) in the definition of $\bar{\mathcal{S}}$ above.

Now from Theorems 1 and 2 and the above discussion, the direct algorithm can be summarized as follows.

It can be assumed without loss of generality that: (*i*) $N \geqq n + (m + 1)n$; that is, there is a sufficient number of observations to determine the internal description uniquely. n is the minimal dimension of the system to be identified. (*ii*) H has full rank. (*iii*) $n \geqq p$.

Step 1: Since $n \leqq \bar{N}/(m + 2)$, let $\bar{N}$ be the largest integer $\leqq N/(m + 2)$. Then $n \leqq N$. In order to arrive at the right S, one starts with an $S \varepsilon \mathcal{S}$ of row dimension $\bar{N}$ and tests the nonsingularity of

$$T \triangleq \sum_{s=1}^{(m+1)\bar{N}} \begin{bmatrix} S\bar{y}(s) \\ \bar{u}(s) \end{bmatrix} [\bar{y}'(s)S' \quad \bar{u}'(s)]$$

for all $S \varepsilon \mathcal{S}$ and having row dimension $\bar{N}$. If T is nonsingular, $\bar{N} = n$. If T is singular, then reduce the row dimension of S by 1 and repeat the test. Repeat the procedure until T becomes nonsingular. The row dimension of S will then be n; let r be as defined in condition (i) in the definition of $\bar{\mathrm{S}}$; that is, S is $n \times rp$.

Step 2: Solve for F, R as follows.

$$[F \quad R] = \left\{ \sum_{s=1}^{(m+1)\bar{N}} Sy(s+1)[\bar{y}(s)S' \quad \bar{u}'(s)] \right\} T^{-1}$$

Step 3: Solve for G from the following formula.

$$G = R_0 + FR_1 + \cdots + F^{r-1}R_{r-1} ,$$

where $S{:}n \times rp$ and R_i are the partitions of R such that

$$R = [R_0 \quad R_1 \cdots R_{r-1}]$$

and $R_i = n \times m\; i = 0, \cdots, r-1$. H can be assumed to be $[I \;0]$ where the identity has dimension p.

In the case when Σ is a continuous-time system, the algorithm presented above applies with appropriate modifications. In the definitions of $\bar{y}(s)$ and $\bar{u}(s)$, s now assumes values in $\mathcal{R}$ and $y(s+i)$ should be replaced by $y^{(i)}(s)$ evaluated at s. The summation signs should be replaced by integration over an interval. The formulas for the parameters become

$$[F \quad R] = \int_t^{t+\epsilon} S\bar{y}^{(r)}(s)[\bar{y}'(s)S' \quad \bar{u}'(s)]\, ds\, T^{-1}$$

$$T = \int_t^{t+\epsilon} \begin{bmatrix} \bar{S}y(s) \\ \bar{u}(s) \end{bmatrix} [\bar{y}'(s)S' \quad \bar{u}'(s)]\, ds.$$

G can be obtained from R exactly as in the above algorithm for the discrete time case.

In the case when observations are contaminated with noise, this method can be generalized to yield consistent estimates for the parameters (see Ref. 5).

REFERENCES

1. Ho, B. L. "On Effective Construction of Realizations from Input-Output Descriptions," Ph.D. dissertation, 1965, Department of Engineering Mechanics, Stanford University, California.
2. Ho, B. L., and Kalman, R. E., "Effective Construction of Linear State-Variable Models from Input/Output Data," Proceedings of the Third Annual Allerton, Conf. on Circuit and Syst. Theory, 1965, pp. 449–459.
3. Youla, D. C., "The Synthesis of Linear Dynamical Systems from Prescribed Weighting Patterns," J. SIAM, *14*, No. 4 (May 1966), pp. 527–549.
4. Kalman, R. E., "Mathematical Description of Linear Systems," SIAM J. on Control, *1*, No. 1 (1963), pp. 152–192.
5. Gopinath, B, "On the Identification and Control of Linear Systems," Ph.D. dissertation, 1968, Department of Electrical Engineering, Stanford University, California.

Uniform Synthesis of Sequential Circuits†

By J. D. ULLMAN and PETER WEINER‡

(Manuscript received July 19, 1968)

In this paper we consider the synthesis of sequential machines by networks of a fixed module with delay. We show that every binary input n *state sequential machine has an isomorphic realization using at most* p *copies of a module with* 2r + 1 *inputs, where* p *is the smaller of* $\frac{2r}{2r-1}(n^{1+\log_r 2} + 4n^{1+\log_r 4})$ *and* $r2^{[n/r]}$. ([x] *is the smallest integer* ≧ x.)

I. INTRODUCTION

The realization of an arbitrary binary output synchronous sequential machine by a network of copies of a fixed sequential machine (module) or copies of a small number of machines is a problem which has received recent attention.[1–5] An equivalent problem has been studied in Ref. 6. A design of this sort is particularly suited to batch fabrication techniques, because it is possible to mass produce a fairly complex integrated circuit (the module) and then wire these circuits together to realize any desired sequential machine.

The machines, so constructed, will be fast; the time between inputs need not be longer than the time it takes a single module to resolve its output after a change in input, no matter how many modules are in the network. The disadvantage of this technique, so far, has been the large number of copies of the module necessary to realize a machine; as many as $2^n - 2$ copies for an n state machine are required when using the modules of Refs. 1 and 2. These modules are shown in Fig. 1 for the binary input case.

Not shown in any of our diagrams is provision for initializing the output of any module to the hot (1) state if desired. Neither is provision for control of the module by a clock shown in this or any other module.

† Portions of this paper appeared in the Proceedings of the IEEE 9th Annual Symposium on Switching and Automata Theory, Schenectady, N. Y., October 1968.

‡ Princeton University, Princeton, New Jersey.

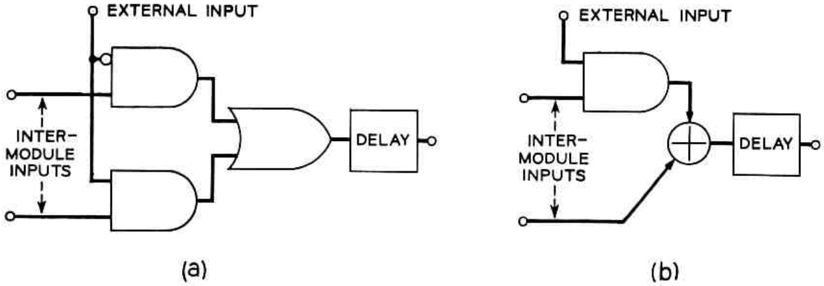


Fig. 1 — Simple modules.

The modules of Fig. 1 each have two intermodule inputs, that is, inputs to which either logical constants or the output of some module will be connected. If a module with two intermodule inputs is universal (can realize any sequential machine having one binary input), then there is a unique minimal network composed of copies of this module realizing a particular sequential machine with a binary input.[1,2] If there are more than two intermodule inputs, there may be more than one network realizing a given machine. We consider a class of modules with different numbers of intermodule inputs and attempt to design small networks consisting of copies of one of the modules in the class.

The class of modules we use for single input machines is represented schematically in Fig. 2a. There is a member of the class with $2r$ intermodule leads for each $r \geqq 1$. Let the module of Fig. 2a with a particular value of r be M_r. Note that M_1 is essentially the same as the module of Fig. 1a. M_2 is shown in Fig. 2b.

In what follows, we restrict ourselves to the design of networks for the realization of machines with one binary input. The generalization to the use of machines having k binary inputs is straightforward when one uses a class of modules represented schematically in Fig. 3.

Notice that conventional designs of sequential circuits, represented schematically in Fig. 4, require the construction of $\log_2 n$ Boolean functions of $k + \log_2 n$ variables, where k and n are the number of input variables and states, respectively, of the machine. The number of gates necessary for a two-level realization of several functions of p variables can be as high as 2^p, so one would expect, even in the case $k = 1$, to require as many as n gates for a realization in the form of Fig. 4. We cannot show, for fixed r, that all n state machines with

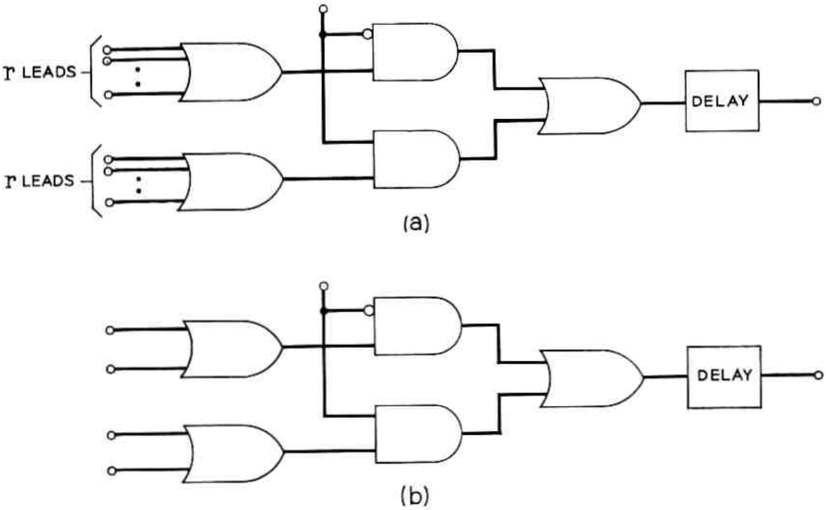


Fig. 2 — (a) The module M_r; (b) the **module M_2.**

single inputs can be realized by networks of as few as n copies of M_r. However, we show that the number of copies of M_r needed to realize any binary input n state sequential machine is bounded above by two functions of n. These functions, to within a constant factor, are $2^{n/r}$ and $n^{1+\log_r 4}$.

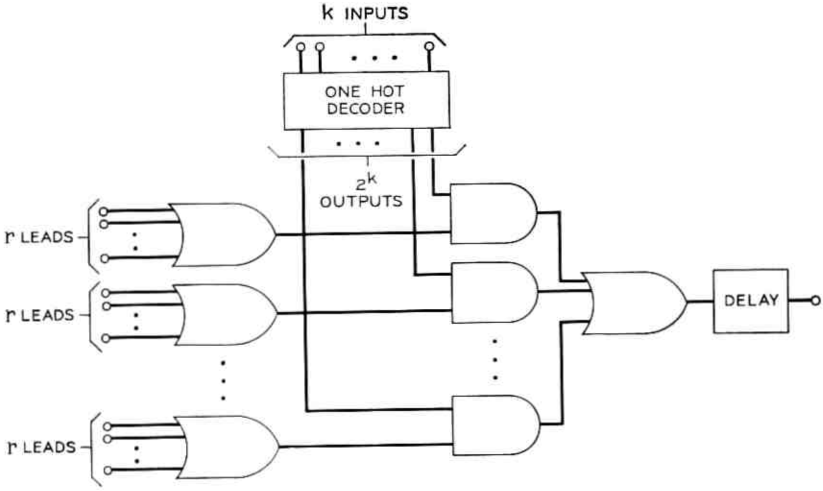


Fig. 3 — Generalization of M_r.

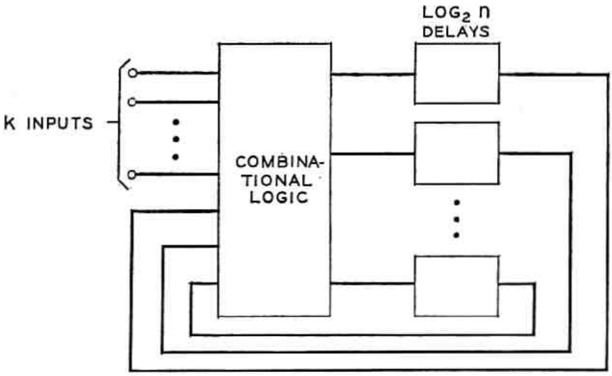


Fig. 4 — Conventional sequential circuit.

II. DEFINITIONS AND BASIC CONCEPTS

A sequential machine will be denoted $A = (K, \Sigma, \delta, q_0, F)$. K and Σ are finite sets of *states* and *inputs,* respectively. F, the *final states,* is a subset of K. It is the set of states for which the output is 1. q_0, the *start state* is a particular element of K. δ maps $K \times \Sigma$ to K. It gives the next state for each combination of state and input symbol. The function δ is usually displayed as a flow table, with a row for each state and a column for each input. The entry in the ith row and jth column is the value of δ for the ith state and jth input. The first state will always be the start state. An example is shown in Table I.

We extend δ to domain $K \times \Sigma^*$ by:†

(*i*) $\delta(q, \epsilon) = q$ for all q in K.
(*ii*) $\delta(q, wa) = \delta\ (\delta\ (q, w), a)$, for all q in K, w in Σ^*, and a in Σ.

The *event defined by* the machine A, denoted $T(A)$, is $\{w \mid \delta\ (q_0, w)$ is in $F\}$. That is, $T(A)$ consists of exactly those input strings which cause A to go from the start state to a final state. For example, if 3 and 5 are the final states of the machine of Table I, then 110 is in $T(A)$, since $\delta(1, 1) = 6$, $\delta(6, 1) = 4$ and $\delta(4, 0) = 5$. 001 is not in $T(A)$ since $\delta(1, 0) = 3$, $\delta(3, 0) = 1$ and $\delta(1, 1) = 6$.

Let R be a subset of Σ^* for some finite set Σ. For each w in Σ^*, define the *derivative* of R with respect to w, denoted R/w to be set of strings x such that xw is in R.‡

† Σ^* is the set of all strings of symbols in Σ, including ϵ, the string of length 0.
‡ This notion of derivative is "backwards" from that used in Ref. 7. It is actually the quotient operation of Ref. 8.

Let $A = (K, \Sigma, \delta, q_0, F)$ be a machine, and let $\Sigma = \{0,1\}$. We can define two "inverses" of δ, denoted μ_0 and μ_1. These functions map sets of states to sets of states by:

$$\mu_0(G) = \{q \mid \delta(q, 0) \text{ is in } G\},$$

$$\mu_1(G) = \{q \mid \delta(q, 1) \text{ is in } G\}.$$

For each subset G of K, let A_G be the machine $(K, \{0, 1\}, \delta, q_0, G)$. Let $H = \mu_0(G)$ and $J = \mu_1(G)$. If $R = T(A_G)$, then $R/0 = T(A_H)$ and $R/1 = T(A_J)$. For w is in $R/0$ if and only if $w0$ is in $T(A_G)$. But $w0$ is in $T(A_G)$ if and only if $\delta(q_0, w)$ is a state p such that $\delta(p, 0)$ is in G. Equivalently, w is in $R/0$ if and only if $\delta(q_0, w)$ is in H. The argument for $R/1$ is analogous.

When talking about a fixed sequential machine, $A = (K, \{0, 1\}, \delta, q_0, F)$, we often identify $T(A_G)$ with G for each subset G of K. We use $G/0$ and $G/1$ for $\mu_0(G)$ and $\mu_1(G)$. For example, if A is the machine of Table I and $G = \{1, 3, 5\}$, then $G/0 = \{1, 2, 3, 4, 5\}$ and $G/1 = \{3\}$.

A *network* of a module M is an interconnection of copies of M such that each intermodule input is connected to either the output of a copy of M in the network or a logical constant (0 or 1). The external inputs of each copy of M (or corresponding external inputs if a copy has more than one) are connected together and receive the input to the network. One copy of M is designated the output of the network; the network accepts an input sequence if the output of the designated copy is hot (1) after receiving the sequence.

The module M_2 of Fig. 2b is repeated as Fig. 5 with certain points marked. Suppose that this module is part of a network realizing the event F of the sequential machine $A = (K, \{0, 1\}, \delta, q_0, F)$. Suppose also, that it has been determined that the output of this copy of the module must be some event $G \subseteq K$. That is, the output of this module

TABLE I—NEXT STATE FUNCTION

		inputs	
		0	1
states	1	3	6
	2	5	4
	3	1	5
	4	5	6
	5	5	2
	6	2	4

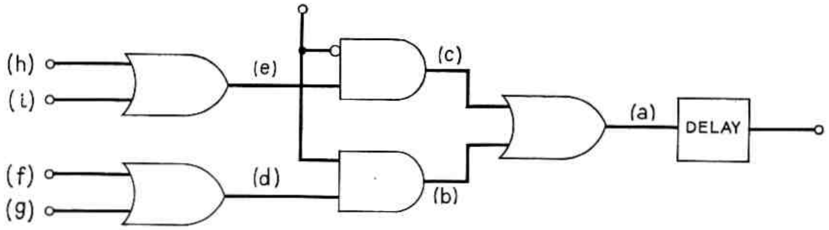


Fig. 5 — Points in the module M_2.

is hot exactly when the sequence of inputs to the network is in the event G. Thus, when the last input appears at the external input of this module, point a, the input to the delay, must immediately become hot if and only if the last input completes a sequence in G.

Observe that point b can be hot only if the last input is 1 and point c can be hot only if the last input is 0. Thus, immediately before the last input appears at the external input terminal, point d must be hot if and only if the previous inputs form a sequence in $G/1$ and point e must be hot if anly only if the previous input sequence is in $G/0$.

The union of the events at f and g must thus be $G/1$ and the union of events at h and i must be $G/0$. We are free to choose the events at the intermodule inputs subject only to these constraints. For example, we could choose the events at f and g to be those strings in $G/1$ of even and odd length, respectively. However, we restrict our choice so that the events at the intermodule leads will be representable as sets of states of A.

Design, using the module M_r, $r > 2$, proceeds the same way. If a given copy of the module is to realize the event G, then the lowest r of the intermodule inputs must be from modules realizing events H_1, H_2, $\cdots$, H_r whose union is $G/1$; the remaining r intermodule inputs must be from modules realizing events J_1, J_2, $\cdots$, J_r, whose union is $G/0$. However, some of $H_1, \cdots, H_r$ or $J_1, \cdots, J_r$ may be the empty set or the set of all states, in which case these events are "realized" by logical constants rather than modules.

The above arguments justify the following reduction in the design problem for the class of modules M_r, $r \geqq 1$:

Let $A = (K, \{0, 1\}, \delta, q_0, F)$ be a sequential machine. An M_r-*synthesis* of A is a set $\mathcal{S}$ of subsets of K having the properties:

(*i*) F is in $\mathcal{S}$.

(*ii*) If G is in $\mathcal{S}$, then there are sets H_1, H_2, $\cdots$, H_r and

$J_1, J_2, \cdots, J_r$ in $\mathcal{S}$, not necessarily all distinct, such that

$$\bigcup_{i=1}^{r} H_i = G/1 \quad \text{and} \quad \bigcup_{i=1}^{r} J_r = G/0.$$

From what we have said concerning the flow of signals in the module M_r, we may conclude that if $\mathcal{S}$ is an M_r-synthesis of A, then there is a network of m copies of M_r realizing $T(A)$, where m is the number of elements of $\mathcal{S}$ that are neither Φ nor K.[†] We call m the *size* of $\mathcal{S}$.

Notice that an M_r-synthesis requires that all modules realize events which are identifiable with a set of states. Such networks are called *isomorphic* to A. There may be networks of copies of M_r which realize $T(A)$, but are not M_r-syntheses of A. However, in our search for small networks we shall not consider any networks except those which are M_r-syntheses. See Ref. 5 for some comments on the existence of non-isomorphic realizations of sequential machines.

III. CONSTRUCTION OF M_r-SYNTHESES

The purpose of this paper is to show that M_r-syntheses of small size exist for an arbitrary n-state sequential machine. The first bound on the size of an M_r-synthesis is straightforward.

Let $A = (K, \{0, 1\}, \delta, q_0, F)$ be an n state sequential machine. We may choose r disjoint subsets of K, say $K_1, K_2, \cdots, K_r$, such that $\bigcup_{i=1}^{r} K_i = K$ and no K_i, $1 \leqq i \leqq r$, contains more than $[n/r]$ states.[‡] Let $\mathcal{S} = \{F\} \cup \{G \mid G \subseteq K_i \text{ for some } i\}$. To see that $\mathcal{S}$ is an M_r-synthesis of A, we have merely to observe that any subset G of K can be expressed as $\bigcup_{i=1}^{r} G_i$, where $G_i = G \cap K_i \subseteq K_i$ for all i. Thus, for any H in $\mathcal{S}$, $H/0$ and $H/1$ are both the union of r elements of $\mathcal{S}$.

The size of $\mathcal{S}$ is no greater than $1 + r\ (2^{[n/r]} - 1)$, which is almost $r2^{[n/r]}$. We thus have:

Theorem 1: If A *is an* n *state sequential machine, with a single binary input, then there is an* $\mathrm{M_r}$*-synthesis of* A *using at most* $\mathrm{r}2^{[\mathrm{n/r}]}$ *copies of* $\mathrm{M_r}$.

Notice that Theorem 1 is not dependent upon the assumption that A has a single binary input. The machine A in that theorem can have any number of binary inputs. Of course, the appropriate generalization of the input module M_r, as given in Figure 3, must be used.

[†] Φ denotes the empty set.

[‡] We use $[x]$ for "the smallest integer equal to or greater than x."

Example: We use a technique suggested by Theorem 1 to design a network for the sequential machine of Table I with final states $\{4, 5, 6\}$. We generate subsets in a sequential manner, and terminate when no new sets are required. Let $r = 2$ and let the states be divided into two sets K_1 $\{1, 3, 5\}$ and $K_2 = \{2, 4, 6\}$. Now $\{4, 5, 6\}/0 = \{2, 4, 5\}$ and $\{4, 5, 6\}/1 = \{1, 2, 3, 4, 6\}$. If we intersect $\{2, 4, 5\}$ and $\{1, 2, 3, 4, 6\}$ each with K_1 and K_2, the inputs to the module realizing $\{4, 5, 6\}$ must be connected to modules realizing $\{5\}$, $\{2, 4\}$, $\{2, 4, 6\}$ and $\{1, 3\}$. The two derivatives of each of these sets are found among $\{2, 4, 5\}$, $\{6\}$, $\{1, 3\}$, $\{3\}$, $\{2, 5, 6\}$, $\{1, 2, 4, 5, 6\}$ and Φ. Intersecting each of these sets with K_1 and K_2, we find that the network needs modules realizing the events $\{6\}$, $\{3\}$, $\{2, 6\}$ and $\{1, 5\}$, in addition to the modules already used. Proceeding in this way, we find that the entire network also requires modules realizing $\{1\}$, $\{4\}$ and $\{3, 5\}$. The completed network is shown in Fig. 6. Modules are labeled with the event (set of states) they realize. Those modules realizing a set including state 1, the start state, must initially give a 1 output. Inputs to the module are shown in no particular order, and inputs not shown are connected to 0.

The second bound uses the concept of partitions on the set of states of a finite automaton.[9] A *partition* on a set of states K is a set of disjoint, nonempty sets, called *blocks* whose union is K. If $A = (K, \{0, 1\}, \delta, q_0, F)$ is a sequential machine, we can associate with every string w in $\{0, 1\}^*$ a partition Π_w as follows:

(*i*) $\Pi_\epsilon = (\{q_1\}, \{q_2\}, \cdots, \{q_m\})$, where $K = \{q_1, q_2, \cdots, q_m\}$.†

(*ii*) For any w in $\{0, 1\}^*$, let Π_w be $(K_1, K_2, \cdots, K_r)$. Let Π_{w0} be the list of nonempty sets G such that $G = K_i/0$ for some i and Π_{w1} be the list of nonempty sets H such that $H = K_i/1$ for some i.

Example: Consider the machine of Table I. $\Pi_\epsilon = 1, 2, 3, 4, 5, 6)$ Π_0 is the list of sets of states that map to a single state under a 0 input. Thus, $\Pi_0 = (1, 245, 3, 6)$. Similarly, $\Pi_1 = (14, 26, 3, 5)$. Proceeding, we can calculate Π_{00} and Π_{01} from Π_0 by seeing which sets of states map onto a single block of Π_0 under inputs 0 and 1, respectively. For example, states 2, 4, 5 and 6 are those which map under a 0 input to one of the states 2, 4 or 5. We find $\Pi_{00} = (1, 2456, 3)$ and $\Pi_{01} = (14, 2356)$. Also, $\Pi_{10} = (1, 245, 3, 6)$ and $\Pi_{11} = (145, 26, 3)$.

A partition Π is said to *represent* a family of sets, namely those sets

† We denote partitions by lists of the blocks. Sometimes it is simpler to represent each block by a string of states not surrounded by brackets. Thus $(\{q_1, q_2\}, \{q_3\})$ will appear as (q_1q_2, q_3).

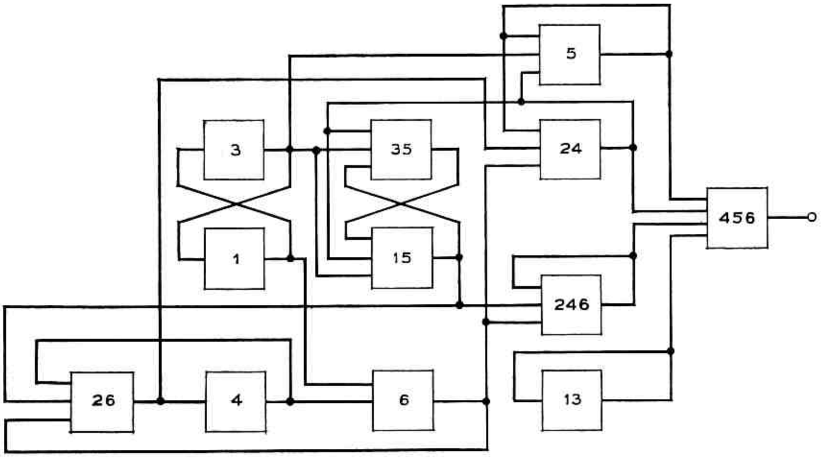


Fig. 6 — Network suggested by Theorem 1.

which are the union of some of the blocks of Π. For example, Π_{01} above represents the sets Φ, {1, 4}, {2, 3, 5, 6} and {1, 2, 3, 4, 5, 6}. Suppose $\Pi_w = (K_1, K_2, \cdots, K_m)$ and G is the union of j of $K_1, K_2, \cdots, K_m$, say $G = K_{i_1} \cup K_{i_2} \cup \cdots \cup K_{i_j}$. Then for $a = 0$ or 1, $G/a = K_{i_1}/a \cup K_{i_2}/a \cup \cdots \cup K_{i_1}/a$ is represented by Π_{wa} and is, in fact, the union of, at most, j blocks of Π_{wa}. Armed with this observation, we prove:

Theorem 2: For every n *state sequential machine with single binary input* A = (K, {0, 1}, δ, q_0, F) *and* r ≧ 2, *there is an* M_r*-synthesis of* A *of size at most* $\frac{2r}{2r-1}(n^{1+\log_r 2} + 4n^{1+\log_r 4})$.

Proof: Let $j = [\log_r n]$. Define the blocks of those partitions Π_w, such that $|w| \leqq j$† to be *basic events*. We will choose $\mathfrak{J}$ to be a set of pairs (G, w), where $G \subseteq K$, w is in $\{0, 1\}^*$, and for each (G, w) in $\mathfrak{J}$, G is represented by Π_w. After constructing $\mathfrak{J}$, we construct $\mathfrak{S}$, an M_r-synthesis of A, from $\mathfrak{J}$ by $\mathfrak{S} = \{\Phi\} \cup \{G \mid (G, w) \text{ is in } \mathfrak{J} \text{ for some } w\}$. We construct $\mathfrak{J}$ by:

(*i*) (F, ε is in $\mathfrak{J}$).

(*ii*) If G is a basic event, then (G, ϵ) is in $\mathfrak{J}$.

(*iii*) Let (G, w) be in $\mathfrak{J}$, $|w| < j$, and let G be the union of k blocks of Π_w. We may choose $H_1, H_2, \cdots, H_r$ such that their union is $G/0$

† $|w|$ denotes the length of w.

and for each i, H_i is the union of from zero to $[k/r]$ blocks of Π_{w0}. Also, choose $J_1, J_2, \cdots, J_r$ such that their union is $G/1$ and for each i, J_i is the union of from zero to $[k/r]$ blocks of Π_{w1}. If H_i is not Φ or a basic event, add $(H_i, w0)$ to $\mathfrak{J}$. If J_i is not Φ or a basic event, add $(J_i, w1)$ to $\mathfrak{J}$.

We say that each $(H_i, w0)$ or $(J_i, w1)$ in $\mathfrak{J}$ is in the *family* of (G, w). We extend the notion of a family by saying that (G, w) is in its own family and if (H, x) is in the family of (G, w) and (J, y) is in the family of (H, x), then (J, y) is in the family of (G, w). The family of (G, ϵ), where G is F or a basic event, can be thought of as the set of elements that must be in $\mathfrak{J}$ because (G, ϵ) is in $\mathfrak{J}$.

We must show that $\mathcal{S}$ is an M_r-synthesis of A. If (G, w) is in $\mathfrak{J}$, then G consists of at most $r^{j-|w|}$ blocks of Π_w. (Since $r^j \geqq n$, we have $r^{j-1} = [r^j/r] \geqq [n/r]$; $r^{j-2} = [r^{j-1}/r] \geqq [[n/r]/r]$ and so on.) We may conclude that if $|w| = j$, then G would be a basic event, and hence, for no G and w of length j is (G, w) in $\mathfrak{J}$. If G is a basic event or F, one can, by rule (*iii*) find $H_1, H_2, \cdots, H_r$ and $J_1, J_2, \cdots, J_r$ such that

$$G/0 = \bigcup_{i=1}^{r} H_i, \qquad G/1 = \bigcup_{i=1}^{r} J_i .$$

For all i, either $(H_i, 0)$ is in $\mathfrak{J}$ or $H_i = \Phi$ or H_i is a basic event, and either $(J_i, 1)$ is in $\mathfrak{J}$ or $J_i = \Phi$ or J_i is a basic event. In any case, all of $H_1, H_2, \cdots, H_r$ and $J_1, J_2, \cdots, J_r$ are in $\mathcal{S}$. If G is in $\mathcal{S}$ but G is neither a basic event not F, then it must be that (G, w) is in $\mathfrak{J}$ and $|w| < j$. But in this case, it again follows immediately from rule (*iii*) that $H_1, H_2, \cdots, H_r$ and $J_1, J_2, \cdots, J_r$ in $\mathcal{S}$ can be found with $\bigcup_{i=1}^{r} H_i = G/0$ and $\bigcup_{i=1}^{r} J_i = G/1$.

We must now put a bound on the size of $\mathcal{S}$. We do so by bounding the number of elements in the families of all (G, ϵ) in $\mathfrak{J}$. The sum of the sizes of all these families bounds the size of $\mathcal{S}$.

Suppose G consists of k states and $m = [\log_r k]$. For each $i \geqq 0$, there are at most $(2r)^i$ elements (H, w) in the family of (G, ϵ) such that $|w| = i$. If (H, w) is in the family of (G, ϵ), then H consists of at most $r^{m-|w|}$ blocks of Π_w. Thus the family of (G, ϵ) contains no pair (H, w) such that $|w| \geqq m$. An upper bound on the size of the family of (G, ϵ) is $1 + 2r + (2r)^2 + \cdots + (2r)^{m-1}$. This number does not exceed $(2r)^m/(2r - 1)$. But $m \leqq 1 + \log_r k$, so $(2r)^{m-1} \leqq k^{1+\log_r 2}$.

We may conclude that the family of (F, ϵ) consists of at most $\dfrac{2r}{2r - 1} n^{1+\log_r 2}$ elements. We must also bound the families of the basic events, and do so by the following argument.

Let N_k be the number of basic events consisting of exactly k states. There are $1 + 2 + 4 + \cdots + 2^j$ partitions Π_w where $|w| \leqq j$. The number of these partitions is at most 2^{j+1}; the blocks of each partition have among them a total of n states. Thus:

$$\sum_{k=1}^{n} kN_k \leqq n2^{j+1}. \tag{1}$$

An upper bound on the sum of the sizes of the families of all the basic events is

$$\sum_{k=1}^{n} N_k \frac{2r}{2r-1} k^{1+\log_r 2}.$$

Since k does not exceed n in the summation, we have

$$\sum_{k=1}^{n} N_k \frac{2r}{2r-1} k^{1+\log_r 2} \leqq \frac{2r}{2r-1} n^{\log_r 2} \sum_{k=1}^{n} kN_k \, .$$

Using equation (1), we see that the sum of the sizes of the families of all basic events is bounded above by $2r/(2r - 1)n^{1+\log_r 2}2^{j+1}$. Since $j \leqq 1 + \log_r n$, this bound becomes $8r/(2r - 1)n^{1+\log_r 4}$.

Including the family of (F, ϵ), we see that the size of $\mathcal{S}$ is no greater than $\dfrac{2r}{2r-1}(n^{1+\log_r 2} + 4n^{1+\log_r 4})$.

We comment that a straightforward generalization of this argument shows that every sequential machine with p binary inputs (2^p symbol input alphabet) can be realized by a network of at most $2^p r/(2^p r - 1) \cdot (n^{1+p \log_r 2} + 4^p n^{1+p \log_r 4})$ copies of the generalization of the module M_r. Thus, for any number of binary inputs p, and any $c > 0$, there are constants r and k such that any n state sequential machine with p binary inputs can be realized by a network of at most kn^{1+c} copies of a module with $2^p r$ intermodule leads.

Example: Theorem 2 suggests the design of a network of copies of M_2 for the machine of Table I with states 4, 5 and 6 final. That machine has 6 states and $[\log_2 6] = 3$. However, in this case the construction of $\mathfrak{J}$ given in Theorem 2 will not require the addition of any pair (G, w) where $|w| > 1$. So we may restrict ourselves to consideration of certain sets represented by the partitions Π_w, for $|w| \leqq 2$. These were calculated in the previous example:

$$\Pi_\epsilon = (1, 2, 3, 4, 5, 6) \qquad \Pi_{00} = (1, 2456, 3)$$

$$\Pi_0 = (1, 245, 3, 6) \qquad \Pi_{01} = (14, 2356)$$

$$\Pi_1 = (14, 26, 3, 5) \qquad \Pi_{10} = (1, 245, 3, 6)$$
$$\Pi_{11} = (145, 26, 3).$$

We begin by placing $(\{4, 5, 6\}, \epsilon)$ in $\mathfrak{I}$. $\{4, 5, 6\}/0$ is the basic event $\{2, 4, 5\}$, and $\{4, 5, 6\}/1$ is the union of three basic events $\{2, 6\}$, $\{3\}$ and $\{1, 4\}$. These three must be formed into two groups; we choose to realize $\{2, 3, 6\}$ and $\{1, 4\}$. We place $(\{2, 4, 5\}, \epsilon)$ and $(\{1, 4\}, \epsilon)$ in $\mathfrak{I}$, since these are basic events, but since $\{2, 3, 6\}$ is not a basic event, we place $(\{2, 3, 6\}, 1)$ in $\mathfrak{I}$.

$\{2, 4, 5\}/0 = \{2, 4, 5\} \cup \{6\}$, so $(\{6\}, \epsilon)$ is placed in $\mathfrak{I}$. $\{2, 4, 5\}/1$ can be expressed as $\{2, 3, 6\} \cup \{5\}$. We thus place $(\{5\}, \epsilon)$ in $\mathfrak{I}$. $\{1, 4\}/0 = \{3\}$ and $\{1, 4\}/1 = \{2, 6\}$. Each of these are basic events, so $(\{3\}, \epsilon)$ and $(\{2, 6\}, \epsilon)$ are placed in $\mathfrak{I}$. $\{2, 3, 6\}/0 = \{1\} \cup \{6\}$. These are basic events, so we add $(\{1\}, \epsilon)$ to $\mathfrak{I}$. $\{2, 3, 6\}/1 = \{1, 4\} \cup \{5\}$; these basic events are each represented in $\mathfrak{I}$ already. Proceeding, we find that the basic events added to $\mathfrak{I}$ require no new events, basic or not. The resulting network is shown in Fig. 7.

IV. CONCLUSIONS

We have considered the design of synchronous sequential machines by networks of a fixed module. This design has various advantages, including speed and ease of production using batch fabrication. It was shown that there is a family of modules M_r, $r \geqq 1$, such that any n state sequential machine with a single binary input can be realized by a network of at most p copies of M_r, when p is the minimum of $r2^{\lceil n/r \rceil}$ and $\frac{2r}{2r-1}(n^{1+\log_r 2} + 4n^{1+\log_r 4})$.

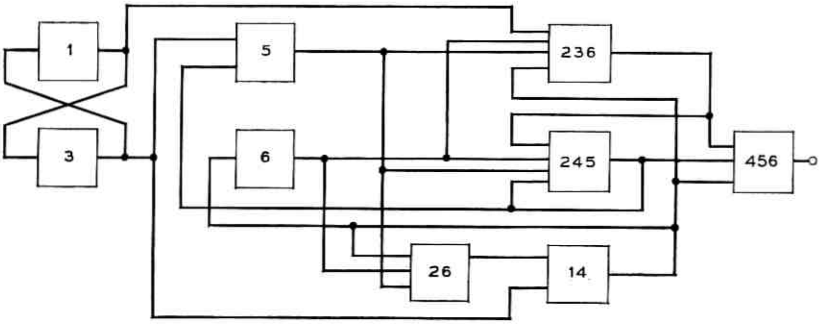


Fig. 7 — Network suggested by Theorem 2.

We feel that the type of design suggested in this paper leads to many interesting questions. In particular, the bounds expressed in Theorems 1 and 2 do not seem to be attained, or even approximated, in most cases. Efficient search techniques will probably yield much better networks than indicated; there is every reason to suspect that the bounds themselves can be improved, even if we restrict consideration to isomorphic networks.

REFERENCES

1. Weiner, P., and Hopcroft, J. E., "Modular Decomposition of Synchronous Sequential Machines," Proc. IEEE 8th Annual Symp. on Switching and Automata Theory, Austin, Texas, October 1967, pp. 233–239.
2. Hsieh, E. P., Tan, C. J., and Newborn, M. M., "Uniform Modular Realization of Sequential Machines," Proceedings of 1968 Association for Computing Machinery Conference, Las Vegas, Nevada, August 1968, pp. 613–621.
3. Weiner, P. and Hopcroft, J. E., "Bounded Fan-in, Bounded Fan-out Uniform Decompositions of Synchronous Sequential Machines," Proc. IEEE, *46,* No. 4 (July 1968), pp. 1219–1220.
4. Arnold, T. F., Tan, C. J., and Newborn, M. M., "Iteratively Realized Sequential Circuits," Proc. IEEE 9th Annual Symp. on Switching and Automata Theory, Schenectady, N. Y., October 1968, pp. 431–448.
5. Ullman, J. D. and Weiner, P., "Universal Two State Machines: Characterization Theorems and Decomposition Schemes," Proc. IEEE 9th Annual Symp. on Switching and Automata Theory, Schenectady, N.Y., October 1968, pp. 413–426.
6. Curtis, H. A., "Polylinear Sequential Circuit Realizations of Finite Automata," IEEE Transactions on Computers, *C-17,* No. 3 (March 1968) pp. 251–258.
7. Brzozowski, J. A., "Derivatives of Regular Expressions," Journal of Association for Computing Machinery *11,* No. 4 (October 1964), pp. 481–494.
8. Elgot, C. C. and Rutledge, J. D., "Operations on Finite Automata," Proc. IEEE 2nd Annual Symp. on Switching Theory and Logical Design, October 1961, pp. 129–132.
9. Hartmanis, J. and Stearns, R. E., Algebraic Structure Theory of Sequential Machines, Englewood Cliffs, N. J.: Prentice Hall, 1966.

Propagation from a Point Source in a Randomly Refracting Medium

By R. T. AIKEN

(Manuscript received August 6, 1968)

This paper considers the propagation of scalar (acoustic) waves from a single-frequency point source imbedded in a medium with random refractive index, in contrast with the usual plane-wave case in which the source is far removed from the medium. With the index being a statistically homogeneous and isotropic function of position, but not a function of time, the average complex field $\mathrm{u}_o(\mathrm{r}) = \langle \mathrm{u}(\mathrm{r}) \rangle$ *and the spatial covariance* $\langle \mathrm{u}_i(\mathrm{r})\mathrm{u}_i^*(\rho) \rangle$ *of the fluctuation field* $\mathrm{u}_i(\mathrm{r}) = \mathrm{u}(\mathrm{r}) - \mathrm{u}_o(\mathrm{r})$ *are calculated. Beyond a few correlation lengths from the source, the average field can be approximated by a spherical wave with the same complex wavenumber found in the plane-wave case. A near-source wave number is also obtained. Under an improved far-field condition, the spatial covariance is reduced to spectral integration formulas for both transverse and longitudinal separation of the receiving points. These formulas reveal that correlation lengths are much longer in the point-source case than in the plane-wave case, even though the relative variances are the same. We illustrate this result with plots for an exponential index spectrum and for a constant spectrum.*

I. INTRODUCTION

For analysis of a detection or communication system which processes signals from an array of sensors, a convenient postulate is that the signal field in the vicinity of the array is a plane wave (or perhaps a finite collection of plane waves in the multipath case). Under such a postulate, coherent addition of the sensor outputs can yield array gain and directivity in the presence of ambient noise. However, there is always some disparity between the predicted performance and the performance realized in practice. In part, the disparity can be attributed to shortcomings in the signal model, the field not being a time-invariant plane wave in the vicinity of the array. The output of a single sensor may not be constant in time but instead is apt to fade. Moreover, the outputs of different sensors do not fade "in step"; that is, after the array is steered, the signals do not fade with the unity correlation predicted by a fading-plane-wave model. Instead, the signals fade with correlation less than unity. The origin of these fading phenomena is the subject of this paper.

A simplified model of fading is considered within the framework of the following assumptions. For a short period of time, the transmission properties of a propagation path are constant; then they undergo small deviations to attain another constant configuration for the next short period of time. These short-term deviations are relative to some nominal or average configuration, as opposed to representing a slow gross trend of the overall path properties. Such short-term deviations are modeled here by the effects of random fluctuations of the index of refraction, which could be associated in the underwater acoustic case, for example, with the temperature microstructure, turbulence, and circulatory motion of water masses. Deviations of path properties associated with fluctuations of a surface of reflection are not incorporated into the model. Thus, the model is most appropriate for short-term deviations of the properties of a pure-refracted path.

In the specific situation analyzed below, the acoustic source is a single-frequency point source suspended far from any boundaries. If the refractive index were nonrandom and not position dependent, the acoustic field would be the usual spherical wave. Instead, the refractive index is a random function of position, but not of time. The average value of the index is not position dependent, so that the average line-of-sight ray path is straight rather than bent. The spatial covariance of the index is a function of the magnitude of the position-difference vector (the index is second-order homogeneous and isotropic). The problem is to find the average and spatial covariance of the acoustic field.

Much of the literature (for example, Refs. 1–3 and most of Ref. 4) is concerned not with the above spherical-wave problem but with a situation in which plane waves impinge upon a half-space with random refractive index. One essential difference is that the spherical-wave source is imbedded in the random medium whereas the plane-wave source is far removed from the random medium. Regardless of how large the distance from the spherical-wave source to an observation point becomes, this difference of configuration is preserved.†

† The configurations are called the "radio link problem" (spherical) and the "radio star problem" (planar) in Ref. 9.

Some aspects of the spherical-wave case have been treated with the Rytov method (Refs. 4–5) and other techniques (Refs. 6–8).

This analysis treats the spherical-wave problem with a version of perturbation theory previously applied to the plane-wave problem.[2,3] For distances greater than a few correlation lengths, the average field can be approximated by a spherical wave with the same complex-valued wave number previously derived in the plane-wave case.[2] A near-source wave number is also obtained. On the other hand, it is found that the covariance function of the fluctuation field exhibits much larger correlation lengths in the spherical-wave case than in the plane-wave case.[3,4] This conclusion follows from simple integration formulas for the covariances and is illustrated by plots of the covariance for special cases.

II. PERTURBATION THEORY

We consider the propagation of acoustic waves in a random time-invariant medium for the case of a monochromatic omnidirectional source. Our model is the Helmholtz equation:

$$[\nabla^2 + (1 + \mu(r))^2 k_o^2]u(r) = -\delta(r) \tag{1}$$

where ∇^2 is the Laplacian $\nabla \cdot \nabla$, $\mu(r)$ is the random deviation of the index of refraction which is a function of position r, $k_o^2 = \omega^2/c^2$, c is the sound velocity for a homogeneous medium if μ were everywhere zero, ω is the angular frequency of the source, $u(r)$ is the complex amplitude (for example, the displacement potential), and δ is the Dirac delta function. The time dependence $\exp(-i\omega t)$ has been suppressed. We assume the source is suspended far from any boundaries; that is, we consider the medium to be unbounded.

Our interest is in both the mean field $\langle u \rangle = u_c$ (coherent field) and the fluctuation field $u - u_c = u_i$ (incoherent field), where $\langle \ \rangle$ denotes expectation.

We develop a pair of equations for u_c and u_i as follows. Consider

$$[L + \epsilon L_1 + \epsilon^2 L_2]u = f \tag{2}$$

where L is a linear deterministic operator, L_1 and L_2 are linear stochastic operators, ϵ is a size parameter, and f is a deterministic forcing function. With μ in (1) replaced by $\epsilon\mu$, the correspondence of (1) and (2) is evident. We put $u = u_c + u_i$ into (2) and operate with $\langle \ \rangle$ to obtain

$$[L + \epsilon\langle L_1 \rangle + \epsilon^2 \langle L_2 \rangle]u_c = f - \epsilon\langle L_1 u_i \rangle - \epsilon^2 \langle L_2 u_i \rangle. \tag{3}$$

We then subtract (3) from (2) in which $u = u_c + u_i$ to obtain

$$[Lu_i + \epsilon(L_1u_i - \langle L_1u_i\rangle) + \epsilon^2(L_2u_i - \langle L_2u_i\rangle)] = -\epsilon(L_1 - \langle L_1\rangle)u_c - \epsilon^2(L_2 - \langle L_2\rangle)u_c . \tag{4}$$

Equation (3) shows the source f of the mean field is countered by the sink $\epsilon\langle L_1u_i\rangle + \epsilon^2\langle L_2u_i\rangle$ describing the effects of scattering into the fluctuation field. Equation (4) is not written to exhibit true sources of u_i as much as to exhibit a zero-mean forcing function and zero-mean terms on the left side. These equations are generalizations of those derived by Keller [Ref. 2, p. 166, equations (12) through (13)] for other purposes.

Solution of (3) and (4) can proceed with perturbation theory for the case of small ϵ. Relative to $\epsilon \to 0$, equation (3) exhibits $u_c = O(1)$ and (4) exhibits $u_i = O(\epsilon)$. Accordingly, (3)–(4) can be rewritten

$$[L + \epsilon\langle L_1\rangle + \epsilon^2\langle L_2\rangle]u_c = f - \epsilon\langle L_1u_i\rangle + O(\epsilon^3) \tag{5}$$

$$Lu_i = -\epsilon(L_1 - \langle L_1\rangle)u_c + O(\epsilon^2). \tag{6}$$

These equations can be partially uncoupled by operating with L^{-1} on (6) and substituting into (5) to obtain

$$[L + \epsilon\langle L_1\rangle + \epsilon^2\langle L_2\rangle]u_c = f + \epsilon^2\langle L_1L^{-1}L_1\rangle u_c - \epsilon^2\langle L_1\rangle L^{-1}\langle L_1\rangle u_c + O(\epsilon^3) \tag{7}$$

$$u_i = -\epsilon L^{-1}(L_1 - \langle L_1\rangle)u_c + O(\epsilon^2). \tag{8}$$

Equation (7) for the mean field u_c is the result obtained by Keller (Ref. 2, p. 148, equation (10)), who used a successive-substitution solution of (2) in conjunction with a crucial, and at-first-glance mysterious, replacement of $L^{-1}f$ by $\langle u\rangle$. Equation (1.8) is a version of Keller's equation (31) on p. 169 of Ref. 2. Thus, we have shown that these equations arise quite naturally from the pair (3) and (4).

We now specialize (7) and (8) to the case of the Helmholtz equation (1). Here,

$$L = \nabla^2 + k_o^2 \qquad L_1 = 2\mu(r)k_o^2 \qquad L_2 = \mu^2(r)k_o^2 . \tag{9}$$

We assume $\langle\mu(r)\rangle = 0$; that is to say, we neglect any systematic dependence of refractive index upon position (the average profile). We have

$$L^{-1}g = -\int \frac{\exp(ik_o \mid r - r' \mid)}{4\pi \mid r - r' \mid} g(r')\, dr', \tag{10}$$

where the integral is over all space. The inverse L^{-1} is an integral operator with kernel corresponding to the Green's function

$$G(r, r') = -\frac{\exp[ik_o \mid r - r' \mid]}{4\pi \mid r - r' \mid}. \tag{11}$$

Thus, the pair (7) and (8) is specialized to

$$[\nabla^2 + k_o^2(1 + \epsilon^2\langle\mu^2(r)\rangle)]u_c(r)$$
$$= -\delta(r) - 4\epsilon^2 k_o^4 \int \frac{\exp[ik_o \mid r - r' \mid]}{4\pi \mid r - r' \mid} \langle\mu(r)\mu(r')\rangle u_c(r')\, dr' + O(\epsilon^3) \tag{12}$$

$$u_i(r) = 2\epsilon k_o^2 \int \frac{\exp[ik_o \mid r - r' \mid]}{4\pi \mid r - r' \mid} \mu(r')u_c(r')\, dr' + O(\epsilon^2). \tag{13}$$

III. THE AVERAGE FIELD

We now develop an approximation of the solution of (12) for the average field u_c. It is assumed that the refractive index is statistically homogeneous and isotropic. The index covariance function is

$$\Gamma(\mid r - r' \mid) = \langle\mu(r)\mu(r')\rangle. \tag{14}$$

Equation (12) becomes

$$\{\nabla^2 + k_o^2[1 + \epsilon^2\Gamma(0)]\}u_c(r)$$
$$= -\delta(r) - 4\epsilon^2 k_o^4 \int \frac{\exp[ik_o \mid \rho \mid]}{4\pi \mid \rho \mid} \Gamma(\mid \rho \mid)u_c(r + \rho)\, d\rho + O(\epsilon^3). \tag{15}$$

We assume an approximation of $u_c(r)$ of the form

$$\frac{\exp[ik \mid r \mid]}{4\pi \mid r \mid} \tag{16}$$

where k is a constant wave number to be determined ($k \neq k_o$). It will be found that (16) is not a global solution, because a constant k cannot exist. Nevertheless, (16) can serve as a useful local approximation of the solution, with k interpreted as a weak and slowly varying function of $\mid r \mid$.

If (16) were the solution, then u_c would satisfy

$$[\nabla^2 + k^2]u_c(r) = -\delta(r). \tag{17}$$

Then (15) and (17) yield

$$\{k_o^2[1 + \epsilon^2\Gamma(0)] - k^2\}u_c(r)$$
$$= -4\epsilon^2 k_o^4 \int \frac{\exp [ik_o \mid \rho \mid]}{4\pi \mid \rho \mid} \Gamma(\mid \rho \mid) u_c(r + \rho)\, d\rho + O(\epsilon^3). \qquad (18)$$

The volume integral in (18) can be evaluated by an integration over the surface of a sphere with radius R followed by a radial integration from $R = 0$ to $R = \infty$. For the surface integration, we need only observe

$$\int_{S:|\rho|=R} u_c(r + \rho) \frac{dS}{4\pi R^2}$$
$$= \begin{cases} u_c(r) \dfrac{\sin kR}{kR}, & 0 < R < \mid r \mid \\[2ex] u_c(r) \sin k \mid r \mid \dfrac{\exp [ik(R - \mid r \mid)]}{kR}, & R > \mid r \mid \end{cases} \qquad (19)$$

where dS is a differential of area on the sphere $S = \{\rho \colon \mid \rho \mid = R\}$. This mean value theorem follows from (16) and (17); see Appendix A. Then (19) inserted into (18) yields

$$\{k^2 - k_o^2[1 + \epsilon^2\Gamma(0)]\}u_c(r)$$
$$= \frac{4\epsilon^2 k_o^4}{k} u_c(r) \left[\int_0^{|r|} \exp (ik_o R)\Gamma(R) \sin kR\, dR \right.$$
$$\left. + \int_{|r|}^{\infty} \exp (ik_o R)\Gamma(R) \sin k \mid r \mid \exp [ik(R - \mid r \mid)]\, dR \right] + O(\epsilon^3). \qquad (20)$$

If (16) were an exact global solution, then $u_c(r)$ could be cancelled in (20); the result would be a relation for the supposedly constant wave number k. But the integrals in (20) suggest that the relation is $\mid r \mid$-dependent, which is a contradiction. Nevertheless, (16) will serve as a local approximation of $u_c(r)$ in regions in which k is virtually constant.

The following manipulations are made upon the integrals in (20). We run the first integral from 0 to ∞ and correct for its contribution from $\mid r \mid$ to ∞ by another term in the second integral. We then change the variable of integration of the resultant second integral. Then (20) becomes

$$k^2 = k_0^2(1 + \epsilon^2\Gamma(0)) + \frac{4\epsilon^2 k_o^4}{k} \left[\int_0^{\infty} \exp (ik_o R)\Gamma(R) \sin kR\, dR \right.$$

$$- \exp [i(k_o - k) \mid r \mid] \int_0^\infty \exp (ik_o R) \Gamma(\mid r \mid + R) \sin kR \, dR \Big] + O(\epsilon^3) \tag{21}$$

The $|r|$-dependence is now confined to the second integral. The large-$|r|$ case occurs when we can assume this integral to be negligible, namely

$$|\exp [i(k_o - k) \mid r \mid]| \, \Gamma(\mid r \mid + R) \ll \Gamma(R), \qquad R \, \varepsilon \, [0, R_o], \tag{22}$$

where we assume the first integration can run from 0 to R_o with little error. The condition (22) shows that $|r|$ must be much larger than a correlation distance; moreover, (22) shows that the increasing function $\exp[(\mathrm{Im}\, k) \ |r|\ ]$ must be taken into account.

Thus, for large $|r|$, the wave number k satisfies

$$k^2 \approx k_o^2[1 + \epsilon^2 \Gamma(0)] + \frac{4\epsilon^2 k_o^4}{k} \int_0^\infty \exp (ik_o R) \Gamma(R) \sin kR \, dR + O(\epsilon^3). \tag{23}$$

This is the relation found by Keller [Ref. 2, p. 151, equation (14)] for the plane-wave problem. As expected, the spherical wave solution far from the source has the same wave number as the plane-wave solution.

The small-$|r|$ case occurs when the integrals in (21) nearly cancel one another; the wave number k is given by

$$k^2 \approx k_0^2[1 + \epsilon^2 \Gamma(0)] + O(\epsilon^3) \tag{24a}$$

or

$$k \approx k_o[1 + \tfrac{1}{2}\epsilon^2 \Gamma(0)] + O(\epsilon^3). \tag{24b}$$

Whereas (24) yields the small-$|r|$ values of k directly, notice that (23) determines the large-$|r|$ values of k in an implicit fashion. However, an explicit approximation of the large-$|r|$ value of k can be obtained. Notice that (23) could be solved by successive substitutions, the first step employing either k_o in the integral below or employing (24) as follows

$$\frac{4\epsilon^2 k_o^4}{k} \int_0^\infty \exp (ik_o R) \Gamma(R) \sin kR \, dR$$
$$\approx \frac{4\epsilon^2 k_o^4}{k_o[1 + \frac{1}{2}\epsilon^2 \Gamma(0)]} \int_0^\infty \exp (ik_o R) \Gamma(R) \sin \{k_o[1 + \tfrac{1}{2}\epsilon^2 \Gamma(0)] R\} \, dR$$

$$\approx 4\epsilon^2 k_0^3 \int_0^\infty \exp(ik_oR)\Gamma(R)\{\sin(k_oR)\cos k_o[\tfrac{1}{2}\epsilon^2\Gamma(0)R] + \cos(k_oR)\sin k_o[\tfrac{1}{2}\epsilon^2\Gamma(0)R]\}\,dR$$

$$\approx 4\epsilon^2 k_0^3 \int_0^\infty \exp(ik_oR)\Gamma(R)\sin k_oR\,dR. \tag{25}$$

Since terms have been discarded consistently insofar as powers of ϵ are concerned, approximations (23) and (25) yield

$$k^2 \approx k_o^2[1 + \epsilon^2\Gamma(0)] + 4\epsilon^2 k_o^3 \int_0^\infty \exp(ik_oR)\Gamma(R)\sin k_oR\,dR + O(\epsilon^3). \tag{26}$$

From (26), it follows that

$$k \approx k_o[1 + \tfrac{1}{2}\epsilon^2\Gamma(0)] + 2\epsilon^2 k_o^2 \int_0^\infty \exp(ik_oR)\Gamma(R)\sin k_oR\,dR + O(\epsilon^3), \tag{27}$$

or equivalently

$$\operatorname{Re} k \approx k_o[1 + \tfrac{1}{2}\epsilon^2\Gamma(0)] + \epsilon^2 k_o^2 \int_0^\infty \Gamma(R)\sin 2k_oR\,dR + O(\epsilon^3), \tag{28}$$

and

$$\operatorname{Im} k \approx \epsilon^2 k_o^2 \int_0^\infty (1 - \cos 2k_oR)\Gamma(R)\,dR + O(\epsilon^3). \tag{29}$$

If Γ has a correlation length L_o and if $k_oL_o \gg 1$ (a large-scale condition not yet imposed), then

$$\operatorname{Im} k \approx \epsilon^2 k_o^2 \int_0^\infty \Gamma(R)\,dR. \tag{30}$$

Also, accuracy of the approximation (25) requires the bracketed factor in the integrand to be equivalent to $\sin k_oR$; this holds when

$$k_o\epsilon^2\Gamma(0)L_o \ll 1. \tag{31}$$

But

$$\int_0^\infty \Gamma(R)\,dR \sim \Gamma(0)L_o\,,$$

and (31) is equivalent to

$$\operatorname{Im} k \ll k_o \, . \tag{32}$$

When (32) is not met, neither (28) nor (30) can be expected to be a good approximation. Also, for the successive-substitution procedure to yield a good approximation at this first step, it appears sufficient that the first step value (28) be well approximated by the initial value (24); equivalently,

$$\Gamma(0) \gg k_o \int_0^\infty \Gamma(R) \sin 2k_oR \, dR, \tag{33}$$

which is a restriction on the large-wave number value of an integral which resembles the spectrum of Γ.

The approximation (16) for the average field $u_c(r)$ together with (23) and (24) for the large-$|r|$ and small-$|r|$ values of k comprise the principal results of this section. The further approximations (28) through (30) for the large-$|r|$ case are more useful than (23), but conditions (31) through (33) must be met. When (28) through (30) are compared with the small-$|r|$ approximation (24), it can be seen that the spherical wave (16) develops attenuation and a change in phase velocity as $|r|$ increases. The transition from small-$|r|$ to large-$|r|$ behavior occurs when (22) begins to hold, namely, when the second integral in (21) begins to become negligible. The order of magnitude of this transitional value of $|r|$ is a few correlation lengths.

IV. COVARIANCE OF THE FLUCTUATION FIELD

The previous section provides a solution of (7) or (25) for the average field $u_c(r)$ which now can be used in (8) or (13) to yield the fluctuation field $u_i(r)$. Thus, (11) and (13) yield

$$u_i(r) = -2\epsilon k_o^2 \int G(r, r')\mu(r')u_c(r') \, dr' + O(\epsilon^2), \tag{34}$$

where $u_c(r')$ is given by (16) in which k is a weak function of $|r'|$.

The spatial covariance function $\langle u_i(r)u_i^*(\rho)\rangle$ is now computed for the case in which the medium is statistically homogeneous and isotropic. Equations (34) and (14) yield

$$\langle u_i(r)u_i^*(\rho)\rangle = 4\epsilon^2k_o^2 \iint G(r, r')G^*(\rho, \rho') \cdot \Gamma(|r' - \rho'|)u_c(r')u_c^*(\rho') \, dr' \, d\rho' + O(\epsilon^3). \tag{35}$$

It is convenient to change to the following variables of integration (with unity Jacobian):

$$y = \frac{r' + \rho'}{2}, \qquad x = r' - \rho', \tag{36}$$

where

$$r' = y + \frac{x}{2}, \qquad \rho' = y - \frac{x}{2}. \tag{37}$$

Moreover, it is convenient to evaluate the fields at the following points:

$$r = \eta + \frac{\xi}{2}, \qquad \rho = \eta - \frac{\xi}{2}, \tag{38}$$

where, by definition,

$$\eta = \frac{r + \rho}{2}, \qquad \xi = r - \rho. \tag{39}$$

The relation of the positions (36)–(39) is shown in Fig. 1. The covariance of the fluctuation field u_i is thus

$$\begin{aligned}\left\langle u_i\left(\eta + \frac{\xi}{2}\right)u_i^*\left(\eta - \frac{\xi}{2}\right)\right\rangle \\ = 4\epsilon^2 k_o^4 \iint G\left(\eta + \frac{\xi}{2}, y + \frac{x}{2}\right)G^*\left(\eta - \frac{\xi}{2}, y - \frac{x}{2}\right)\Gamma(\mid x \mid) \\ \cdot u_c\left(y + \frac{x}{2}\right)u_c^*\left(y - \frac{x}{2}\right) dx\, dy + O(\epsilon^3).\end{aligned} \tag{40}$$

In words, the second-moment of the fields at observation center η with observation position-difference vector ξ comprises the integrated effect of scattering of the average field by the refractive index at scattering center y with scattering position-difference vector x.

We now approximate the integrand of (40). Although the approxi-

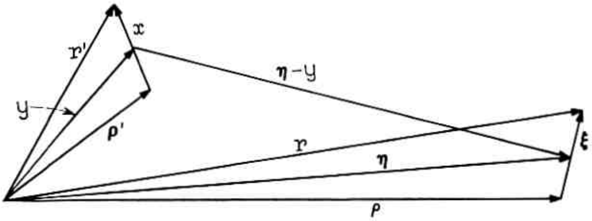


Fig. 1 — Scattering points r', ρ' and receiver points r, ρ.

mations are not valid over all space, they are valid for a region which can account for the major contribution to (40) in the case to be described later. The first approximation involves

$$G(r, r')G^*(\rho, \rho') = \frac{\exp [ik_o(| r - r' | - | \rho - \rho' |)]}{(4\pi)^2 \, | r - r' | \cdot | \rho - \rho' |}. \tag{41}$$

But

$$\begin{aligned} | r - r' | &= | \eta - y + \tfrac{1}{2}(\xi - x)| \\ &= | \eta - y | + \frac{\frac{1}{2}(\xi - x)\cdot(\eta - y)}{| \eta - y |} + \cdots \end{aligned} \tag{42a}$$

and

$$\begin{aligned} | \rho - \rho' | &= | \eta - y - \tfrac{1}{2}(\xi - x)| \\ &= | \eta - y | - \frac{\frac{1}{2}(\xi - x)\cdot(\eta - y)}{| \eta - y |} + \cdots . \end{aligned} \tag{42b}$$

The above expansions in powers of $(\xi - x)$ are appropriate for a large vector $\eta - y$ as perturbed by the small vectors $\pm\frac{1}{2}(\xi - x)$. Our approximation of (41) is

$$G\left(\eta + \frac{\xi}{2}, y + \frac{x}{2}\right)G^*\left(\eta - \frac{\xi}{2}, y - \frac{x}{2}\right)$$

$$\approx \frac{\exp ik_o\left[\dfrac{(\eta - y)\cdot(\xi - x)}{| \eta - y |}\right]}{(4\pi)^2 \, | \eta - y |^2} \tag{43}$$

$$= \frac{\exp [ik_s(y)\cdot(\xi - x)]}{(4\pi)^2 \, | \eta - y |^2} \tag{44}$$

where the relation

$$k_s(y) = k_o \frac{\eta - y}{| \eta - y |} \tag{45}$$

defines a scattering wavevector .

The second approximation involves replacing the coherent-field factor $u_c(y + x/2)u_c^*(y - x/2)$ by a function that locally represents the fields as plane waves. Thus, it follows from (16) that

$$u_c\left(y + \frac{x}{2}\right)u_c^*\left(y - \frac{x}{2}\right) = \frac{\exp \left[i\left(k \left| y + \dfrac{x}{2} \right| - k^* \left| y - \dfrac{x}{2} \right|\right)\right]}{(4\pi)^2 \left| y + \dfrac{x}{2} \right| \left| y - \dfrac{x}{2} \right|} \tag{46}$$

where the wavenumber k is a weak function of position. But

$$\left| y + \frac{x}{2} \right| = | y | + \frac{1}{2} \frac{y}{| y |} \cdot x + \cdots \tag{47a}$$

$$\left| y - \frac{x}{2} \right| = | y | - \frac{1}{2} \frac{y}{| y |} \cdot x + \cdots , \tag{47b}$$

leads to the approximation

$$u_c\left(y + \frac{x}{2}\right) u_c^*\left(y - \frac{x}{2}\right) \approx | u_c(y) |^2 \, e^{ik(y) \cdot x}, \tag{48}$$

where

$$k(y) = (\mathrm{Re}\, k) \frac{y}{| y |} \tag{49}$$

defines an incident wavevector, and

$$| u_c(y) |^2 = \frac{\exp \left(-2 \ \mathrm{Im}\, k \ | y |\right)}{(4\pi)^2 \ | y |^2}. \tag{50}$$

Collecting these approximations into (40) yields

$$\begin{aligned} \left\langle u_i\left(\eta + \frac{\xi}{2}\right) u_i^*\left(\eta - \frac{\xi}{2}\right) \right\rangle & \\ \approx 4\epsilon^2 k_0^4 \int dy \, & \frac{\exp \left[ik_s(y) \cdot \xi\right] \exp \left[-2 \ \mathrm{Im}\, k \ | y |\right]}{(4\pi)^4 \ | \eta - y |^2 \ | y |^2} \\ & \cdot \int dx \ \Gamma(| x |) \exp \left\{i[k(y) - k_s(y)] \cdot x\right\}. \end{aligned} \tag{51}$$

This is the central result of this section. Equation (51) has the physical interpretation of a volume distribution of sources. The source at y generates a plane wave at the receiver with correlation $\exp[ik_s(y) \cdot \xi]$. The strength of this wave is proportional to $| \eta - y |^{-2}$ $| y |^{-2}$ and to the value of the spectrum

$$S(| \kappa |) = \int dx \ \Gamma(| x |) \exp \left\{i\kappa \cdot x\right\} \tag{52}$$

as evaluated at the local wavevector $k(y) - k_s(y)$. At this wave vector, the spectrum is a measure of the amplitude of those components of refractive index with the orientation and the periodicity required for constructive interference (Bragg scattering; compare with Ref. 4, pp. 68–69).

The physical justification of the above approximations follows from (40) by noticing the role played by the index covariance Γ in the integrand. The weighting introduced by Γ means that scattering from center y depends upon the neighborhood of y with linear extent L_o, where L_o is the outer scale. First, the local plane-wave approximation (48) is poorest near the origin where the wavefronts are most curved. With a criterion of not more than $\pi/16$ radian departure from plane-wave phase, Fig. 2 shows that $| y |$ must be larger than $4L_0^2/\lambda$. In fact this usual far-field condition can be replaced by $| y | > 4L_o(L_o/\lambda)^{\frac{1}{2}}$ which is less restrictive when $L_o > \lambda$. This weaker condition, derived in Appendix B, follows from an overbound of the phase error in (48) caused by eliminating the remainder of (47). Second, the scattering approximation (43) is poorest near the observation center η where the phase (and amplitude) of (41) can experience large excursions as r', ρ' range over a neighborhood of linear size L_o. A usual far-field condition is $| \eta - y | > 4L_o^2/\lambda$ or $| \eta - y | > 4(| \xi | + L_o)^2/\lambda$. Again, when $L_o > \lambda$, only a weaker condition,

$$| \eta - y | > 4(| \xi | + L_o)\left(\frac{| \xi | + L_o}{\lambda}\right)^{\frac{1}{2}}, \tag{53}$$

need be met. Condition (53) follows from an overbound of the phase error in (43) associated with the remainder in (52), (see Appendix B). Strictly speaking, the y-integration in (51) must exclude the near-source and near-receiver spheres of radius $4L_o(L_o/\lambda)^{\frac{1}{2}}$, and their contributions must be evaluated separately. In Section V we give a condition necessary for this contribution to be negligible.

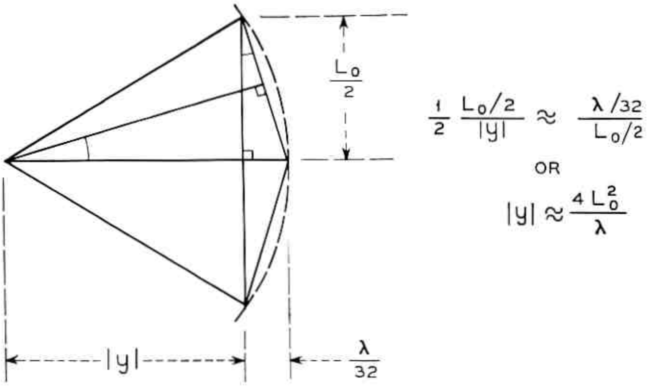

$\frac{L_o}{2}$
$\frac{1}{2}\,\frac{L_o/2}{|y|} \approx \frac{\lambda/32}{L_o/2}$
OR
$|y| \approx \frac{4L_o^2}{\lambda}$
$|y|$
$\frac{\lambda}{32}$

Fig. 2 — Distance for the plane-wave approximation [in fact, only $4L_o(L_o/\lambda)^{1/2}$ required].

Apart from the excluded regions of integration, the validity of approximation (51) does not rely upon a "large-scale" condition requiring the wavelength λ to be much smaller than some refractive-index scale size. But when such a condition is met, (51) yields both a maximum angle of important scattering and a finite volume of important scattering. In the approximation (51), the refractive-index spectrum (52) is evaluated at the local wave vector,

$$k(y) - k_s(y). \tag{54}$$

Suppose there exists an inner scale l_o such that for $|\kappa| > 2\pi/l_o$ the spectrum (52) is negligible. Since the maximum magnitude (54) can attain is of order $4\pi/\lambda$, whereas $2\pi/\lambda \gg 2\pi/l_o$, it follows that the integrand of (51) is large only for values of y such that

$$|k(y) - k_s(y)| < 2\pi/l_o . \tag{55}$$

Under the assumption that $|k(y)| = k_o = 2\pi/\lambda$, condition (55) yields the maximum angle of important scattering. With $\psi(y)$ the angle between $k(y)$ and $k_s(y)$, as shown in Fig. 3, we have

$$|k(y) - k_s(y)|^2 = 2k_o^2 - 2k_o^2 \cos \psi(y) = 4k_o^2 \sin^2 \frac{\psi(y)}{2}. \tag{56}$$

Then (55) and (56) yield $\cos \psi(y) > 1 - \lambda^2/2l_o^2$ or $2l_o/\lambda \sin \psi(y)/2 < 1$ or approximately $\psi(y) < \lambda/l_o$.

These conditions may be used to find the region of important scattering. Figure 4 shows cylindrical coordinates with origin at the midpoint between transmitter and receiver; there is rotational symmetry around the transmitter-receiver axis. With $\tan \psi$ constant, we have

$$\tan \psi = \tan (\alpha + \beta) = \frac{\tan \alpha + \tan \beta}{1 - \tan \alpha \tan \beta}. \tag{57}$$

But

$$\tan \alpha = \frac{b}{\frac{L}{2} + a} \tag{58}$$

$$\tan \beta = \frac{b}{\frac{L}{2} - a}. \tag{59}$$

Algebraic manipulations which include completing a square yield

$$a^2 + \left(b + \frac{L}{2 \tan \psi}\right)^2 = \left(\frac{L}{2 \sin \psi}\right)^2. \tag{60}$$

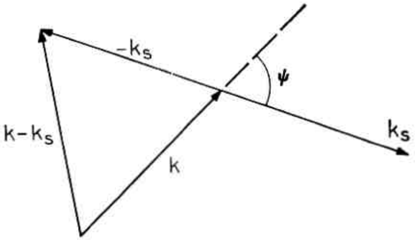


Fig. 3 — Angle of scattering.

Equation (60) is a circle in the a-b plane passing through the transmitter and receiver-center locations, Fig. 5. The slope of the curve is

$$\frac{db}{da} = -\frac{a}{b + \dfrac{L}{2\tan\psi}}. \tag{61}$$

Thus, near the transmitter,

$$\frac{b}{\dfrac{L}{2} + a} \approx \left.\frac{db}{da}\right|_{(-L/2,0)} = \tan\psi, \tag{62}$$

and near the receiver

$$\frac{b}{\dfrac{L}{2} - a} \approx -\left.\frac{db}{da}\right|_{(L/2,0)} = \tan\psi. \tag{63}$$

Also, at the midpoint $a = 0$, (60) yields

$$b = \frac{L}{2}\left(\frac{1}{\sin\psi} - \frac{1}{\tan\psi}\right) = \frac{L}{2}\tan\frac{\psi}{2}. \tag{64}$$

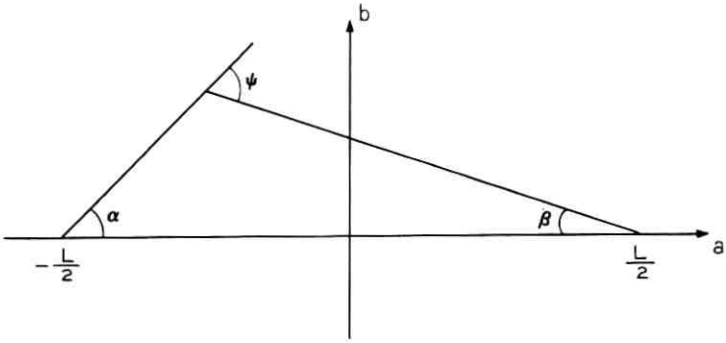


Fig. 4 — Coordinates for the region of important scattering.

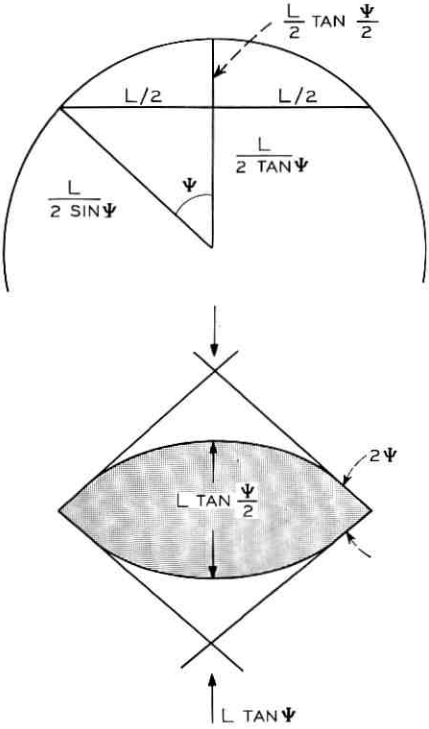


Fig. 5 — Region of important scattering (spherical waves).

The condition for important scattering is

$$\cos \psi > 1 - \frac{\lambda^2}{2l_o^2} \triangleq \cos \Psi. \tag{65}$$

The volume specified by (65) is enclosed within the surface generated by rotating the arc of the circle (60), with $\psi = \Psi$, around the transmitter-receiver axis. This volume lies within the volume common to two cones with apexes at transmitter and receiver, each cone with half-angle Ψ.

For the large-scale case, $\lambda \ll l_o \leqq L_o$, a condition necessary for (51) to be an accurate approximation of the covariance (40) is now apparent. The volume of important scattering shown in Fig. 5 must be much larger than the volumes in which the integrand of (51) is a poor approximation of the integrand of (40). These comprise a near-source cone of axial length $4L_o^{3/2}/\lambda^{\frac{1}{2}}$ and a near-receiver cone of axial length $4(L_o + |\xi|)^{3/2}/\lambda^{\frac{1}{2}}$.

An equivalent condition is seen to be that the transmitter-receiver distance must be much larger than the axial lengths of these cones, that is to say,

$$L = |\eta| \gg 4(L_o + |\xi|)^{3/2}/\lambda^{\frac{1}{2}}. \tag{66}$$

It remains to observe that the covariance expression (40) is itself an accurate relation provided that the average field is not severely attenuated by virtue of scattering into the fluctuation field. The attenuation exhibited in (50), as evaluated throughout the above region of important scattering, must be small; that is to say,

$$(\mathrm{Im}\, k)\, |\eta| \ll 1, \tag{67}$$

where Im k is given by (30). Combining conditions (66) and (67) yields an interval for validity of (51). When $|\xi| = 0$, this interval is

$$L_o^{3/2}/\lambda^{\frac{1}{2}} \ll |\eta| \ll (\mathrm{Im}\, k)^{-1}. \tag{68}$$

In other words, the transmitter-to-receiver distance must be (*i*) sufficiently large so that far-field approximations of the covariance are valid, and (*ii*) sufficiently small so that single-scatter perturbation approximations are valid.

V. REDUCTION OF THE INTEGRATION FOR THE COVARIANCE-SPHERICAL AND PLANAR CASES

5.1 *Spherical-Wave Case*

The central result of the previous section is the approximation (51) of the covariance. The problem remains to evaluate the integral specified by this approximation. In this section, we introduce a set of coordinates which simplifies the integration, the result being (77). Although Section V indicates the extent of the important region of integration, the result (77) is equivalent to integration over all space rather than over only the important region. Under the large-scale approximation, the formula (77) is specialized to (83) and to (85) for transverse and longitudinal receiver separations.

For simplicity, we first observe that (51) can be replaced by an expression employing the unperturbed field u_o rather than the average field u_c. We need only observe from (7) that

$$u_c = L^{-1}f + O(\epsilon^2)$$

where $\langle L_1 \rangle = 0$. It immediately follows that (8) can be replaced by

$$u_i = -\epsilon L^{-1}L_1 u_o + O(\epsilon^2)$$

where $u_o = L^{-1}f$ is the field that would exist in the nonrandom medium $(\epsilon = 0)$. In our special case,

$$u_o(r) = \frac{\exp [ik_o \mid r \mid]}{4\pi \mid r \mid},$$

and accordingly

$$u_i(r) = -2\epsilon k_o^2 \int G(r, r')\mu(r')u_o(r') \, dr' + O(\epsilon^2)$$

can replace (34). Equivalently, (51) can be approximated by

$$\left\langle u_i\left(n + \frac{\xi}{2}\right)u_i^*\left(\eta - \frac{\xi}{2}\right)\right\rangle$$

$$\approx 4\epsilon^2 k_o^4 \int dy \frac{\exp [ik_s(y) \cdot \xi]}{(4\pi)^4 \mid y \mid^2 \mid \eta - y \mid^2} S(\mid k_o\hat{y} - k_s(y) \mid), \tag{69}$$

where the spectrum S is defined by (52) and where $\hat{y} = y/\mid y \mid$. That is to say, the replacement of u_c by u_o corresponds to the replacement of $k(y)$ by $k_o\hat{y}$.

The volume integration can be carried out with spherical coordinates which have the receiving center η as their origin. In such coordinates, the differential of volume of $d\Omega dR \, R^2$, where $R = \mid \eta - y \mid$ and $d\Omega$ is the differential of the solid angle. Since $k_s(y)$ is a function only of the direction of an element $d\Omega$ relative to the origin at η, it follows that (69) equals

$$\frac{4\epsilon^2 k_o^4}{(4\pi)^4} \int d\Omega \exp [ik_s \cdot \xi] \int dR \, \{\mid y \mid^{-2} S(\mid k_o\hat{y} - k_s(y) \mid)\} \tag{70}$$

where the factor in braces is to be evaluated as a function of R with k_s fixed.

The angular integration in (70) will use the coordinates in Fig. 6, where $\theta = 0$ corresponds to the direction of the transmitter. The radial integration in (70) will employ the angle ψ shown in Fig. 7. The argument of the spectrum is the square root of

$$\mid k_o\hat{y} - k_s \mid^2 = k_o^2(2 - 2 \cos \psi) = 4k_o^2 \sin^2 \frac{\psi}{2}. \tag{71}$$

The law of sines is

$$\frac{\mid y \mid}{\sin \theta} = \frac{L}{\sin (\pi - \psi)} = \frac{R}{\sin (\psi - \theta)}, \tag{72}$$

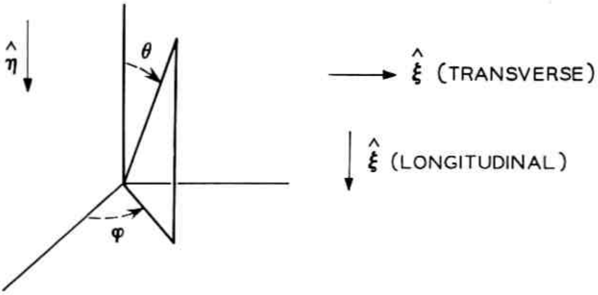


Fig. 6 — Spherical polar coordinates at the receiver.

and thus

$$| y | = L \frac{\sin \theta}{\sin \psi} \tag{73}$$

$$\frac{dR}{d\psi} = L \frac{\sin \theta}{\sin^2 \psi}. \tag{74}$$

The radial integral in (70) is thus

$$(L \sin \theta)^{-1} \int_\theta^\pi d\psi \, S\left(2k_o \sin \frac{\psi}{2}\right), \tag{75}$$

and (70) becomes

$$\frac{4\epsilon^2 k_o^4 L}{(4\pi)^2 (4\pi L)^2} \int_0^\pi d\theta \int_0^{2\pi} d\varphi \exp (ik_s \cdot \xi) \int_\theta^\pi d\psi \, S\left(2k_o \sin \frac{\psi}{2}\right). \tag{76}$$

The θ-ψ integration is over the triangle $\{0 \leqq \theta \leqq \pi, \theta \leqq \psi \leqq \pi\} = \{0 \leqq \psi \leqq \pi, 0 \leqq \theta \leqq \psi\}$, so that interchange of the order of integration yields

$$\frac{\epsilon^2 k_o^4 L}{(2\pi)^2 (4\pi L)^2} \int_0^\pi d\psi \, S\left(2k_o \sin \frac{\psi}{2}\right) \int_0^\psi d\theta \int_0^{2\pi} d\varphi \exp (ik_s \cdot \xi). \tag{77}$$

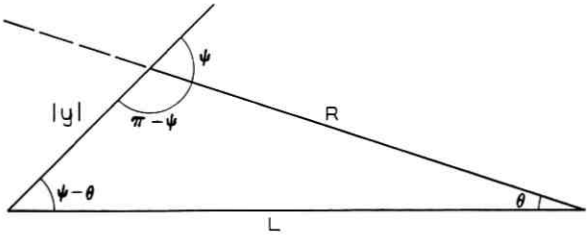


Fig. 7 — Radial variable R related to angle ψ.

Further specialization of (77) is made to the cases in which the receiving displacement ξ is transverse and is longitudinal, Fig. 6. In the transverse case, $k_s \cdot \xi = -k_o \mid \xi \mid \sin \theta \sin \varphi$, and (77) becomes

$$\frac{\epsilon^2 k_o^4 L}{2\pi(4\pi L)^2} \int_0^{\pi} d\psi \; S\left(2k_o \sin \frac{\psi}{2}\right) \int_0^{\psi} d\theta \; J_o(k_o \mid \xi \mid \sin \theta). \tag{78}$$

In the longitudinal case, $k_s \cdot \xi = k_o \mid \xi \mid \cos \theta$, and (77) becomes

$$\frac{\epsilon^2 k_o^4 L}{2\pi(4\pi L)^2} \int_0^{\pi} d\psi \; S\left(2k_o \sin \frac{\psi}{2}\right) \int_0^{\psi} d\theta \; \exp\,(ik_o \mid \xi \mid \cos \theta). \tag{79}$$

Expressions (77) to (79) correspond to integration over all space, rather than over only the region of important scattering. Further approximations rely upon the cutoff provided by $S(\kappa)$ for $\kappa > 2\pi/l_o$, where l_o is the inner scale size and $l_o \gg \lambda$. For the transverse case, (78) becomes

$$\frac{\epsilon^2 k_o^2 L}{2\pi(4\pi L)^2} \int_0^{\infty} dx \; S(x) \int_0^{x} d\kappa \; J_o(\kappa \mid \xi \mid), \tag{80}$$

and in the longitudinal case, (79) yields

$$\frac{\epsilon^2 k_o^2 L}{2\pi(4\pi L)^2} \exp\,(ik_o \mid \xi \mid) \int_0^{\infty} dx \; S(x) \int_0^{x} d\kappa \; \exp\left(-i \frac{\mid \xi \mid \kappa^2}{2k_o}\right). \tag{81}$$

The κ-integral in (80) can be evaluated in closed form, namely

$$\int_0^{x} d\kappa \; J_o(\kappa \mid \xi \mid) = xJ_o(x \mid \xi \mid) + \frac{\pi}{2} xJ_1(x \mid \xi \mid)H_o(x \mid \xi \mid) - \frac{\pi}{2} xJ_o(x \mid \xi \mid)H_1(x \mid \xi \mid) \tag{82}$$

where H_ν are the Struve functions. Thus, for the transverse case, (80) is

$$\frac{\epsilon^2 k_o^2 L}{2\pi(4\pi L)^2} \int_0^{\infty} dx \; xS(x)\left[J_o(x \mid \xi \mid) + \frac{\pi}{2} J_1(x \mid \xi \mid)H_o(x \mid \xi \mid) - \frac{\pi}{2} J_o(x \mid \xi \mid)H_1(x \mid \xi \mid)\right]. \tag{83}$$

The κ-integral in (81) is related to the Fresnel integrals, namely

$$\begin{aligned}\int_0^{x} d\kappa \; \exp\left(-i \frac{\mid \xi \mid \kappa^2}{2k_o}\right) &= \left(\frac{\pi k_o}{\mid \xi \mid}\right)^{\frac{1}{2}} \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \int_0^{(x^2 \mid \xi \mid / 2k_o)^{\frac{1}{2}}} dt \; \exp\,(-it^2) \\ &= \left(\frac{\pi k_o}{\mid \xi \mid}\right)^{\frac{1}{2}} \left[\hat{C}\left(\frac{x^2 \mid \xi \mid}{2k_o}\right)^{\frac{1}{2}} - i\hat{S}\left(\frac{x^2 \mid \xi \mid}{2k_o}\right)^{\frac{1}{2}}\right].\end{aligned} \tag{84}$$

Thus, for the longitudinal case, (81) is

$$\frac{\epsilon^2 k_o^2 L}{2\pi(4\pi L)^2} \exp\,(ik_o \mid \xi \mid) \int_0^\infty dx\; xS(x)\left(\frac{\pi}{2}\right)^{\frac{1}{2}}\left(\frac{2k_o}{x^2 \mid \xi \mid}\right)^{\frac{1}{2}} \cdot\left[\hat{C}\left(\frac{x^2 \mid \xi \mid}{2k_o}\right)^{\frac{1}{2}} - i\hat{S}\left(\frac{x^2 \mid \xi \mid}{2k_o}\right)^{\frac{1}{2}}\right]. \qquad (85)$$

5.2 *Plane-Wave Case*

For the spherical-wave case, the spatial covariance $\langle u_i(\eta + \xi/2)u_i^*(\eta - \xi/2)\rangle$ is given by (77) and its transverse and longitudinal specializations (80) and (81) or (83) and (85). By way of contrast, we derive the corresponding expressions for the plane-wave case.

Approximations (45) and (48) show that the covariance (40) is approximately

$$\left\langle u_i\left(\eta + \frac{\xi}{2}\right)u_i^*\left(\eta - \frac{\xi}{2}\right)\right\rangle \approx 4\epsilon^2 k_o^4 \int dy\, \frac{\exp\,[ik_s(y)\cdot\xi]}{(4\pi)^2 \mid \eta - y \mid^2}\, S(\mid k(y) - k_s(y) \mid) \qquad (86)$$

where $k(y)$ is a constant wavevector, $\mid k(y) \mid = k_o$ in keeping with the interchange of u_e and u_o, and $\mid u_o(y) \mid^2 = 1$. Here the integral is over the volume of a half-space with k perpendicular to the face.

With spherical coordinates centered at the receiving center η, (86) becomes

$$\frac{\epsilon^2 k_o^4}{(2\pi)^2} \int d\Omega \exp\,(ik_s\cdot\xi) S(\mid k - k_s \mid) \int dR \qquad (87)$$

where the radial integral has k_s-dependent integration limits corresponding to the half-space interface. Under the large-scale approximation, $\lambda/l_o \ll 1$, the radial integral is approximately L. For transverse separation, (87) is

$$\frac{\epsilon^2 k_o^4 L}{(2\pi)^2} \int_0^\pi \int_0^{2\pi} d\theta\, d\varphi \sin\theta \exp\,(-ik_o \mid \xi \mid \sin\theta \sin\varphi) S\left(2k_o \sin\frac{\theta}{2}\right) \qquad (88)$$

or

$$\frac{\epsilon^2 k_o^4 L}{2\pi} \int_0^\pi d\theta \sin\theta J_o(k_o \mid \xi \mid \sin\theta) S\left(2k_o \sin\frac{\theta}{2}\right). \qquad (89)$$

Under the large-scale approximation, with small angles yielding the significant part of (89) the covariance becomes

$$\frac{\epsilon^2 k_o^2 L}{2\pi} \int_0^\infty dx\, xS(x) J_o(x \mid \xi \mid). \tag{90}$$

This expression was obtained by Tatarski [Ref. 4, equation (7.64)] to be equal to twice the correlation function for either the log-amplitude or the phase fluctuation of the total field. But

$$S(x) = \frac{4\pi}{x} \int_0^\infty dr\, r\Gamma(r) \sin xr, \tag{91}$$

because (91) is a function of $\mid \kappa \mid$, and

$$\int_0^\infty dx \sin (xr) J_o(x \mid \xi \mid) = \begin{cases} (r^2 - \mid \xi \mid^2)^{-1/2}, & r^2 > \mid \xi \mid^2, \\ 0 \qquad , & r^2 < \mid \xi \mid^2. \end{cases} \tag{92}$$

Substituting (91) and (92) into (90) and changing the variable of integration shows that

$$\left\langle u_i\left(\eta + \frac{\xi}{2}\right) u_i^*\left(\eta - \frac{\xi}{2}\right)\right\rangle \approx 2\epsilon^2 k_o^2 L \int_0^\infty dr\, \Gamma(r^2 + \mid \xi \mid^2)^{\frac{1}{2}} \tag{93}$$

for transverse separation. This is a central result of much of the literature (for example, Ref. 3); we have obtained this result in a simple and novel way.

For the case of longitudinal separation, (87) is

$$\frac{\epsilon^2 k_o^4 L}{(2\pi)^2} \int_0^\pi \int_0^{2\pi} d\theta\, d\varphi \sin \theta \exp (ik_o \mid \xi \mid \cos \theta) S\left(2k_o \sin \frac{\theta}{2}\right) \tag{94}$$

or

$$\frac{\epsilon^2 k_o^4 L}{2\pi} \int_0^\pi d\theta\, 2 \sin \frac{\theta}{2} \cos \frac{\theta}{2} \exp \left[ik_o \mid \xi \mid \left(1 - 2 \sin^2 \frac{\theta}{2}\right)\right] S\left(2k_o \sin \frac{\theta}{2}\right). \tag{95}$$

Under the large-scale approximation, the upper limit of the variable of integration $\kappa = 2k_o \sin \theta/2$ can be replaced by infinity. Thus,

$$\left\langle u_i\left(\eta + \frac{\xi}{2}\right) u_i^*\left(\eta - \frac{\xi}{2}\right)\right\rangle \approx \frac{\epsilon^2 k_o^2 L}{2\pi} \exp (ik_o \mid \xi \mid) \int_0^\infty dx\, xS(x) \exp \left(-i \frac{\mid \xi \mid x^2}{2k_o}\right). \tag{96}$$

It does not appear possible to simplify (96) by using (91) together with the sine-transform of the exponential in (96). However, the special case

$$\Gamma(r) = \exp(-r^2/2l_o^2) \tag{97}$$

is of interest. Then

$$S(x) = (2\pi)^{3/2} l_o^3 \exp(-l_o^2 x^2/2). \tag{98}$$

Inserting (98) into (96) and using the variable $u = l_o^2 x^2/2$ for integration yields

$$\left\langle u_i\left(\eta + \frac{\xi}{2}\right) u_i^*\left(\eta - \frac{\xi}{2}\right)\right\rangle \approx \epsilon^2 k_o^2 L l_o (2\pi)^{\frac{1}{2}} \exp(ik_o \mid \xi \mid) \frac{1 - i\dfrac{\mid \xi \mid}{k_o l_o^2}}{1 + \left(\dfrac{\mid \xi \mid}{k_o l_o^2}\right)^2}. \tag{99}$$

This expression corresponds to a result of Chernov [Ref. 1, p. 94, equation (187)] for longitudinal log-amplitude or phase fluctuations. The magnitude of the last factor in (99) is reduced by $5^{\frac{1}{2}}$ when $\mid \xi \mid = 2k_o l_o^2$.

5.3 *Comparison of the Spherical and Planar Cases*

Notice that the relative variance (zero receiver separation) is the same for the spherical and planar cases. That is to say, expressions (80) and (81) yield the same variance, relative to the spherical-wave power $(4\pi L)^{-2}$, as do expressions (90) and (96), relative to the unity plane-wave power.

For transverse receiver separation, the planar-case result (90) can be compared with the spherical-case result (80). The weighting of the spectral function $xS(x)$ is $j_o(x \mid \xi \mid)$ in the planar case; this $\mid \xi \mid$-function has its first zero at $\mid \xi \mid = 2.4/x$ with subsequent zeros spaced $3.1/x$ apart. In the spherical case, the weighting is

$$x^{-1} \int_0^x d\kappa \, J_o(\kappa \mid \xi \mid);$$

this $\mid \xi \mid$-function is a mixture of functions with "periodicity" larger than that of $J_o(x \mid \xi \mid)$. Presumably, correlation lengths would usually be larger in the spherical case than in the planar case.

For longitudinal receiver separation, the planar-case result (96) can be compared with the spherical-case result (81). The weighting of $xS(x)$ is

$$\exp\left(-i\frac{\mid \xi \mid x^2}{2k_o}\right)$$

in the planar case; this $|\xi|$-function has period $4\pi k_o/x^2$. In the spherical case, the weighting is

$$x^{-1}\int_0^x d\kappa \exp\left(-i\frac{|\xi|\kappa^2}{2k_o}\right);$$

this $|\xi|$-function is a mixture of longer-period functions, and again correlation lengths would presumably be larger in the spherical case than in the planar case.

Physical reasoning also suggests that correlation lengths are larger in the spherical case than in the planar case. First, compare the regions of important scattering. For the spherical case, this region is sketched in Fig. 5, where the angle Ψ is given by (65). For the planar case, this region is a cone with half-angle Ψ and axial length L (the transmitter-receiver separation being replaced by the distance the receiver is imbedded into a half-space of random refractive index), Fig. 8. Comparison of the two regions suggests the the fluctuation field in the spherical case is more directive than the fluctuation field in the planar case.

Second, consider the implication of a more directive fluctuation field. The directionality function N can be defined by

$$\frac{\left\langle u_i\left(\eta+\frac{\xi}{2}\right)u_i^*\left(\eta-\frac{\xi}{2}\right)\right\rangle}{\langle|u_i(\eta)|^2\rangle} = \int d\Omega \exp(ik_s\cdot\xi)N(k_s). \qquad (100)$$

A wave in direction k_s contributes a correlation $\exp(ik_s\cdot\xi)$, and the total correlation is a weighted average of such constituents. The form (100) is exhibited by (70) in the spherical case and by (87) in the planar case. An idealized directionality function would be constant with $k_s\cdot\eta$ above a threshold and would be zero elsewhere. That is to say, (100) would be

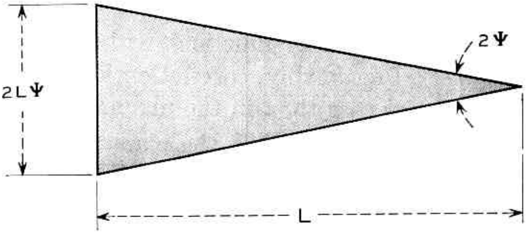


Fig. 8 — Region of important scattering (plane waves).

$$\frac{1}{\Delta\Omega}\int_{\Delta\Omega} d\Omega \exp\left(i\mathbf{k}_s\cdot\xi\right) \tag{101}$$

where $\Delta\Omega$ is a small cap on the unit sphere of size $2\pi(1 - \cos\Theta) \approx \pi\Theta^2$. When ξ is transverse, (101) becomes

$$[2\pi(1 - \cos\Theta)]^{-1}\int_0^{\Theta}\int_0^{2\pi} d\theta\, d\varphi \sin\theta \exp\left(-ik_o \mid \xi \mid \sin\theta \sin\varphi\right) \tag{102}$$

$$(1 - \cos\Theta)^{-1}\int_0^{\Theta} d\theta \sin\theta J_o(k_o \mid \xi \mid \sin\theta).$$

Under the small-angle approximation, (102) yields

$$\frac{2J_1(k_o \mid \xi \mid \Theta)}{k_o \mid \xi \mid \Theta}. \tag{103}$$

The correlation function (103) is unity at $\mid \xi \mid = 0$, is 0.88 at $\mid \xi \mid = \lambda/2\pi\Theta \approx 0.16\ \lambda/\Theta$, and is zero at $\mid \xi \mid \approx 0.61\ \lambda/\Theta$.

When ξ is longitudinal, (101) becomes

$$[2\pi(1 - \cos\Theta)]^{-1}\int_0^{\Theta}\int_0^{2\pi} d\theta\, d\varphi \sin\theta \exp\left(ik_o \mid \xi \mid \cos\theta\right)$$

$$= \exp\left(ik_o \mid \xi \mid\right)\frac{1 - \exp\left[ik_o \mid \xi \mid (1 - \cos\theta)\right]}{ik_o \mid \xi \mid (1 - \cos\Theta)} \tag{104}$$

$$= \exp\left(ik_o \mid \xi \mid\right)\frac{1 - \exp\left[-i2k_o \mid \xi \mid \Delta\Omega/4\pi\right]}{i2k_o \mid \xi \mid \Delta\Omega/4\pi}. \tag{105}$$

The correlation $\exp\left(ik_o \mid \xi \mid\right)$ associated with a plane wave is modulated by a function having ripple in its numerator with period $\lambda(2\pi/\Delta\Omega)$.

The λ-dependence exhibited in (103) and (107) must be tempered by the λ-dependence of the angle Θ. Figures 5 and 8 suggest that Θ would be at most λ/l_o ($\Delta\Omega$ at most $\pi\lambda^2/l_o^2$) for equivalence of the idealized and true directionality functions. For the transverse case, the null of (103) would be at $\mid \xi \mid = 0.61\ l_o$ or more; the λ-dependence disappears as in (80) and (90). For the longitudinal case, the period in (105) would be at least $2l_o^2/\lambda$; this period is to be compared with the width of the last factor in (99) for the plane-wave gaussian-index correlation case.

Both (103) and (105) exhibit the fact that correlation lengths are inversely proportional to $\Delta\Omega$, the width of the directionality function. But the scattering volumes depicted in Figs. 5 and 8 show that this width is smaller in the spherical case than in the planar case. This physical reasoning corroborates the previous interpretation of the integration formulas which showed larger correlation lengths for the spherical case.

VI. EXAMPLES OF TRANSVERSE COVARIANCES

It has been shown that, for transverse receiver separation, the covariance $\langle u_i(\eta + \xi/2)u_i^*(\eta - \xi/2)\rangle$ is given by (80) in the spherical case and by (90) in the planar case. These expressions are now evaluated in closed form for two illustrative spectra.

Recall that the spectrum $S(\kappa)$ is related to the refractive index covariance $\Gamma(r)$ by (52) which becomes (91) for the statistically isotropic case. For convenience, (91) is repeated here, together with its inverse:

$$S(\kappa) = \frac{4\pi}{\kappa}\int_0^\infty dr\, r\Gamma(r)\sin\kappa r \tag{106}$$

$$\Gamma(r) = \frac{1}{2\pi^2 r}\int_0^\infty d\kappa\, \kappa S(\kappa)\sin\kappa r. \tag{107}$$

Also, the planar-case covariance (90) is

$$\frac{\epsilon^2 k_o^2 L}{2\pi}\int_0^\infty d\kappa\, \kappa S(\kappa) J_o(\kappa \mid \xi \mid), \tag{108}$$

and the spherical-case covariance (80) is

$$\frac{\epsilon^2 k_o^2 L}{(4\pi L)^2 2\pi}\int_0^\infty d\kappa\; J_o(\kappa \mid \xi \mid)\int_\kappa^\infty dx\; S(x); \tag{109}$$

the order of integration has been changed.

The normalization of the spectrum follows from (107) evaluated at $r = 0$,

$$1 = \frac{1}{2\pi^2}\int_0^\infty d\kappa\, \kappa^2 S(\kappa), \tag{110}$$

so that ϵ^2 plays the role of the variance of the refractive index.

Our first example is the case of an exponential spectrum:

$$S(\kappa) = \pi^2\Lambda^3 \exp\left[-\Lambda\kappa\right], \tag{111}$$

$$\Gamma(r) = [1 + (r/\Lambda)^2]^{-2} \tag{112}$$

where $\Lambda \gg \lambda$. Then, the planar covariance (108) is

$$\frac{\epsilon^2 k_o^2 L \pi^2 \Lambda}{2\pi}[1 + (\mid \xi \mid/\Lambda)^2]^{-3/2} \tag{113}$$

and the spherical covariance (109) is

$$\frac{\epsilon^2 k_o^2 L\pi^2 \Lambda}{(4\pi L)^2 2\pi}\,[1 + (|\,\xi\,|/\Lambda)^2]^{-1/2}. \tag{114}$$

The respective correlation functions in (112) to (114) are plotted in Fig. 9. The correlation length for the spherical case is larger than the comparable correlation lengths for the planar case and for the refractive index.

Our second example is the case of a constant spectrum:

$$S(\kappa) = \begin{cases} \dfrac{3\Lambda^3}{4\pi}, & 0 \leqq \kappa \leqq 2\pi/\Lambda \\ 0, & \kappa > 2\pi/\Lambda \end{cases} \tag{115}$$

$$\Gamma(r) = \frac{\sin\dfrac{2\pi r}{\Lambda} - \dfrac{2\pi r}{\Lambda}\cos\dfrac{2\pi r}{\Lambda}}{\dfrac{1}{3}\left(\dfrac{2\pi r}{\Lambda}\right)^3}. \tag{116}$$

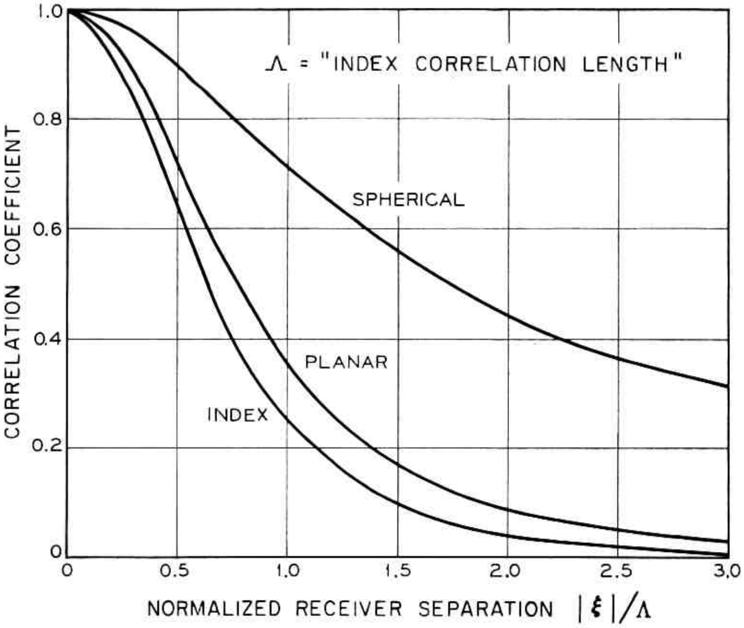


Fig. 9 — Correlations for the exponential spectrum.

Then, the planar covariance (108) becomes

$$\frac{\epsilon^2 k_o^2 L}{2\pi} \frac{3\pi\Lambda}{2} \left[\frac{2J_1\left(\frac{2\pi \mid \xi \mid}{\Lambda}\right)}{2\pi \mid \xi \mid / \Lambda} \right]. \tag{117}$$

The correlation function in brackets agrees with (103) with $\Theta = \lambda/\Lambda$, which determines the angular extent of the constant directionality function. The spherical covariance (109) is

$$\frac{\epsilon^2 k_o^2 L}{(4\pi L)^2} \frac{3\pi\Lambda}{2} \left\{ 2J_o\left(\frac{2\pi \mid \xi \mid}{\Lambda}\right) - \frac{2J_1\left(\frac{2\pi \mid \xi \mid}{\Lambda}\right)}{2\pi \mid \xi \mid / \Lambda} \right.$$
$$\left. + \pi J_1\left(\frac{2\pi \mid \xi \mid}{\Lambda}\right) H_o\left(\frac{2\pi \mid \xi \mid}{\Lambda}\right) - \pi J_o\left(\frac{2\pi \mid \xi \mid}{\Lambda}\right) H_1\left(\frac{2\pi \mid \xi \mid}{\Lambda}\right) \right\}, \tag{118}$$

where H_ν are Struve functions. The correlation functions in (116) to (118) are plotted in Fig. 10. As before, the correlation length for the spherical case is larger than the comparable correlation lengths for the planar case and for the refractive index.

VII. SUMMARY

In Section II, the perturbation theory of Keller is developed in a novel way.[2] This development shows that the nearly uncoupled equations (12) and (13) arise naturally from the fundamental pair (3) and (4). In Section III, the equation for the average field (12) is solved for the case in which the refractive index is statistically homogeneous and isotropic. The spherical wave (16) is shown to be a good local approximation of the average field; the wavenumber k is a weak function of position and satisfies (21). Beyond a few correlation lengths from the source, the wavenumber is a constant given approximately by (28) to (30).

In Section IV, the equation for the fluctuation field (13) is shown to imply (40) for the spatial covariance of the field. A useful approximation of the covariance is (51) which is then justified on physical grounds. For the large-scale case $\lambda \ll l_o$, where λ is the wavelength and l_o is the inner scale for the refractive index, this approximation shows that the region of important scattering is given by (60) with (65) and lies within the volume common to two cones with apexes at transmitter and receiver, each cone with half-angle approximately λ/l_o. The interval for validity of (51) is given by (68) which states that the transmitter-to-receiver distance is sufficiently large for far-field covariance approxima-

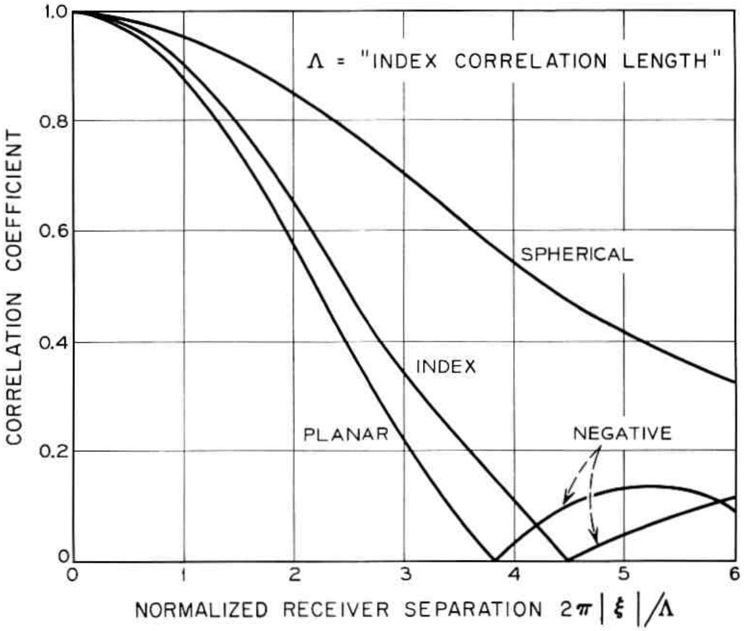


Fig. 10 — Correlations for the constant spectrum.

tions to be valid but is sufficiently small for single scatter perturbation approximations to be valid. In Appendix B, it is shown that a far-field condition relative to the covariance is less restrictive than a far-field condition relative to the field itself.

In Section V, the volume integration of (51) for the covariance is transformed to the angular integrations exhibited in (77). For the large-scale case, the covariance is given by (80) and (81) for transverse and longitudinal separation of receivers. These expressions for our spherical-wave model are contrasted with (90) and (96) for the plane-wave model, showing that relative variances are the same but that correlation lengths are larger in the spherical-wave case than in the plane-wave case. This is to be expected on physical grounds, for comparison of the volumes of important scattering for the two cases indicates that the fluctuation field is more directive in the spherical case. But a more directive field has longer correlation lengths; this is illustrated by (103) and (105) for transverse and longitudinal separations under the idealized directionality function in (101). In Section VI, two special cases of refractive-index correlation which correspond to an exponential spectrum and a constant spectrum are considered. For transverse separation, the covariance functions are derived in closed form. Plots of the correlation functions show that correlation lengths for the spherical wave case are larger than the plane-wave correlation lengths, which are comparable to the correlation lengths of the refractive index.

VIII. ACKNOWLEDGMENT

Discussions with B. H. Bharucha, T. L. Ho, J. N. Lahti, and D. C. Stickler are gratefully acknowledged.

APPENDIX A

A Mean-Value Theorem

We show that any solution of

$$[\nabla^2 + k^2]u(r) = -\delta(r) \tag{119}$$

satisfies the mean-value relation

$$\frac{1}{4\pi R^2}\int dS\, u(r+\rho) = \begin{cases} u(r)\dfrac{\sin kR}{kR}, & 0 < R < |r|, \\ u(r)\dfrac{\sin kR}{kR} + \dfrac{\sin k(|r| - R)}{4\pi |r| kR}, & R > |r|, \end{cases} \tag{120}$$

where the integration is over the surface of the sphere $\{\rho : |\rho| = R\}$. In particular, when $u(r)$ is of the form (16), then (120) becomes (19).

Introduce a function $\psi(r)$ that satisfies

$$[\nabla^2 + k^2]\psi(r) = -\delta(r - r_o). \tag{121}$$

Then (119) and (121) imply that

$$\nabla\cdot(\psi\nabla u - u\nabla\psi) = u\delta(r - r_o) - \psi\delta(r). \tag{122}$$

For the sphere $\{r : |r - r_o| = R\}$ with the outward unit normal $\hat{\rho} = (r - r_o)/|r - r_o|$, the divergence theorem yields

$$\int dS\,(\psi\hat{\rho}\cdot\nabla u - u\hat{\rho}\cdot\nabla\psi) = \begin{cases} u(r_o), & 0 < R < |r_o|, \\ u(r_o) - \psi(0), & R > |r_o|. \end{cases} \tag{123}$$

We choose ψ to be a linear combination of

$$\frac{\exp\,(\pm ik\,|\,r - r_o\,|)}{4\pi\,|\,r - r_o\,|} \tag{124}$$

such that ψ is zero on the surface of the sphere and satisfies (121). This choice is

$$\psi(r) = -\frac{\sin k(|\,r - r_o\,| - R)}{4\pi\,|\,r - r_o\,|\sin kR}. \tag{125}$$

The radial component of the gradient of (125) evaluated on the surface of the sphere is

$$\hat{\rho}\cdot\nabla\psi = -\frac{kR}{4\pi R^2\sin kR}. \tag{126}$$

Then (125) and (126) in conjunction with (123) yield (120).

APPENDIX B

An Improved Far-Field Condition

The kernel (41) used in the integral (40), yielding the covariance, is

$$G(r, r')G^*(\rho, \rho') = \frac{\exp\,[ik_o(|\,r - r'\,| - |\,\rho - \rho'\,|)]}{(4\pi)^2\,|\,r - r'\,|\,|\,\rho - \rho'\,|}. \tag{127}$$

With the definitions and inverse relations (36) to (39),

$$\begin{aligned} \eta &= \frac{r + \rho}{2}, \qquad & \xi &= r - \rho, \\ y &= \frac{r' + \rho'}{2}, \qquad & x &= r' - \rho', \end{aligned} \tag{128}$$

$$\begin{aligned} r &= \eta + \frac{\xi}{2} \qquad & \rho &= \eta - \frac{\xi}{2}, \\ r' &= y + \frac{x}{2}, \qquad & \rho' &= y - \frac{x}{2} \end{aligned} \tag{129}$$

the kernel is

$$\frac{\exp\,[ik_o(|\,\eta - y + \tfrac{1}{2}(\xi - x)| - |\,\eta - y - \tfrac{1}{2}(\xi - x)|)]}{(4\pi)^2\,|\,\eta - y + \tfrac{1}{2}(\xi - x)|\,|\,\eta - y - \tfrac{1}{2}(\xi - x)|}. \tag{130}$$

The far-field (Fraunhofer) approximation arises from the series expansions (42)

$$\begin{aligned} | \eta - y + \tfrac{1}{2}(\xi - x)| &= | \eta - y | + \frac{1}{2} \frac{\eta - y}{|\eta - y|} \cdot (\xi - x) + \cdots , \\ | \eta - y - \tfrac{1}{2}(\xi - x)| &= | \eta - y | - \frac{1}{2} \frac{\eta - y}{| \eta - y |} \cdot (\xi - x) + \cdots , \end{aligned} \tag{131}$$

and the approximation kernel (44) is

$$\frac{\exp ik_o \left[\dfrac{\eta - y}{| \eta - y |} \cdot (\xi - x) \right]}{(4\pi)^2 \, | \eta - y |^2}. \tag{132}$$

A usual condition for the validity of a far-field approximation is

$$| \eta - y | \gg \frac{L_o^2}{\lambda}, \tag{133}$$

where L_o is an outer scale of the scatering medium. This condition is relative to approximation of the field. But relative to approximation of the covariance of the field, the condition of validity is

$$| \eta - y | \gg L_o \left(\frac{L_o}{\lambda} \right)^{\frac{1}{3}}. \tag{134}$$

In the case $L_o \gg \lambda$, condition (134) is considerably less restrictive than condition (133). The reason for this improved state of affairs is that, rather than approximating Green's function G, we are approximating the kernal GG^*. In the computation of the phase of this kernel with expansion (131), there is cancellation of terms that ordinarily remain when computing the phase of G itself. Overbounding the effect of all neglected terms, not just the first one, leads to condition (134).

Our task is to approximate the phase in (130), namely, the argument of the exponential. We put

$$Y = \eta - y, \qquad X = \xi - x, \tag{135}$$

so that

$$\begin{aligned} | r - r' | &= | Y + \tfrac{1}{2}X |, \\ | \rho - \rho' | &= | Y - \tfrac{1}{2}X |. \end{aligned} \tag{136}$$

Then, we observe

$$| Y \pm \tfrac{1}{2} | = | Y | \left(1 \pm \frac{Y \cdot X}{Y^2} + \frac{X^2}{4Y^2} \right)^{\frac{1}{2}}. \tag{137}$$

Next we put

$$\alpha = \frac{Y\cdot X}{Y^2}, \qquad \beta = \frac{X^2}{4Y^2}, \tag{138}$$

so that

$$\begin{aligned} |\, r - r' \,| &= |\, Y \,| (1 + \alpha + \beta)^{\frac{1}{2}}, \\ |\, \rho - \rho' \,| &= |\, Y \,| (1 - \alpha + \beta)^{\frac{1}{2}}. \end{aligned} \tag{139}$$

The next step is to assume $|\pm \alpha + \beta| < 1$ and expand $|r - r'|$ and $|\rho - \rho'|$ with a binomial series. Then,

$$\begin{aligned} (1 \pm \alpha + \beta)^{\frac{1}{2}} = 1 + \tfrac{1}{2}(\pm\alpha + \beta) - \frac{1\cdot1}{2\cdot4}(\pm\alpha + \beta)^2 \\ + \frac{1\cdot1\cdot3}{2\cdot4\cdot6}(\pm\alpha + \beta)^3 - \frac{1\cdot1\cdot3\cdot5}{2\cdot4\cdot6\cdot8}(\pm\alpha + \beta)^4 + \cdots . \end{aligned} \tag{140}$$

In the above expression, only terms with differing signs contributed to the difference $|r - r'| - |\rho - \rho'|$. Thus,

$$\begin{aligned} &(1 + \alpha + \beta)^{\frac{1}{2}} - (1 - \alpha + \beta)^{\frac{1}{2}} \\ &\qquad = \alpha - 2\,\frac{1\cdot1}{2\cdot4}\,2\alpha\beta + 2\,\frac{1\cdot1\cdot3}{2\cdot4\cdot6}\,\alpha^3 + 2\,\frac{1\cdot1\cdot3}{2\cdot4\cdot6}\,3\alpha\beta^2 + R, \end{aligned} \tag{141}$$

where the remainder R has the series expansion

$$\begin{aligned} R = -\frac{1\cdot1\cdot3\cdot5}{2\cdot4\cdot6\cdot8}\,[(\alpha + \beta)^4 - (-\alpha + \beta)^4] \\ + \frac{1\cdot1\cdot3\cdot5\cdot7}{2\cdot4\cdot6\cdot8\cdot10}\,[(\alpha + \beta)^5 - (-\alpha + \beta)^5] - \cdots . \end{aligned} \tag{142}$$

The series is readily "majorized," with the result

$$\begin{aligned} |\, R \,| < 2 \,\Bigg|\, \frac{1\cdot1\cdot3\cdot5}{2\cdot4\cdot6\cdot8}\,(|\,\alpha\,| + |\,\beta\,|)^4 \\ + \frac{1\cdot1\cdot3\cdot5\cdot7}{2\cdot4\cdot6\cdot8\cdot10}\,(|\,\alpha\,| + |\,\beta\,|)^5 + \cdots \Bigg|, \end{aligned} \tag{143}$$

or, with $\gamma = |\,\alpha\,| + |\,\beta\,|$ and $0 < \gamma < 1$,

$$\frac{|\,R\,|}{2} < \frac{5}{128}\,\gamma^4\left[1 + \frac{7}{10}\,\gamma + \frac{7\cdot9}{10\cdot12}\,\gamma^2 + \cdots\right]. \tag{144}$$

But the series in brackets is overbounded by

$$\frac{1}{1-\gamma} = 1 + \gamma + \gamma^2 + \cdots . \tag{145}$$

Thus, $\gamma < \frac{1}{2}$ implies

$$| R | < \frac{5}{32}\gamma^4 = \frac{5}{32}(| \alpha | + | \beta |)^4. \tag{146}$$

The above calculations show that

$$| r - r' | - | \rho - \rho' | = | Y | \left[\frac{Y\cdot X}{Y^2} - \frac{1}{8}\frac{Y\cdot X}{Y^2}\frac{X^2}{Y^2} + \frac{1}{8}\left(\frac{Y\cdot X}{Y^2}\right)^3 + \frac{3}{128}\frac{Y\cdot X}{Y^2}\left(\frac{X^2}{Y^2}\right)^2 + R \right], \tag{147}$$

where

$$| R | < \frac{5}{32}\left(\left| \frac{Y\cdot X}{Y^2} \right| + \frac{X^2}{4Y^2} \right)^4. \tag{148}$$

The conditions

$$\left| \pm \frac{Y\cdot X}{Y^2} + \frac{X^2}{4Y^2} \right| < 1 \tag{149}$$

and

$$\left| \frac{Y\cdot X}{Y^2} \right| + \frac{X^2}{4Y^2} < \frac{1}{2} \tag{150}$$

have been imposed. Since condition (150) implies (149) it is clear that both are met when

$$\frac{| X |}{| Y |} < \frac{1}{3}. \tag{151}$$

The far-field approximation of the kernel employs the leading term of (147); that is to say, the Fraunhofer phase is

$$k_o \frac{Y}{| Y |}\cdot X = \frac{2\pi}{\lambda}\frac{\eta - y}{| \eta - y |}\cdot(\xi - x). \tag{152}$$

The phase error is then

$$-k_o\left\{ \frac{| X |}{8}\cdot\frac{Y\cdot X}{| Y | | X |}\cdot\frac{X^2}{Y^2} \cdot\left[1 - \left(\frac{Y\cdot X}{| Y | | X |}\right)^2 - \frac{3}{16}\frac{X^2}{Y^2} \right] - | Y | R \right\}. \tag{153}$$

Our task now is to overbound the magnitude of the phase error. We impose $| X | \leqq L_o$, an outer scale of the refractive-index correlation function Γ. This is appropriate for zero receiver separation, $| \xi | = 0$; later, L_o could be replaced by $| \xi | + L_o$ for nonzero receiver separation. Apart from the remainder, (153) is seen to be a function of the cosine $u \cdot v$, with $u = X/| X |$ and $v = Y/| Y |$, and of the ratio $w = | X |/| Y |$. We overbound the product in (153) by a product of overbounds in which $u \cdot v$ has distinct values, and overbound the second factor by unity (unity is greater than both the largest positive value $1 - \frac{3}{16}w^2$ and the largest magnitude, $\frac{3}{16}w^2$, of negative values, since $w < 1$). It follows that the magnitude of the phase error is overbounded by

$$k_o\left(\frac{L_o^3}{8Y^2} + | Y | | R |\right). \tag{154}$$

We now overbound our previous estimate of $| R |$. We assume $| Y | > L_o$, even though tighter overbounds can be obtained under $| Y | > 3L_o$ as previously assumed. Then, (148) yields

$$| R | < \frac{5}{32}\left(\frac{L_o}{| y |} + \frac{L_o^2}{4Y^2}\right)^4 < \frac{5}{32}\left(\frac{5L_o}{4 | Y |}\right)^4, \tag{155}$$

or

$$| R | < \frac{1}{10}\left(\frac{25}{16}\right)^3\left(\frac{L_o}{| Y |}\right)^4 < 0.382\left(\frac{L_o}{| Y |}\right)^4. \tag{156}$$

We use

$$| R | < \frac{2}{5}\left(\frac{L_o}{| Y |}\right)^4. \tag{157}$$

The above phase-error bound (154) is less than

$$k_o \frac{L_o^3}{8Y^2}\left(1 + \frac{16}{5}\frac{L_o}{| Y |}\right), \tag{158}$$

which is less then

$$\frac{k_o L_o^3}{8Y^2} \cdot \frac{21}{5} < \frac{5}{8}\frac{k_o L_o^3}{Y^2}. \tag{159}$$

All of the above calculations required $| Y | > L_o$. With the stronger condition $| Y | > 4L_o$, the next -to-last calculation becomes

$$k_o \frac{L_o^3}{8Y^2}\left(1 + \frac{16}{5}\frac{L_o}{| Y |}\right) < \frac{k_o L_o^3}{8Y^2}\frac{9}{5}. \tag{160}$$

This overbound is less than

$$\frac{k_o L_o^3}{4 Y^2}. \tag{161}$$

Suppose we impose the condition that the phase error be less than $\pi/32$ and we ask what value of $| Y |$, the scattering range, is required. The last overbound yields

$$| \eta - y | = | Y | > 4\left(\frac{L_o}{\lambda}\right)^{\frac{1}{2}} L_o \, . \tag{162}$$

This is a less restrictive condition than one specifying the far-field range to be much greater than L_o^2/λ. When $L_o \gg \lambda$, it is a considerably weaker condition. On the other hand, with $L_o > \lambda$ (but now $L_o \approx \lambda$), the condition (162) still implies the assumption $| Y | < 4L_o$ under which it was obtained. When $L_o < \lambda$, the results are valid but vacuous, since the pertinent condition is then $| Y | > 4L_o$.

We turn to the approximation of the coherent-field function (46),

$$u\left(y + \frac{x}{2}\right) u_c^*\left(y - \frac{x}{2}\right) = \frac{\exp\left[i\left(k \left| y + \frac{x}{2} \right| - k^* \left| y - \frac{x}{2} \right|\right)\right]}{(4\pi)^2 \left| y + \frac{x}{2} \right| \left| y - \frac{x}{2} \right|}, \tag{163}$$

in which k is a weak function of position. With the identification $y = Y$, $x = X$, the previous analysis is applicable. The approximation (48) to (50) is

$$\frac{\exp\left(-2 \operatorname{Im} k \, | y |\right) \exp\left[i(\operatorname{Re} k) \dfrac{y}{| y |} \cdot x\right]}{(4\pi)^2 \, | y |^2}, \tag{164}$$

with a phase error less than

$$\frac{\operatorname{Re} k L_o^3}{4 \, | y |^2}. \tag{165}$$

In view of the interchangeability of k and k_o, the phase error (165) is comparable to

$$\frac{k_o L_o^3}{4 \, | y |^2}. \tag{166}$$

Then, a phase error in (164) less than $\pi/32$ requires that the source-to-scattering distance $| y |$ satisfy

$$| y | > 4L_o(L_o/\lambda)^{\frac{1}{2}}. \tag{167}$$

Our calculation is similar to one by Lahti and Ishimaru, but the calculation (and result) is simpler and the final conditions are less restrictive.[10] Simplicity is obtained in part because we use variables with symmetric form, (128) and (129), and we overbound a quartic remainder rather than modify a cubic remainder. Our quartic-remainder overbound also yields less restrictive conditions, say $| Y | > 4L_o(L_o/\lambda)^{\frac{1}{2}}$ implying a phase error less than $\pi/32$ in comparison with $| Y | > 7L_o(L_o/\lambda)^{\frac{1}{2}}$ yielding an error less than $\pi/10$.

REFERENCES

1. Chernov, L. A., *Wave Propagation in a Random Medium,* New York: McGraw-Hill Book Company, Inc., 1960.
2. Keller, J. B., "Stochastic Equations and Wave Propagation in Random Media," Proc. Symp. Appl. Math., *16,* Amer. Math. Soc., Providence, R. I., 1964.
3. Taylor, L. S., "Decay of Mutual Coherence in Turbulent Media," J. Opt. Soc. Amer., *57,* No. 3 (March 1967), pp. 304–308.
4. Tatarski, V. I., *Wave Propagation in a Turbulent Medium,* New York: McGraw-Hill, 1961.
5. Fried, D. L., "Propagation of a Spherical Wave in a Turbulent Medium," J. Opt. Soc. Amer., *57,* No. 2 (February 1967), pp. 175–180.
6. Bergmann, P. G., "Propagation of Radiation in a Medium with Random Inhomogeneities," Phys. Rev., *70,* Nos. 7 and 8 (October 1946), pp. 486–492.
7. Mintzer, D., "Wave Propagation in a Randomly Inhomogeneous Medium," J. Acoust. Soc. Amer., part I, *25,* No. 5 (September 1953), pp. 922–927; part II, *25,* No. 6 (November 1953), pp. 1107–1111; and part III, *26,* No. 2 (March 1954), pp. 186–190.
8. Beran, M. J., "Propagation of a Spherically Symmetric Mutual Coherence Function Through a Random Medium," IEEE Trans. on Antennas and Propagation, *AP-15,* No. 1 (January 1967), pp. 66–69.
9. Wheelon, A. D., "Radio-Wave Scattering by Tropospheric Irregularities," J. Res. Nat. Bureau of Standards-part D, Radio Propagation, *63D,* No. 2 (September-October 1959), pp. 205-233.
10. Lahti, J. N., and Ishimoru, Akira, unpublished work.

Response of Delta Modulation to Gaussian Signals

By M. R. AARON, J. S. FLEISCHMAN, R. A. McDONALD, and E. N. PROTONOTARIOS

(Manuscript received October 30, 1968)

Analytical, experimental, and computer simulation results are given for the error spectrum of a delta modulator when probed by stationary, band limited, gaussian noise. These three complementary methods are used to increase our quantitative understanding of the nonlinear system with memory. The error is conveniently split into two components: one linearly dependent on the input signal and one linearly independent of the input signal. In order to isolate these two types of errors we use two measurement techniques. For purposes of analysis we show that the delta modulator can be replaced by an equivalent linear system with additive noise at its output which is linearly uncorrelated with the input. The equivalent linear system may be approximated by using methods involving statistical linearization or the deterministic describing function. Alternately, the equivalent linear system may be obtained from computer simulation.

I. INTRODUCTION

1.1 *General Background and Broad Objectives*

Delta modulation (DM) has been known for almost two decades; yet, little has been published comparing experiment with theory particularly for random inputs.† On the surface this might seem strange because of the apparent simplicity of the delta modulation system blocked out in Fig. 1(a). The waveforms depicted in Fig. 1(b) and the mathematical model in Fig. 1(c) should suffice to explain how the system operates. The principal difficulty of the analysis is the absence of general tools for handling random processes in nonlinear systems with memory. From this viewpoint the simplicity of the delta

† The first reference to delta modulation appeared in French patent literature (see Ref. 1) in 1946, but the first readily available description in English appeared in 1952 (see Ref. 2).

modulator is deceptive. However, if we make some inroads into the quantitative understanding of this seemingly simple case, it may give us courage to go on to more complicated situations.

In this paper we concentrate on the development and exploitation of analytical, experimental, and computational techniques to enhance our understanding of the objective performance of delta modulation. We do not consider the correlation of objective measures with subjective effects for applications to either voice or video; rather, our main aim is to correlate what is known in theory, including our own developments with what has been achieved experimentally.

Renewed interest has come from at least two sources. First, differential systems of which delta modulation is the simplest member, have long been known to be well suited to handling signals whose spectra fall off at high frequencies.[3–5] This is particularly true of black and white video; there is substantial interest in transmitting such signals digitally.[6] Interest also has been generated by the desire to produce inexpensive time division switching and transmission systems for voice.[7] In this application, delta modulation is attractive because of its simplicity and great compatibility with the emerging integrated circuit technology.

1.2 *Use of the Random Noise Probe*

Reasons for characterizing a delta modulator with a random noise probe are twofold. First, the envelope of a scanned video signal has a power spectral density that is close to that obtained by passing gaussian noise through an RC filter.† Therefore objective performance measures obtained in response to this signal bear some relationship to subjective performance. Second, use of the established noise-loading procedure for determining the spectrum of the noise in a nonlinear system yields a "signature" that is useful for verifying that the delta modulator is performing as designed. Verification of prescribed performance is an essential prelude to careful subjective testing as well as an absolute necessity for production control. To avoid measurement problems that result from the low spectral density of the RC noise source at the upper end of the band, we deal with flat band-limited white noise almost exclusively.

1.3 *Chronology and Summary*

At the start of our work the signal-to-noise ratios obtained by computer, experiment, and analysis disagreed substantially, partic-

† We use the terms power spectral density, power spectrum, and spectrum interchangeably throughout.

ularly when the signal was changing more rapidly than the delta modulator could follow. In this region, known as the region of slope overload, two methods of computing slope overload noise differed markedly.[6,8] Reference 9 gives a reconciliation of the differences and the development of an analytical expression for the mean square value of the slope overload noise. By using the best features of the previous conflicting theories, an analytical result was obtained that yielded good agreement with computer results.[9] Granular noise, as computed from Van de Weg's approach, agreed with both simulation and experiments.[10] Results obtained by noise-loading experiments continued to disagree with both theory and simulation in the slope overload region. It quickly became apparent that the difference resulted from the fact that this measurement procedure did not measure the noise as defined by theory.

To clarify the differences it is desirable to consider the spectrum of the noise introduced by the delta modulator. Two definitions of noise are possible; the simplest is that the noise is the error, that is, the difference $[x(t) - y(t)]$ between the input signal and the local output signal as defined in Fig. 1. This error is correlated statistically with the input signal. In other words, the error may be considered to be made up of two components, one linearly dependent on the input signal and one linearly independent of the input signal. The linearly dependent component may be regarded as being caused by passing the signal through a noise-free linear filter.

This equivalent linear filter does not introduce noise but merely introduces frequency distortion, as for example in producing selective attenuation and phase shift, particularly for the higher frequency components of the signal that the delta modulator cannot follow. The noise component linearly independent of the signal may be viewed as equivalent to additive uncorrelated noise just as in the case of a nonfeedback type of pulse code modulation quantizer.[11] When the noise is split up this way, the components have distinctly different subjective effects and are thus meaningfully studied separately. In fact the spectrum of the uncorrelated component of the noise is measured by the noise-loading test pictured schematically in Fig. 2. This test procedure is commonly used to test transmission systems, primarily for nonlinear distortion, in this manner.[12] With the switch in the upper position, the colored gaussian noise is passed through a narrowband elimination filter prior to exciting the system under test. At the system output, noise generated by system nonlinearities is measured by the bandpass filter in the receiver of the noise-loading set. This

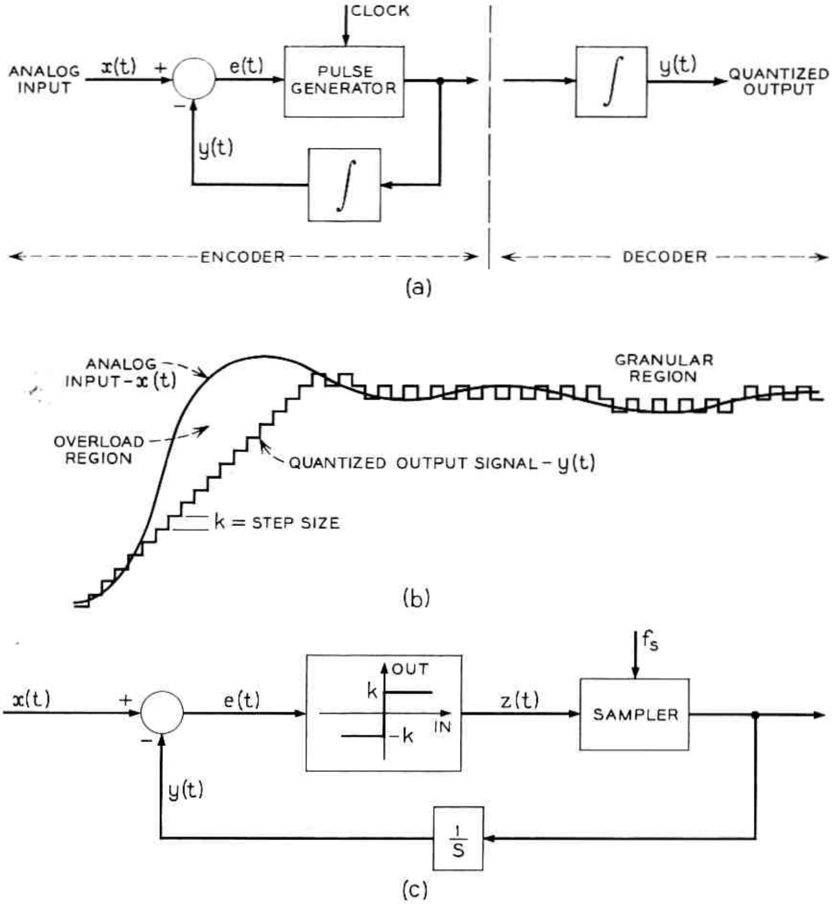


Fig. 1 — Delta modulation: (a) delta modulation system, (b) waveforms, (c) mathematical model.

filter passes only those frequency components eliminated from the input signal. This differs from the total noise as computed by analysis and simulation.

Two approaches were taken to reconcile measurements with the paper and pencil results. First, we used a straight-forward, but tedious, measurement procedure called the cancellation test to measure the total noise as defined by theory, that is, the difference between output and input. The results achieved substantiated the theoretical results. Unfortunately the cancellation or "feed-around" technique, as discussed in detail in Section IV, is tedious and difficult to perform accurately. This made it necessary to rely on the more convenient noise loading measurements. To compare theory and experiment it became necessary to remove the correlated components of the noise as obtained from theory and simulation. It was not possible to do so for the purely theoretical approach, but the results of the simulation were modified, as described in Section III, to agree with the measurement made with the noise-loading technique.

The equivalent linear filter, defined in Section 1.4, cannot be obtained analytically, but it may be determined using computer simulation. An approximate analytical method for arriving at the equivalent linear filter is statistical linearization to replace the quantizer (signum function, threshold circuit) of Fig. 1 with a "suitable" linear gain. This approach is discussed in Section 2.2 where comparisons are made of the equivalent linear filters obtained by the statistical linearization and simulation approaches. Most of the manipulations regarding the statistical linearization are relegated to Appendix A. Section 2.3 is concerned with harmonic analysis useful in its own right as well as a part of the cancellation test. The prelude to the Fourier analysis relevant to the sinusoidal response in the overload region is given in Appendix B.

In Section III we cover the highlights of the simulation program with emphasis on the spectral calculations. Estimates of accuracy are given in Section 3.2. Section IV is devoted to a discussion of the techniques used for measuring the spectrum of the error. We also show how the delta modulator parameters are measured and discuss the realization of a laboratory model delta modulator. Throughout the paper we compare experiment with theory and simulation. In Section V we make some general comments about the various sets of results.

1.4 *System Definition, Terminology, and Symbols*

The following are the terms and symbols used. Our input signal $x(t)$ in Fig. 1 is chosen from a stationary, zero mean, gaussian random process, band-limited to f_b. Its correlation function is $R_{xx}(\tau)$ and cor-

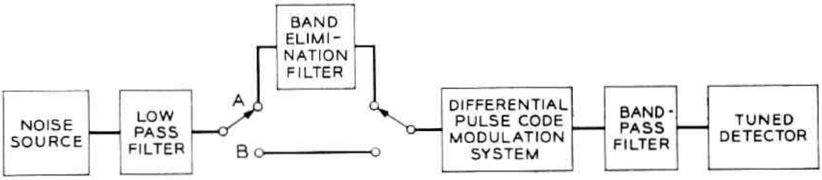


Fig. 2 — Noise loading test: When the switch is at A the uncorrelated noise is measured; when at B the signal plus total noise is measured.

responding spectral density $S_{xx}(\omega)$. The delta modulator is characterized by the step size k of the quantizer, sampling frequency f_s, or normalized sampling frequency $f_s/f_b = F_s$, and an ideal integrator with transfer function $1/s$.† Clearly the maximum slope that the delta modulator can follow is $kf_s \equiv x'_o$, which corresponds to a string of one's at its output. As k approaches zero with x'_o fixed, the granular noise tends to zero and the noise primarily results from slope overload. Under these conditions it will be convenient and quite accurate to represent the delta modulator as a continuous feedback loop with a step size x'_o. We make this assumption in much of the analysis to follow.

Throughout we use $e(t)$ for the total noise, $n(t)$ for that component of the noise uncorrelated with the input signal, $z(t)$ for the unsampled output of the threshold circuit, and $y(t)$ for the output of both the local integrator and the remote integrator (error free transmission). Other symbols are defined as needed.

II. DEFINITION OF THE UNCORRELATED NOISE—AN EQUIVALENT LINEAR SYSTEM

2.1 *General*

In this section we define an equivalent linear system and an additive uncorrelated noise which produce statistical behavior identical with that of the delta modulator up to second moments. Notice that any time invariant linear transformation of the input signal contained in the output may be considered as useful signal because, at least in principle, the input may be recovered by passing it through a fixed linear filter corresponding to the inverse linear transformation.

Definition 1: Equivalent Linear System. We compare the output of the delta modulator $y(t)$ with the output of an "equivalent linear system," defined by Figure 3, whose impulse response $g(t)$ is defined such that the difference

$$y(t) - g(t)*x(t) \triangleq n(t) \tag{1}$$

is uncorrelated with the input $x(t)$; that is,

$$R_{xn}(\tau) = \langle x(t + \tau)n(t)\rangle_{\text{av}} = \langle x(t + \tau)[y(t) - g(t)*x(t)]\rangle_{\text{av}} = 0 \tag{2}$$

where * denotes convolution.

Definition 2: Additive Uncorrelated Noise. The difference $n(t)$

† In Section 2.2 we consider the more practical case of a leaky integrator with transfer function $1/(s + a)$.

given in equation (1), with equation (2) satisfied, is defined as the additive uncorrelated noise. Equation (2) is satisfied when

$$R_{xx}(\tau)*g(-\tau) = R_{xy}(\tau). \tag{3}$$

Taking the Fourier transforms of both sides of this equation and then the complex conjugates we get

$$G(j\omega) = \frac{S_{yx}(\omega)}{S_{xx}(\omega)} = 1 - \frac{1}{S_{xx}(\omega)} \{\mathrm{Re}\,[S_{xe}(\omega)] - j\,\mathrm{Im}\,[S_{xe}(\omega)]\}. \tag{4}$$

We remark here that the transfer function $G(j\omega)$ does not have to be causal; that is, $g(t)$ may be nonzero for $t < 0$.

Applying the orthogonality principle we can see that $g(t)$, thus found, also satisfies†

$$\langle [y(t) - g(t)*x(t)]^2 \rangle_{\mathrm{av}} = \text{minimum}.$$

Notice that from Fig. 3, we can write

$$S_{yy}(\omega) = |\,G(j\omega)\,|^2\, S_{xx}(\omega) + S_{nn}(\omega). \tag{5}$$

If the input process $[x(t)]$ has a spectrum $S_{xx}(\omega)$ such that

$$S_{xx}(\omega) = 0 \quad \text{for} \quad \omega \,\varepsilon \left\{\omega_o - \frac{\Delta\omega}{2}, \omega_o + \frac{\Delta\omega}{2}\right\} \tag{6}$$

where ω_o is a given radian frequency and $\Delta\omega$ a small radian frequency slot, then applying (5) we get

$$S_{nn}(\omega) = S_{yy}(\omega) \quad \text{for} \quad \omega \,\varepsilon \left\{\omega_o - \frac{\Delta\omega}{2}, \omega_o + \frac{\Delta\omega}{2}\right\}. \tag{7}$$

So that for the noise power in this frequency slot we will have

$$\frac{1}{2\pi} \int_{\omega_o - \Delta\omega/2}^{\omega_o + \Delta\omega/2} S_{nn}(\omega)\, d\omega = \frac{1}{2\pi} \int_{\omega_o - \Delta\omega/2}^{\omega_o + \Delta\omega/2} S_{yy}(\omega)\, d\omega. \tag{8}$$

The noise-loading measurement described in Section I applies the technique mentioned here. Thus the noise spectrum and noise power measured are $S_{nn}(\omega)$ and $\langle n^2(t)\rangle_{\mathrm{av}}$. In order to compare experiment and analysis we have to find $G(j\omega)$ and $\langle n^2(t)\rangle_{\mathrm{av}}$. When we are slightly in slope overload, $G(j\omega)$ is practically equal to 1 and all noise definitions so far used are equivalent. When well into slope overload,

† Kazakov used this approach to obtain $g(t)$ or equivalently $G(j\omega)$ in equation (4) in 1960.[13] We were unaware of his work at the time we conceived of the additive uncorrelated noise approach which for our purposes has real physical appeal.

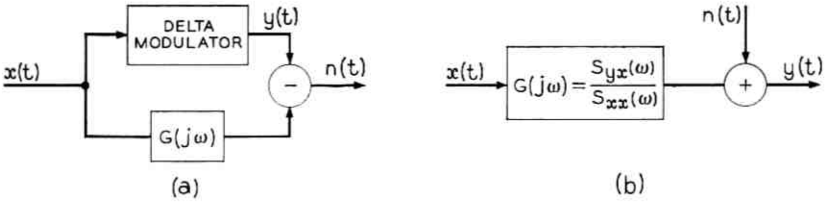


Fig. 3 — Equivalent linear system.

$G(j\omega)$ deviates markedly from unity, thus accounting for the differences between experiment and analysis. To find $G(j\omega)$ we need the cross-spectrum $S_{yx}(\omega)$ which is not presently available analytically. We can find $S_{yx}(\omega)$ using computer simulation; this is what is done in Section III.

2.2 *The Method of Statistical Linearization*

Even though the equivalent transfer function $G(j\omega)$ cannot be found analytically, it may be approximated through the method of statistical linearization.[14] Statistical linearization can be applied to the corresponding continuous system (without sampling) as shown in Fig. 4. The study of the slope overload noise corresponds to the study of this feedback loop, with the nonlinear element in the forward path being a hard limiter with saturation levels $\pm x'_o = \pm kf_s$. The use of the continuous system is not a substantial limitation since the correlated component of the overall noise $e(t)$ is conjectured to be mainly overload noise.

The nonlinearity in the loop will be replaced by a linear gain K chosen according to criteria given in this section and in Appendix A. Independent of the choice of criterion, the equivalent linear system will have the form,

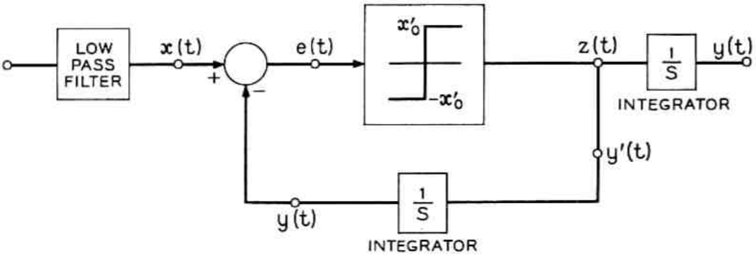


Fig. 4 — Continuous feedback system for the study of slope overload noise.

$$H(f) = A/(s + A) \tag{9}$$

or

$$H(f) = \frac{1}{1 + j(f/f_c)} \tag{10}$$

where f_c is the corner frequency (3dB frequency) of the filter ($f_c = K/2\pi$).

In a real system, we generally have a leaky integrator whose transfer function is of the form $1/(s + a)$. Then it is easy to show that

$$H(f) = \frac{H(0)}{1 + j\dfrac{f}{f_c}} \tag{11}$$

where

$$H(0) = \frac{K}{K + a}, \quad \text{and} \quad f_c = \frac{K + a}{2\pi}.$$

The variety of ways by which one can determine the equivalent gain K, are presented in detail in Appendix I. Let us call K_1 the equivalent gain found with the assumption that the input to the nonlinearity is gaussian with variance σ^2 equal to the overload noise power. Denote by K_2 the equivalent linear gain when the gaussian assumption is removed. Let K_3 be the equivalent gain determined under the requirement that the difference between the overload error and the input to the linearized element be uncorrelated (for $\tau = 0$) with the input signal. In order to compare the equivalent linear filter transfer functions with the computer simulation results we plot the magnitude $|G(j\omega)|$ of the transfer function [calculated using equation (4) and the computer generated cross-spectra] for $kf_s/f_b = kF_s = 2$ and 4 in Fig. 5. From these figures we find that the equivalent linear system may be approximated by a one-pole tranfer function with corner frequency, $f_c = 0.358\ f_b$ and $0.94\ f_b$, respectively, and corresponding dc gains (caused by the small leak in the integrators) $H(0) = 0.89$ and 0.98. The results of the comparison are summarized in Table I. Thus there is reasonable agreement between the equivalent linear system transfer function obtained from computer simulation and all the approximate statistical linearization methods.

2.3 *Describing Function Method*

In Appendix B a method is outlined for obtaining an equivalent frequency dependent complex gain for a delta modulation system with

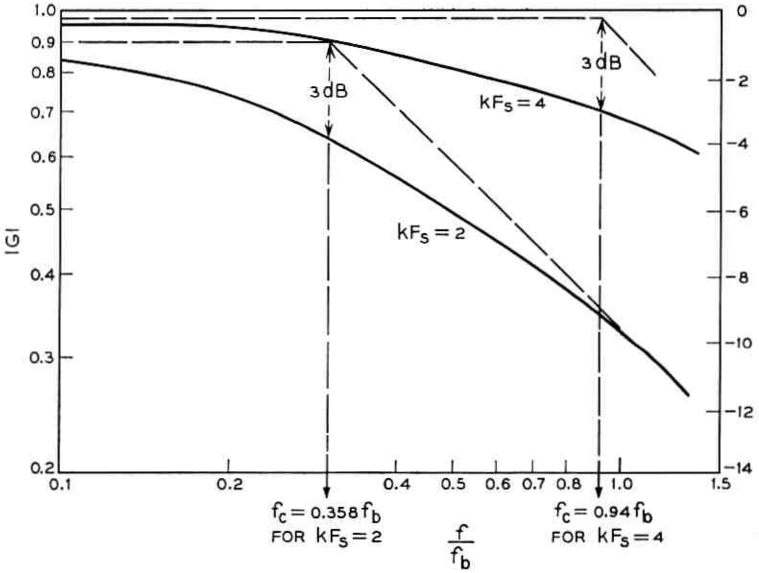


Fig. 5 — Bode plot of the gain of the equivalent linear system (computer results, $a = 0.16\ \omega_b$).

a leaky integrator, subject to a sinusoidal input of amplitude X_o and frequency ω_o, under pure slope overload conditions. This complex gain is defined to be the ratio of the complex amplitude of the fundamental of the output to the complex amplitude of the input sinusoid. This deterministic equivalent linearization method is well known as the describing

Table I—Parameters of Equivalent Linear Systems

	$kF_s = 2$		$kF_s = 4$	
	K/f_b Equivalent linear gain	f_c/f_b Corner frequency	K/f_b Equivalent linear gain	f_c/f_b Corner frequency
Computer	$K_0 = 2.10$	0.358	$K_0 = 5.90$	0.94
Gaussian assumption	$K_1 = 2.34$	0.398	$K_1 = 7.13$	1.13
Without gaussian assumption	$K_2 = 1.48$	0.262	$K_2 = 4.00$	0.64
Correlated noise† Approach	$K_3 = 2.70$	0.43	$K_3 = 5.90$	0.94

† The leaky integrator effect is neglected; if taken into account the results would be somewhat smaller. For $kF_s = 4$, the effect of leak is negligible.

function method. The corresponding magnitude of the equivalent gain is given as a function of "normalized" frequency ($\omega_o X_o / x_o'$) in Fig. 6 when the leak in the integrator goes to zero. For $x_o' = kF_s = 2$ the 3 dB point (corner frequency) is at $f_c \cong 0.4 f_b$, which is in good agreement with the results in Table I. Measured values of equivalent gain shown on Fig. 6 agree well with theoretical predictions.

III. COMPUTER SIMULATION TECHNIQUE

3.1 *Basic Concepts*

Computer simulation provides a convenient method of studying the characteristics of delta modulation systems without actually building them. The computer can also provide accurate numerical results against which to compare experimental results from laboratory or

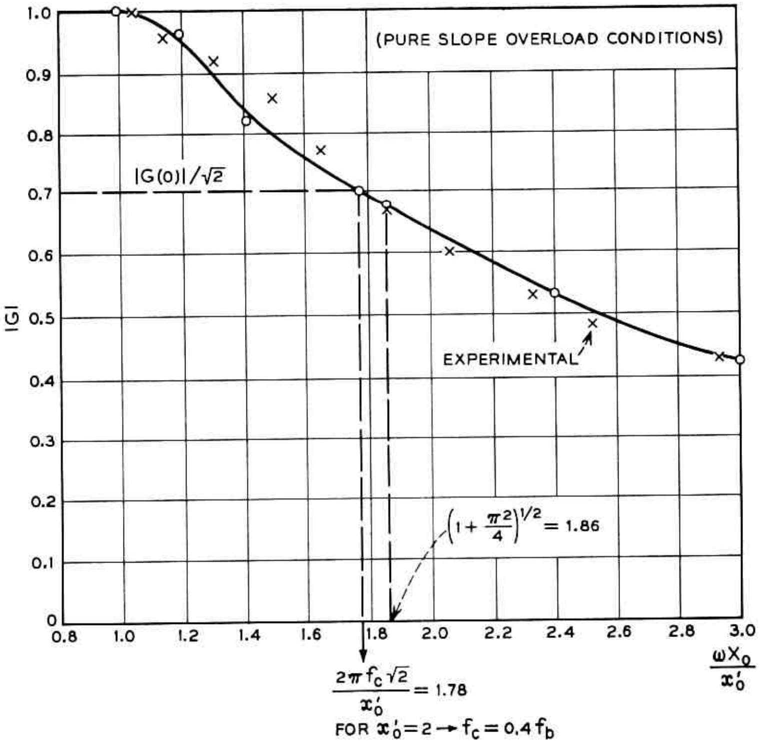


Fig. 6 — Harmonic response equivalent gain of a single integration ΔM.

production models. Computer simulation is a compromise between laboratory techniques and analytical techniques in that it is easy to change the program in order to study a variety of system parameters or to introduce defects similar to those expected in practical systems. On the other hand, the simulated system is an idealized abstraction which does not represent the practical system in full detail.

The BLØDI programming system, used for the simulation, results in a program which processes a sequence of samples by whatever set of mathematical operations may be specified by a block diagram.[15] BLØDI flexibility allows the use of FORTRAN for such things as computing estimates of signal statistics, for which FORTRAN is more efficient. Figure 7 indicates the basic philosophy: a FORTRAN program supervises the entire operation calling the various subprograms as needed. By structuring the simulation programs as a hierarchy of modules, changes in one area of the model could be effected without involving the entire program. The program was purposely written with extensive use of subroutines. This for example, makes it applicable to differential pulse code modulation (DPCM) by simply changing the subroutine for the quantizer. The actual programs are of interest to only a few people, and are not listed here. Appendix C gives a discussion of the computational formulas used to estimate correlation functions and spectra.

3.2 *Accuracy of Computer Estimates of Spectrum of Error*

In order to estimate the expected accuracy of the spectrum estimates from the computer simulation, the following example is given: In the simulation the estimate $S_e(k)$ is made on the basis of 10,000 input samples. Here k is an integer index related to frequency, and

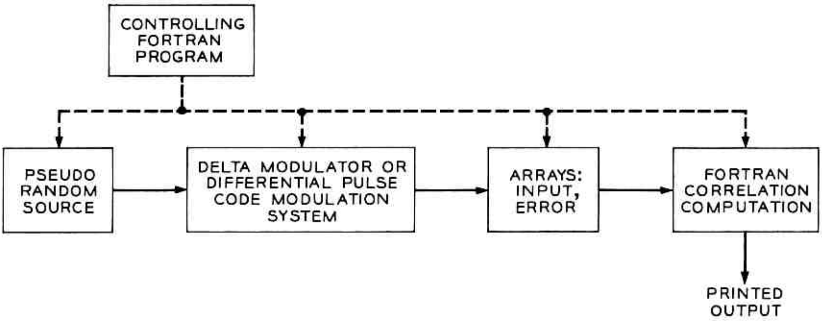

CONTROLLING FORTRAN PROGRAM
PSEUDO RANDOM SOURCE
DELTA MODULATOR OR DIFFERENTIAL PULSE CODE MODULATION SYSTEM
ARRAYS: INPUT, ERROR
FORTRAN CORRELATION COMPUTATION
PRINTED OUTPUT

Fig. 7 — Computer simulation.

the subscript e refers to the total noise $e(t)$. In one particular run, ten intermediate estimates based on 1,000 input samples each were made. Using the notation $S_{e1}(k)$ through $S_{e10}(k)$ for these we have

$$\text{sample mean} = S_e(k) = \tfrac{1}{10}[S_{e1}(k) + \cdots + S_{e10}(k)] \tag{12}$$

$$E[S_e(k)] = \mu,$$

$$\text{sample variance} = \text{var} = \tfrac{1}{10}\{[S_{e1}(k) - S_e(k)]^2 + \cdots + [S_{e10}(k) - S_e(k)]^2\}. \tag{13}$$

One can then show that the variance of the estimate $S_e(k)$ relative to μ is estimated by

$$E\{[S_e(k) - \mu]^2\} = \tfrac{1}{9}\,\text{var}. \tag{14}$$

For the cancellation technique and one particular value of k, representing a low frequency point in the spectrum, a numerical computation yielded:

$$\frac{(\frac{1}{9}\,\text{var})^{\frac{1}{2}}}{S_e(k)} = 0.066.$$

Although the result may in general depend on k, spot checks at other points yielded similar results.

Assuming the estimate is a gaussian random variable with 0.066 = the ratio of standard deviation to mean, the result indicates that the estimate is within ±½ dB of the true mean with probability 0.9.

Other sources that could contribute errors in the results of the simulation include: (*i*) random error in measurements caused by finite averaging time constant, estimated as ±½ dB, (*ii*) round-off errors in computation, which are most significant in the region of high noise, and (*iii*) systematic error resulting from differences between the spectral shape of the simulated input and the output of the laboratory noise generator used in the experiments.

IV. EXPERIMENTAL TECHNIQUES

The extensive analytical and computer work that has been presented was undertaken to a large extent to gain a better understanding of an actual laboratory delta modulation system.

4.1 *Description of the Delta Modulator*

The delta modulator used for the measurements is a variable parameter system in which the step size, leak, and sampling rate are independently variable. Figure 8 is a block diagram of the encoder. The difference between the input and the local integrator output is amplified and presented to the threshold detector. This circuit controls the output of the pulse generator.

The local integrator has circuit elements which can be changed to vary the important parameters of the system. The capacitor C controls the step size; since the amplifier has a high input impedance, the resistor R_L controls the leak.

The decoder consists of a regenerator for amplitude and phase regeneration and a decoder integrator which is a duplicate of the local integrator. The system was operated at a 12.5 MHz sampling rate.

Waveforms in a delta modulator are rather simple; nevertheless, some are shown in Fig. 9 to illustrate the actual operation of the system. Figure 9a indicates the output of the decoder and the pulse output of the coder when no input is presented to the system. A delta modulator should change state every clock period with no input; the photograph illustrates this. This waveform can be used to measure the step size.

Figure 9b illustrates the output of the system when it is in overload. The slope of the input sinusoidal signal is greater than the slope that the delta modulator can follow. Therefore, the output is a triangular wave whose slope is a measure of the normalized step size. An interesting feature can be seen by observing the slopes of the flat steps in this picture. In the lower half, they slant upward and in the upper half they slant downward, illustrating the leaking off of the capacitor voltage.

Figure 9c illustrates the response of the delta modulator to a sine wave whose amplitude is below overload. The rather blurred trace

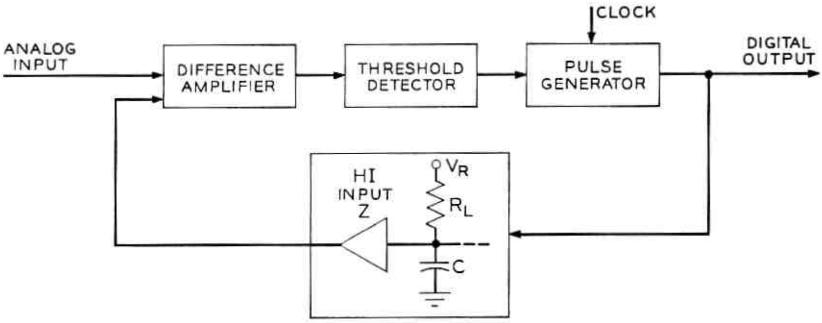


Fig. 8 — Delta modulator encoder.

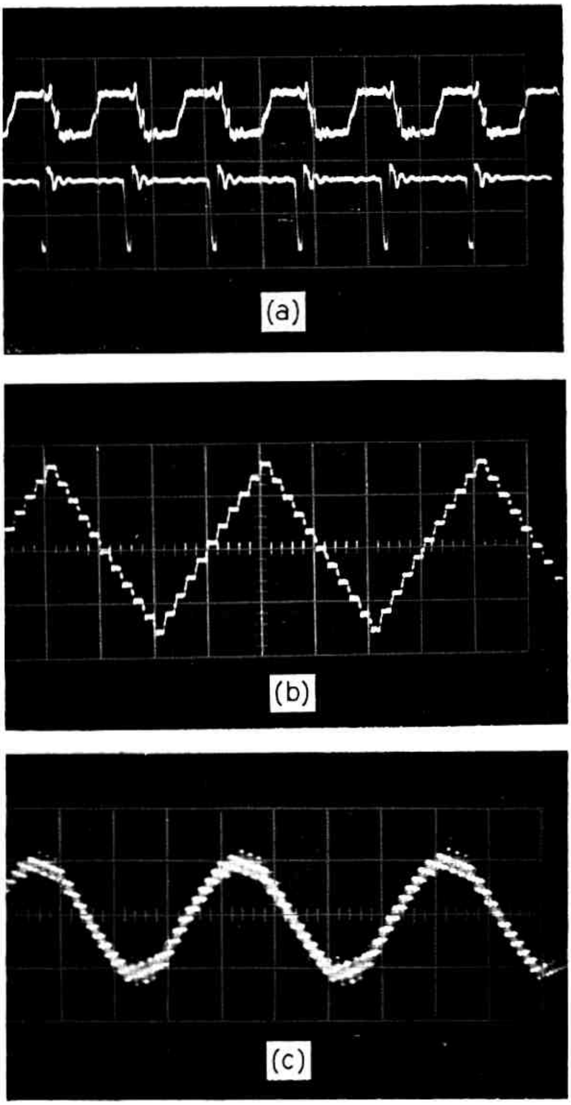


Fig. 9 — (a) Analog output $y(t)$ and digital output with no input (100 ns/cm); (b) $y(t)$ with system in overload (400 ns/cm); (c) $y(t)$ with system not in overload (2 μs/cm).

results because the frequency of the sine wave is not a submultiple of the sampling frequency.

4.2 *Noise Loading Test*

The use of the noise loading test to measure nonlinearities in a transmission system has been mentioned in Section 1.3. In this test, as shown in Fig. 2, a wideband of gaussian noise is applied to a low-pass filter to band limit the input to the delta modulator. With the switch in the upper position, a narrow band of noise can be eliminated from the input signal. Several band elimination filters are available to cover the input spectrum. This signal is fed into the system and only that band from which signal has been eliminated is allowed to pass to the tuned detector. With the switch in the upper position, only noise introduced by nonlinearities in the delta modulator and uncorrelated with the input is passed into the detector. The power spectrum of the uncorrelated noise component $n(t)$ can be measured by changing the center frequency of the band elimination and bandpass filters. With the switch in the lower position, the full signal enters the system and the tuned detector reads signal and noise within the passband.

4.3 *Cancellation Technique*

To measure the total noise output, $e(t)$, and its spectrum, the arrangement shown in Fig. 10 was set up. The signal is fed to the delta modulator and the output of the delta modulator and the attenuated and delayed input are compared, their difference being the noise introduced by the system.

The immediate problem encountered in this technique is the adjustment of the variable attenuator and the delay to cancel the signal

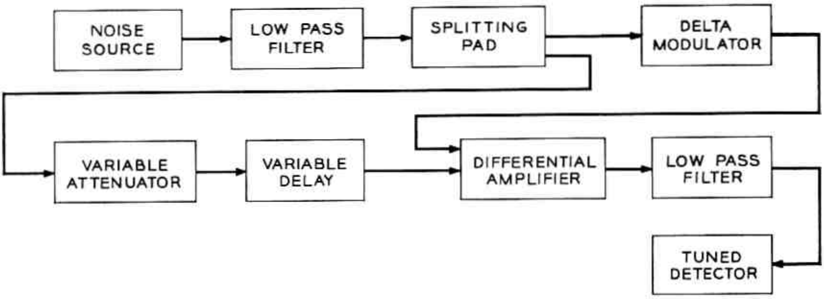


Fig. 10 — Cancellation technique.

component at the output of the delta modulator. The delay is not the same for all frequencies and will have more of an effect at high rather than low frequencies. A sine-wave input whose amplitude was less than that required to overload the system was used to correctly null the system, making the equivalent gain unity. The frequency was chosen as high as conveniently possible (within the signal band) so that the effects of delay could be observed on the nulling procedure. Attenuation and delay were adjusted to produce a null at the input frequency at the tuned detector. Then the noise source was used to replace the sine wave and the output noise measured as a function of frequency by the tuned detector. The gain and delay should be adjusted at each frequency where the noise spectrum is measured. The rather broad null, particularly at the lower frequencies, makes this measurement both tedious and inaccurate. Consequently only the high-frequency approach was used.

The noise-free output signal is measured by removing that input to the difference amplifier that comes from the delta modulator.

4.4 *Accuracy of the Measurements*

The tuned detector used to measure the noise in these experiments was a 37B transmission measuring set. It has a frequency window of about 400 Hz. Therefore, when the noise is measured at a particular frequency, a 400 Hz band is actually measured and the meter reading must be averaged, ignoring peaks. It is estimated that the readings are accurate to about ±0.5 dB.

Another source of error arises in the determination of the normalized step size kF_s. As mentioned above kF_s can be found from direct measurement on an oscilloscope, or by using a square wave input that overloads the system. A small error in this measurement is equivalent to a displacement in the noise curve (or signal-to-noise ratio) when plotted against kF_s. The noise changes in the overload and granular regions about 1 dB for every dB change in kF_s. Furthermore, the spectrum in overload also changes very rapidly with kF_s.

Therefore, it is fair to conclude that the experimental results in Section V are accurate to about ±1 dB.

V. RESULTS

5.1 *Noise Loading Results—Uncorrelated Noise Component*

In Fig. 11 we have plotted the spectrum of the signal uncorrelated component of the noise as obtained by the noise loading test for three

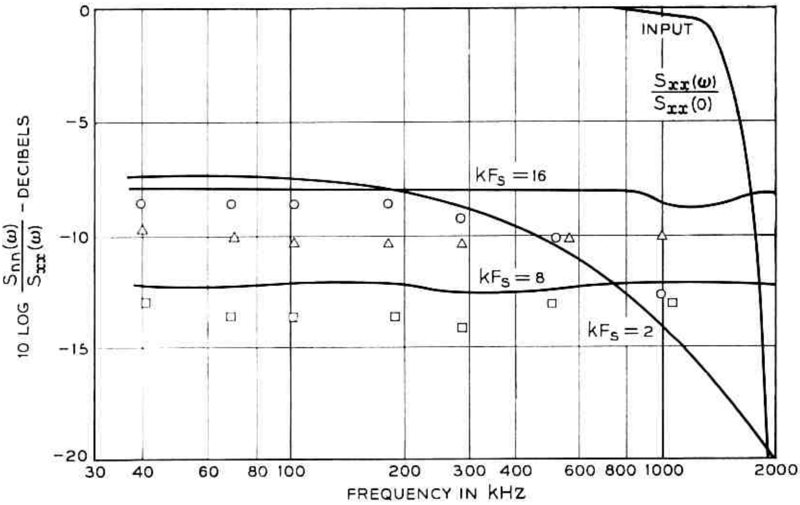


Fig. 11 — Δ-mod uncorrelated noise spectrum, $F_s = 8$.
○ $kF_s = 2$
□ $kF_s = 8$
◇ $kF_s = 16$
— computer results.

values of kF_s. For $kF_s = 8$ and 16 notice that the noise spectrum is flat, as expected, since granular noise is predominant. When $kF_s = 2$, overload noise is controlling, and the noise spectrum is largest at low frequencies. Agreement between the computer generated spectrum and the measured spectrum is good except where the granular noise is small. In this region, it is believed that round-off errors in the computer simulation account for the discrepancy. Integration of the noise spectrum yields the signal to noise curve of Fig. 12 plotted as a function of kF_s.

5.2 *Cancellation Technique—Total Noise*

Noise spectrum measurements obtained by the cancellation technique are compared with computer results in Fig. 13. As before, for $kF_s = 8$ and 16 the spectrum is flat and nearly identical in level with the noise loading results. When well into overload ($kF_s = 2$), the total noise spectrum peaks at the high frequency end. This behavior is readily explained in terms of our equivalent linear system. Consider the difference $e(t)$ between the input signal and the output of the

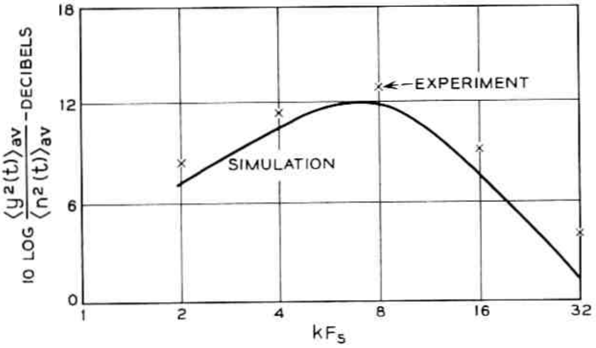


Fig. 12 — Δ-mod signal to uncorrelated noise ratio, $F_s = 8$.

equivalent linear system of Fig. 3

$$e(t) = x(t) - y(t) = x(t) - n(t) - \int_{-\infty}^{\infty} g(t - \tau)x(\tau)\,d\tau \tag{15}$$

or

$$e(t) = \int_{-\infty}^{\infty} [\delta(t - \tau) - g(t - \tau)]x(\tau)\,d\tau - n(t). \tag{16}$$

Since $n(t)$ and $x(t)$ are uncorrelated by definition, it is an easy matter to show that the error spectrum of the total noise is

$$S_{ee}(\omega) = S_{nn}(\omega) + |\,1 - G(j\omega)\,|^2 S_{xx}(\omega). \tag{17}$$

Substituting $H(j\omega)$, obtained by statistical linearization and given in equation (10) for the equivalent linear system function $G(j\omega)$ in equation (17), we get

$$S_{ee}(\omega) = S_{nn}(\omega) + \frac{\dfrac{\omega^2}{\omega_c^2}}{1 + \dfrac{\omega^2}{\omega_c^2}} S_{xx}(\omega). \tag{18}$$

From either equation (17) or (18) we can see that when $G(j\omega)$ is essentially unity (in the granular region) that the total noise is given by $S_{nn}(\omega)$. On the other hand, when well into overload, the low frequency portion of the total noise is determined by $S_{nn}(\omega)$ and the noise at high frequencies increases due to the second term in equation (18), the term linearly dependent on the input. Indeed, we can use the measured noise spectrum in Fig. 13 for $kF_s = 2$ along with equa-

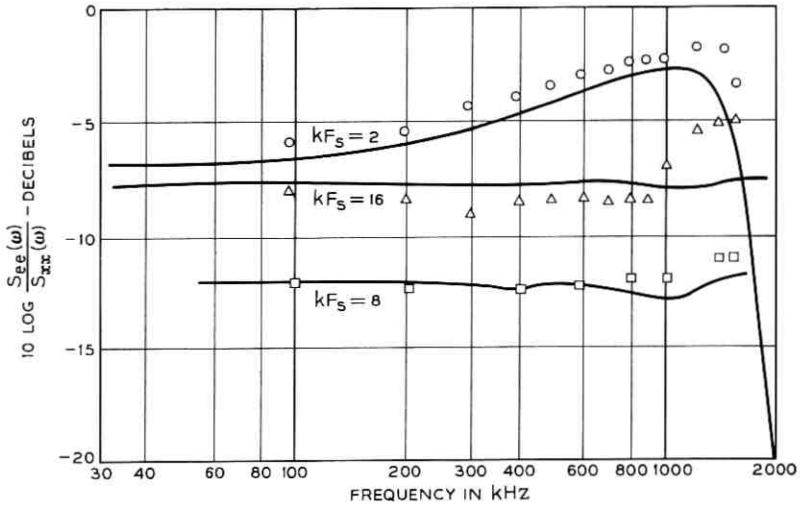


Fig. 13 — Δ-mod error spectrum, cancellation technique, $F_s = 8$.
○ kF$_s$ = 2
□ kF$_s$ = 8
◇ kF$_s$ = 16
— computer results.

tion (18) to determine the corner frequency for the equivalent linear system. The f_c so obtained is about 0.4 f_b in agreement with the analysis.

For completeness, we present in Fig. 14 the signal-to-total-noise ratio obtained by integrating the curves of Fig. 13. In addition, we have noted the corresponding analytical results obtained by using the

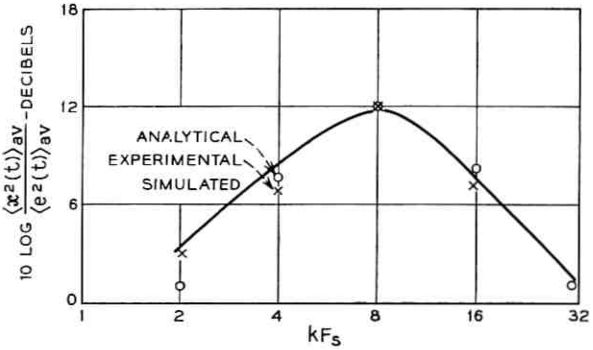


Fig. 14 — Δ-mod signal to error ratio, cancellation technique, $F_s = 8$.

results of Refs. 9 and 10. Agreement is good except when far into the overload region where it is known that the mean square value of the total noise obtained analytically is a coarse upper bound.

VI. ACKNOWLEDGMENT

We are indebted to R. W. Stroh for his initial efforts in the development of the computer program.

APPENDIX A

Statistical Linearization

A.1 *General*

In this appendix we consider the delta modulation system under pure slope overload conditions. Our objective is to replace the hard limiter in the encoder loop with a linear amplifier. We give three methods for the determination of the gain in this linear approximation.

A.2 *Conventional Statistical Linearization—Gaussian Assumption*

First, we use the statistical linearization method attributed to Booton.[18] We isolate the hard-limiter in Fig. 4 with input $e(t)$ and output $z[e(t)]$ in order to replace it with an ideal linear amplifier of gain K_{eq}. This gain factor is chosen such that $K_{eq}e(t)$ differs least in the mean square sense from $z[e(t)]$. It is readily shown the optimum K_{eq} satisfies

$$K_{eq} = \frac{\langle ez\rangle_{\text{av}}}{\langle e^2\rangle_{\text{av}}}. \tag{19}$$

For the hard limiter, under the assumption that $e(t)$ is gaussian, we get the well known result[14]

$$K_{eq} = x_o'\left(\frac{2}{\pi\langle e^2\rangle_{\text{av}}}\right)^{\frac{1}{2}} \equiv K_1 . \tag{20}$$

A.3 *Removal of Gaussian Assumption*

In general, $e(t)$ will not be gaussian; though this is commonly assumed in all references to the statistical linearization method. We remove this assumption in this section since we can determine both $\langle ez\rangle_{\text{av}}$ and $\langle e^2\rangle_{\text{av}}$ using the approach given in Ref. 9. Since $\langle e^2\rangle_{\text{av}}$ was found in that reference, we need only consider $\langle ez\rangle_{\text{av}} = R_{ez}(0)$.

Notice that when

and where

$$\left.\begin{array}{ll} e(t) > 0, & z(t) = x_o' \\ e(t) < 0, & z(t) = -x_o' \end{array}\right\}. \tag{21}$$

Hence

$$\langle ez \rangle_{\text{av}} = x_0' \langle | \, e(t) \, | \rangle_{\text{av}} = x_0' \text{ ave } [e(t)] \, |_{\text{p.b.}} \tag{22}$$

that is, the average of $e(t)$ over the positive bursts (p.b.) only of the slope overload noise.

Following the procedure developed in Ref. 9, we obtain

$$x_0' \langle | \, e(t) \, | \rangle_{\text{av}} = \frac{1}{\pi} \frac{b_1^{\frac{3}{2}}}{b_2^{\frac{1}{2}}} \left[\frac{3b_1^{\frac{1}{2}}}{x_0'} \right] \exp\left[-\frac{(x_0')^2}{2b_1} \right] \Omega(\chi_1) \tag{23}$$

where

$$\begin{aligned} \chi_1 &= \frac{x_0'\sqrt{2}}{3(b_1)^{\frac{1}{2}}} \\ \Omega(\chi_1) &= 1 - (1 - \chi_1^2) \exp\left(-\frac{\chi_1^2}{2} \right) - \chi_1^3 \Phi(\chi_1) \\ \phi(\chi_1) &= \int_{\chi_1}^{\infty} \exp\left(-\frac{z^2}{2} \right) dz \\ b_n &= \int_{-f_o}^{f_o} \omega^{2n} S_{xx}(\omega) \, df. \end{aligned} \tag{24}$$

In Ref. 9 it was found that

$$\langle e^2(t) \rangle_{\text{av}} = \frac{1}{4(2\pi)^{\frac{1}{2}}} \left(\frac{b_1^2}{b_2} \right) \left(\frac{3b_1^{\frac{1}{2}}}{x_0'} \right)^5 \exp\left\{ -\left[\frac{(x_0')^2}{2b_1} \right] \right\} A(\chi) \tag{25}$$

where $A(\chi)$ is given in equation (66) of Ref. 9. Hence

$$K_{eq} = 4 \left(\frac{2}{\pi} \right)^{\frac{1}{2}} \left(\frac{b_2}{b_1} \right)^{\frac{1}{2}} \left(\frac{x_0'}{3b_1} \right)^3 \frac{\Omega(\chi_1)}{A(\chi)} \triangleq K_2 \, . \tag{26}$$

A.4 *Equivalent Gain from Definition of Equivalent Linear System*

Among the many other viewpoints that might be adopted to find K_{eq}, we single out one that makes use of the definition of the equivalent linear system given in the text. Recall that

$$G(j\omega) \equiv \frac{S_{yx}(\omega)}{S_{xx}(\omega)} = 1 - \frac{S_{ex}(\omega)}{S_{xx}(\omega)} \tag{27}$$

or

$$S_{ex}(\omega) = S_{xx}(\omega) - G(j\omega)S_{xx}(\omega). \tag{28}$$

If we integrate equation (27) over $(-2\pi f_o$ to $+2\pi f_o)$ and choose $G(j\omega) = K_3/K_3 + j\omega$, we obtain the following equation defining K_3.

$$\langle x(t)e(t)\rangle_{\text{av}} = R_{ex}(0) = \langle x^2(t)\rangle_{\text{av}} - \frac{1}{2\pi}\int_{-2\pi f_o}^{2\pi f_o} \frac{S_{xx}(\omega)}{1 + j\dfrac{\omega}{K_3}}\,d\omega. \tag{29}$$

Noticing that $S_{xx}(\omega)$ is an even function of ω, using

$$F(f) = 2S_{xx}(2\pi f) \quad \text{for} \quad f > 0, \tag{30}$$

and defining $f_c = K_3/2\pi$, we get

$$h(f_c) = \int_0^{f_o} \frac{F(f)\,df}{1 + \left(\dfrac{f}{f_c}\right)^2} = \langle x^2(t)\rangle_{\text{av}} - \langle x(t)e(t)\rangle_{\text{av}}\,. \tag{31}$$

The left side of equation (31) is a function of f_c only, and hence of K_3, while the right side of equation (30) is known; a formula for $\langle x(t)e(t)\rangle_{\text{av}}$ has been found.[17] Equation (31) can be shown to always have a solution.

A little reflection will convince the reader that equation (31) could have been obtained from scratch by preselecting the form of the equivalent linear system, and requiring that $x(t)$ be uncorrelated with $n(t)$ at $\tau = 0$. The approach we have taken could be generalized to match various spectral moments of the processes under consideration. This would entail multiplying equation (28) by ω^{2n} prior to integration and choosing the number of parameters in $G(j\omega)$ equal to the number of moments matched. In general a set of simultaneous nonlinear equations would have to be solved and quantities such as $\langle d^n x(t)/dt^n e(t)\rangle_{\text{av}}$ obtained using the techniques of Ref. 9. Fortunately, no such generalization is required. As we see below and from Table I all of the techniques used in this Appendix give good agreement with computer simulation.

Example: Application of equation (31) to flat band limited signals

$\langle x^2\rangle_{\text{av}} = 1, f_o = 1$ gives

$$h(f_c) = \int_0^1 \frac{df}{1 + \left(\dfrac{f}{f_c}\right)^2} = f_c \tan^{-1}\left(\frac{1}{f_c}\right) = 1 - \langle xe\rangle_{\text{av}}\,. \tag{32}$$

APPENDIX B

Harmonic Response of a Delta Modulator with a Leaky Integrator Under Pure Slope Overload Conditions

B.1 *Introduction*

Consider the single-integration delta modulator with a leaky integrator under pure slope overload conditions. The problem is to find the steady-state response of this nonlinear system to a sinusoidal input. The analysis is applicable to differential pulse code modulation and delta modulation with a more complicated linear network in the feedback path.

Consider a sinusoidal input signal:

$$x(t) = X_o \cos \omega_o t \tag{33}$$

with

$$\omega_o = 2\pi f_o . \tag{34}$$

In the steady-state the output $y(t)$ will be a periodic function of t with period $1/f_0$. The maximum value of the magnitude of the slope of the input sinusoidal signal is clearly equal to $\omega_0 X_o$ so that if

$$\frac{\omega_o X_o}{x_o'} \leqq 1 \tag{35}$$

the output will follow the input and we will have

$$y(t) = x(t) = X_o \cos \omega_o t. \tag{36}$$

Suppose now that $x_o' < \omega_o X_o$. In this case slope overload occurs. Call ϕ the value of $\omega_o t - 2n\pi$ (where n is a positive integer) for which slope overload occurs for the first time after the beginning of the nth period. Assuming that we have reached the steady-state, the value of ϕ will be the same for all periods.

Clearly $0 < \phi < \pi/2$. The slope of the input signal at the transition point A (Fig. 15) will be negative and equal to $-X_o\omega_o \sin \phi$. (The second derivative at A is also negative and equal to $-X_o\omega_o^2 \cos \phi$.)

For slope overload to begin at A we should have;† $-\omega_o X_o \sin \phi = -x_o'$, so that

$$\sin \phi = \frac{x_o'}{\omega_o X_o}. \tag{37}$$

† A similar analysis may be made in the asymmetric case, that is, when the positive overloading slope is not equal to the negative overloading slope.

At this transition point the output signal begins to follow an exponential curve such that

$$y(t) = X_o \cos \phi - x_o' \frac{1 - \exp\left(-a \dfrac{\omega_o t - \phi - n\pi}{\omega_o}\right)}{a} \tag{38}$$

as long as $y(t)$ exceeds $x(t)$. The exponential segment ends when $y(t)$ and $x(t)$ once again become equal as shown in Fig. 15. For small leak, the response in overload is clearly linear in time. As long as $|\theta| < \phi$ we have for all n

$$y(t) = \begin{cases} X_o \cos \omega_o t \quad \text{for} \quad \theta + n\pi \leqq \omega_o t \leqq \phi + n\pi \\ (-1)^n X_o \cos \phi + (-1)^{n+1} x_o' \dfrac{1 - \exp\left(-a \dfrac{\omega_o t - \phi - n\pi}{\omega_o}\right)}{a}. \\ \qquad \text{for} \quad \phi + n\pi \leqq \omega_o t \leqq \theta + (n+1)\pi \end{cases} \tag{39}$$

It it easy to show that the region where equation (39) is true may be translated to the condition

$$1 \leqq \frac{\omega_o X_o}{x_o'} \leqq \left\{1 + \frac{\pi^2}{4}\left[\frac{1 - \exp\left(-\dfrac{a\pi}{\omega_o}\right)}{\dfrac{a\pi}{\omega_o}}\right]^2\right\}^{\frac{1}{2}}. \tag{40}$$

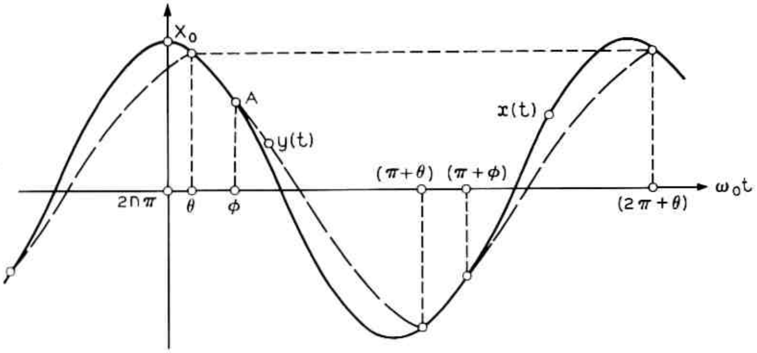


Fig. 15—Slope overload for

$$1 < \frac{\omega_0 x_0}{x_0'} \leqq \left\{1 + \frac{\pi^2}{4}\left[\frac{1 - \exp\,(-a\pi/\omega_0)}{a\pi/\omega_0}\right]^2\right\}^{\frac{1}{2}}$$

(leaky integrator).

In the limit when a goes to zero (no leak) equations (39) and (40) reduce to results obtained previously by Baikovskii.[16] The quantities θ and ϕ coalesce when

$$\frac{\omega_o X_o}{x_o'} \geqq \left\{1 + \frac{\pi^2}{4}\left[\frac{1 - \exp\left(-\dfrac{a\pi}{\omega_o}\right)}{\dfrac{a\pi}{\omega_o}}\right]^2\right\}^{\frac{1}{2}} \tag{41}$$

and the output is made up of segments of an exponential curve as shown in Fig. 16. From Fig. 16 we see that

$$\cos \phi_o = x_o' \frac{1 - \exp\left(-\dfrac{a\pi}{\omega_o}\right)}{2aX_o} \tag{42}$$

and for all n

$$y(t) = (-1)^n x_o' \left\{\frac{1 - \exp\left(-\dfrac{a\pi}{\omega_o}\right)}{2a} - \frac{1 - \exp\left[-\dfrac{a}{\omega_o}(\omega_o t - \phi_o - n\pi)\right]}{a}\right\}. \tag{43}$$

Notice that in this case the magnitude of the output depends only on the frequency of the input sinusoidal waveform and not on its

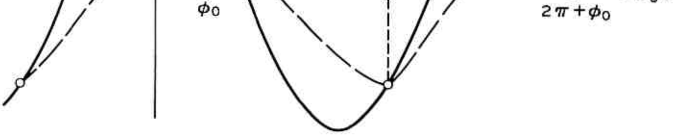


Fig. 16 — Slope overload for

$$\frac{\omega_0 x_0}{x_0'} > \left\{1 + \frac{\pi^2}{4}\left[\frac{1 - \exp\,(-a\pi/\omega_0)}{a\pi/\omega_0}\right]^2\right\}^{\frac{1}{2}}$$

(leaky integrator).

amplitude. Only the phase of $y(t)$ depends on X_o. Clearly when the leak goes to zero $(a = 0)$ the response is triangular.

B.2 *Harmonic Analysis of* $y(t)$

In all three regions above the output $y(t)$ is a periodic function of t with period $2\pi/\omega_o$ such that

$$y\left(t + \frac{\pi}{\omega_o}\right) = -y(t). \tag{44}$$

Hence $y(t)$ contains only odd harmonics; it is a straight-forward matter to compute the Fourier coefficients. The complex equivalent gain is given by the ratio of the coefficient of the fundamental in the output to X_o. We leave this manipulation to the interested reader and merely provide a curve of equivalent gain computed for the case of a perfect integrator $(a = 0)$, in Fig. 6. Experimental points on the curve are seen to be in close agreement with the analysis.

APPENDIX C

Computational Formulas to Estimate Correlation Function and Spectra

From the sample sequences x_i and e_i for signal and error produced by the simulator, autocorrelation and cross-correlation functions $R_e(j)$, $R_x(j)$, $R_{xe+}(j)$, and $R_{xe-}(j)$ were estimated as the arithmetic means of $e_m e_{m+j}$, $x_m x_{m+j}$, $e_m x_{m+j}$ and $e_m x_{m-j}$, respectively. In the computations, sample sequences of length 10,000 were used. Correlations were computed up to $j = 30$. It is easy to show that spectrum estimates may be obtained by using the correlation estimates as coefficients of a Fourier series. In the case of the cross spectrum, real and imaginary parts must be computed. For clarity, the formulas are listed below. Using the relationship derived in Section 2.1, the uncorrelated noise spectrum may be estimated by:

$$S_n(j) = S_e(j) - \frac{1}{S_x(j)} (\{\text{Re}\,[S_{xe}(j)]\}^2 + \{\text{Im}\,[S_{xe}(j)]\}^2). \tag{45}$$

To smooth possible ripples in the spectrum estimates due to time truncation of the correlation functions, a hanning window function was used.[19] This smoothing amounts to replacing each spectrum estimate by a linear sum of the estimate and the two adjacent estimates, with weights $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$.

Error spectrum:

$$S_e(j) = R_e(0) + 2 \sum_{l=1}^{N-2} R_e(l) \cos \frac{lj\pi}{N-1} + R_e(N-1) \cos j\pi. \tag{46}$$

Signal spectrum:

$$S_x(j) = R_x(0) + 2 \sum_{l=1}^{N-2} R_x(l) \cos \frac{lj\pi}{N-1} + R_x(N-1) \cos j\pi. \tag{47}$$

Cross spectrum:

$$\begin{aligned} \operatorname{Re} \{S_{xe}(j)\} = R_{xe}(0) &+ \sum_{l=1}^{N-2} [R_{xe^+}(l) + R_{xe^-}(l)] \cos \frac{lj\pi}{N-1} \\ &+ \tfrac{1}{2}[R_{xe^+}(N-1) + R_{xe^-}(N-1)] \cos j\pi \end{aligned} \tag{48}$$

$$\operatorname{Im} \{S_{xe}(j)\} = \sum_{l=1}^{N-2} [R_{xe^+}(l) - R_{xe^-}(l)] \cos \frac{lj\pi}{N-1}. \tag{49}$$

REFERENCES

1. Deloraine, E. M., Van Merlo, S., and Derjavitch, B., "Method and System of Impulse Transmission," French Patent No. 932–140, August 10, 1946, p. 140.
2. de Jager, F., "Delta Modulation, a Method of PCM Transmission Using 1-Unit Code," Philips Res. Rep. 7, 1952, pp. 442–466.
3. Cutler, C. C., "Differential Quantization of Communications Signals," U.S. Patent No. 2,605,361, applied for June 29, 1950, issued July 29, 1952.
4. Graham, R. E., "Predictive Quantization of Television Signals," IRE Wescon Record, Part IV, August 1958, pp. 147–156.
5. *Special Issue on Redundancy Reduction,* C. C. Cutler, ed., Proc. IEEE, *55,* No. 3 (March 1967).
6. O'Neal, J. B., Jr., "Delta Modulation Quantizing Noise Analytical and Computer Simulation Results for Gaussian and Television Input Signals," B.S.T.J., *45,* No. 1 (January 1966), pp. 117–141.
7. Inose, H., Yasuda, Y., Kawai, Y., and Takagi, M., "Subscriber Line Circuits for an Experimental Time Division Multiplex Exchange System Featuring Delta Modulation Techniques," J. Elec. Commun. Engineers of Japan, *44,* (1961), pp. 1322–1328.
8. Zetterberg, L. A., "A Comparison Between Delta and Pulse Code Modulation," Ericsson Technics, *11,* No. 1 (January 1955), pp. 95–154.
9. Protonotarios, E. N., "Slope Overload Noise in Differential Pulse Code Modulation Systems," B.S.T.J., *46,* No. 9 (November 1967), pp. 2119–2162.
10. Van de Weg, H., "Quantizing Noise of a Single Integration Delta Modulation System with an N-Digit Code," Philips Res. Rep. 8, 1953, pp. 367–385.
11. Bennett, W. R., "Spectra of Quantized Signals," B.S.T.J., *27,* No. 3 (July 1948), pp. 446–472.
12. Schwartz, M., Bennett, W. R., and Stein, S., *Communications Systems and Techniques,* New York: McGraw Hill, Inc., 1966, p. 224.
13. Kazkov, I. Y., "Problems of the Theory of Statistical Linearization and its Applications," Proc. First Int. Congress, Int. Federation of Automatic Control, Moscow, 1960, pp. 717–723.
14. Smith, H. W., *Approximate Analysis of Randomly Excited Nonlinear Controls,* Cambridge, Massachusetts: MIT Press, 1966.

15. Karafin, B. J., "A Sampled-Data Simulation Language," Chapter 8 in *System Analysis by Digital Computer,* Kuo, F. F. and Kaiser, F. F., eds., New York: John Wiley and Sons, 1966, pp. 286–312.
16. Baikovskii, V. M., "Application of the Describing Function Method to the Study of Automatic Unitary Code Systems," Automation and Remote Control, *26,* No. 9 (September 1966), pp. 1494–1504.
17. Protonotarios, E. N., unpublished work.
18. Booton, R. C., Jr., "Nonlinear Control Systems with Random Inputs," Trans. of IRE Professional Group on Circuit Theory, *CT-1,* No. 9 (September 1954).
19. Blackman, R. B., and Tukey, J. W., *The Measurement of Power Spectra,* New York: Dover Publications, Inc., 1958.

Delta Modulation Granular Quantizing Noise

By DAVID J. GOODMAN

(Manuscript received September 25, 1968)

We present a statistical analysis of a single integration delta modulation system in which slope overload effects are negligible. In defining the delta modulation signal ensemble, we identify a binary phase parameter and show that when this parameter is random, the signal statistics are stationary, provided the input is stationary. Thus the delta modulation correlation functions depend on a single time variable and have Fourier transforms that are the power spectra of the delta modulation signals.

After deriving the delta modulation correlation statistics and power density spectra, we use these functions to investigate the properties of the delta modulation granular quantizing noise. We demonstrate the ratio of input signal power to the quantizing noise power of three signals that approximate the system input. These signals are the integrated delta modulation signal, the signal at the output of the ideal low-pass interpolation filter usually considered in delta modulation studies, and the signal at the output of the optimum interpolation filter. We determine the properties of this filter by referring to the derived spectral density functions.

I. BACKGROUND

Delta modulation (ΔM) systems are subject to two types of quantizing distortion, generally referred to as granular quantizing noise and slope overload noise. The overload noise arises when the analog input to the delta modulator changes at a rate greater than the maximum average slope of the signal generated in the delta modulator feedback loop. The granular noise is analogous to pulse code modulation (PCM) quantizing noise; it arises because the ΔM signal is a discrete-time discrete-amplitude representation of a continuous-amplitude process.

After the discovery of ΔM in the early 1950's, two statistical analyses of distortion effects appeared.[1] Van de Weg considered a delta modulator, constrained so that slope overload effects are negligible, and analyzed the effects of granular quantizing noise in a manner that paralleled Bennett's analysis of quantizing noise effects in a pulse code modulation (PCM) system constrained to be free of overload.[2,3] Zetterberg, in 1955, published a study of both types of distortion as part of an extensive mathematical analysis of the ΔM process.[4] Zetterberg's expression for granular noise power is less precise than van de Weg's. His results pertaining to slope overload have recently been revised.[5]

Eleven years after the appearance of Zetterberg's paper an independent analysis of slope overload noise was published by O'Neal whose effort was supported by S. O. Rice.[6] O'Neal used van de Weg's formula to predict the granular noise power but obtained slope overload characteristics that differed from those derived by Zetterberg. The reason for the two solutions to the same problem is investigated in a recent paper by Protonotarios.[5] This paper gives new expressions for the slope overload noise that are more accurate than any previously obtained. Like O'Neal, Protonotarios uses van de Weg's characterization of the granular quantization effects.

Although van de Weg's formula for granular quantizing noise power has been experimentally verified over an important range of operating conditions, his statistical characterization is inadequate for certain analytical purposes. A principal difficulty in this characterization is the nonstationarity of the ΔM signal ensemble. Because the statistics are nonstationary it is not possible to calculate correlation coefficients by Fourier transformation of the power spectral density function, derived by van de Weg as a mean square amplitude spectrum.

To admit the techniques of stationary time series analysis to the study of ΔM signals, we generalize the signal ensemble by defining a binary phase parameter. We derive correlation statistics directly as average products and show that if the phase is random with both values equiprobable, the ensemble is stationary. Thus we are able to compute power density spectra as Fourier transforms of the correlation functions and to compare the new formula for granular quantizing noise with that given by van de Weg. We find that over the range of operating speeds considered by van de Weg and O'Neal that van de Weg's formula is a good approximation to the one presented here. For very low speeds van de Weg's approximations break down while the formulas we present in this paper are applicable to all ΔM sampling rates.

An additional advantage of this analysis is the presentation of cross-correlation statistics and the cross-power spectrum of the ΔM signal and the analog waveform it represents. We use the cross-power spectrum to derive the transfer function of the optimal interpolation filter for ΔM. We compare the output noise power of this filter with that of the ideal low-pass filter usually considered in ΔM studies. The correlation statistics presented here have also been used in the synthesis of optimal digital filters.[7]

II. THE ΔM SYSTEM

The delta modulator shown in Fig. 1 transforms the continuous signal $y(t)$ to the binary sequence

$$\cdots , b_{-1} , b_0 , b_1 , \cdots$$

in which b_n may have the value $+1$ or -1. The modulator generates binary symbols at τ second intervals according to the sign of $e(t)$, the error signal. This error is the difference of $y(t)$ and $x(t)$, the integrated ΔM signal generated in the modulator feedback loop. The term $x(t)$ is the integral of the binary impulses weighted by the "step size," δ. Thus $x(t)$ has a step of $+\delta$ or $-\delta$ at each sampling instant and is otherwise constant. At the ΔM receiver, this integrated ΔM signal is recovered by a replica of the modulator feedback loop and an analog signal, $\hat{y}(t)$, is generated by means of the interpolating filter with impulse response $h(\cdot)$. The signal $\hat{y}(t)$ is an approximation to the system input, and in this paper the fidelity of the ΔM system will be measured by the mean square error,

$$\eta = E\{[y(t) - \hat{y}(t)]^2\}, \tag{1}$$

in which $E\{\cdot\}$ is the expectation operator. We assume that the binary signal processed by the receiver is identical to the one generated at the modulator. The effects of transmission errors are not considered.

The two ΔM parameters are τ, the sampling interval, and δ, the step size. The quantizing distortion decreases monotonically with increasing

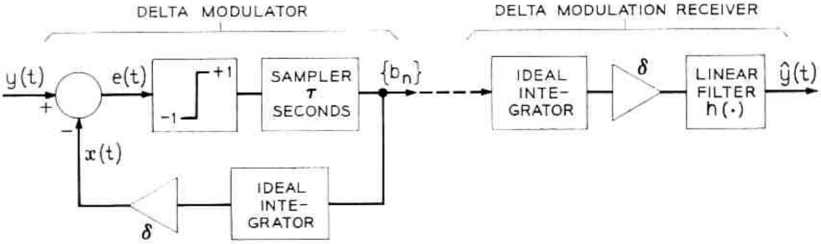


Fig. 1 — The delta modulation system.

sampling rate, $f_s = 1/\tau$, while for a fixed rate the value of the step size determines the mix of granular quantizing noise and slope overload noise in the quantizing noise signal, $y(t) - \hat{y}(t)$. In this paper we consider only the granular quantizing noise; thus we postulate a system in which δ is set such that δ/τ, the maximum average slope of $x(t)$, is exceeded by the slope of $y(t)$ with very low probability. To serve this aim we follow van de Weg and establish the condition that δ/τ is four times the root mean square slope of $y(t)$. This condition is analogous to the "4σ loading" assumed by Bennett in his analysis of a PCM system with negligible overload effects.[3] For gaussian signals, the probability that the slope of $y(t)$ is greater than δ/τ is less than 4×10^{-5}.

If $y(t)$ is a sample function of a stationary stochastic process, the stated design condition may be expressed in terms of $S_{yy}(f)$, the power spectral density of the process. The important parameters of $S_{yy}(f)$ are its average,

$$\sigma^2 = \int_{-\infty}^{\infty} S_{yy}(f)\, df = E\{[y(t)]^2\}, \tag{2}$$

the mean square signal, and its effective bandwidth,[8]

$$f_e = \left[\frac{\int_{-\infty}^{\infty} f^2 S_{yy}(f)\, df}{\int_{-\infty}^{\infty} S_{yy}(f)\, df} \right]^{\frac{1}{2}}. \tag{3}$$

The rms slope of $y(t)$ is $2\pi\sigma f_e$. Thus the condition that the maximum average slope of $x(t)$ equal four times the rms slope of $y(t)$ may be expressed as

$$\delta/\tau = 8\pi\sigma f_e$$

or

$$\beta = \delta/\sigma = 8\pi f_e \tau = 8\pi/F \tag{4}$$

in which we have related the ΔM parameters to the important signal parameters. Thus, β is the step size as a multiple of the rms signal and $F = f_s/f_e$ is the sampling rate as a multiple of the effective bandwidth.

Equation (4) establishes β for each sampling rate; in the analysis of granular quantizing noise to be presented, it is the sampling rate that is considered to be the independent variable of the ΔM system. Studies of slope overload indicate that for minimal total quantizing noise, βF, instead of remaining constant as it does here, should increase with increasing F.[5,6] In the numerical examples given by O'Neal and by Protonotarios, the value of βF that results in minimal total quantizing noise approximates 8π for the highest sampling rate considered.

III. THE SCOPE OF THE ANALYSIS

The signals processed in the ΔM system have been analyzed as realizations of discrete-time (sampled-data) random processes. The transmitted binary sequence, $\{b_n\}$, the integrated ΔM signal, $x(t)$, and the analog output, $\hat{y}(t)$, are all determined by the values of the analog input at the sampling instants, $n\tau$ ($n = \cdots, -1, 0, 1, \cdots$). Thus the analysis reported here consists of derivations of the statistical properties of $\{x_n\} = \{x(n\tau)\}$, the integrated ΔM sequence and $\{e_n\} = \{e(n\tau)\}$, the error sequence, from the statistics of $\{y_n\} = \{y(n\tau)\}$, the input signal sequence.

If $y(t)$ is drawn from a stationary process with auto-covariance function $\sigma^2\rho(\cdot)$ [the Fourier transform of $S_{yy}(f)$], the covariance coefficients of the ΔM signals may be expressed as functions of the statistics, $\rho_n = \rho(n\tau)$. The derived covariance functions are $E\{x_ix_j\}$, the autocovariance of the integrated ΔM signal, and $E\{y_ix_j\}$, the cross-covariance of this signal and the analog input. A property of the definition (in Section 5.1) of the ensemble of sequences $\{x_n\}$ is its stationarity in the wide sense. (Van de Weg considers a somewhat different ensemble, one that has nonstationary statistics.) Thus the covariances are functions of the single time variable, $\mu = j - i$, and we denote them r_μ (the autocovariance) and ϕ_μ (the cross-covariance) respectively. Also of interest is Q_μ, the error covariance function given by

$$Q_\mu = E\{e_ne_{n+\mu}\} = \sigma^2\rho_\mu + r_\mu - \phi_\mu - \phi_{-\mu}.$$

It is shown in Section 5.4 that the covariance statistics, ϕ_μ, are proportional to $\sigma^2\rho_\mu$, the autocovariance of the continuous input. Thus $\phi_\mu = \phi_{-\mu}$ and the error covariance function is given by

$$Q_\mu = \sigma^2\rho_\mu + r_\mu - 2\phi_\mu. \tag{5}$$

Because the processes under consideration are stationary, their power density spectra are Fourier cosine series with coefficients given by the covariance statistics defined above. The spectra are periodic in frequency over intervals of $1/\tau$ Hz; they are denoted with asterisks in keeping with conventions of sampled data analysis. We apply the

Fourier series representation:

$$A^*(f) = a_0 + 2\sum_{n=1}^{\infty} a_n \cos 2\pi n f \tau$$

so that

$$a_n = 2\tau \int_0^{f_s/2} A^*(f) \cos 2\pi n f \tau \, df. \tag{6}$$

In the sequel we will denote these Fourier transform relationships between $A^*(f)$ and a_n by $A^*(f) \longleftrightarrow a_n$.

The power density spectrum, $S^*_{yy}(f)$, of the samples of the analog input is related to $S_{yy}(f)$, the power spectrum of the continuous input signal, by

$$\sigma^2 \rho_\mu \leftrightarrow S^*_{yy}(f) = \frac{1}{\tau} \sum_{n=-\infty}^{\infty} S_{yy}(f + nf_s). \tag{7}$$

It follows that if $y(t)$ is bandlimited to $W < f_s/2$ Hz, there is no aliasing distortion and

$$S^*_{yy}(f) = \frac{1}{\tau} S_{yy}(f), \quad \text{for} \quad |f| < f_s/2. \tag{8}$$

The other transform pairs of interest are $S^*_{xx}(f) \leftrightarrow r_\mu$, $S^*_{ee}(f) \leftrightarrow Q_\mu$, and $S^*_{xy}(f) \leftrightarrow \phi_\mu$. $S^*_{xx}(f)$ and $S^*_{ee}(f)$ are the power spectral density functions of the integrated ΔM signal and the error signal, respectively. $S^*_{xy}(f)$ is the cross-power spectrum of the integrated ΔM signal and the analog input. Equation (5) implies that the four power density spectra are related by

$$S^*_{ee}(f) = S^*_{xx}(f) + S^*_{yy}(f) - 2S^*_{xy}(f). \tag{9}$$

These spectral density functions and $H(f)$, the transfer function of the interpolating filter, determine the value of the output quantizing noise power defined in equation (1).† Thus,

$$\eta = 2\tau \int_0^{f_s/2} \{S^*_{yy}(f) - 2\mathrm{Re}\,[H(f)S^*_{xy}(f)] + |H(f)|^2 S^*_{xx}(f)\} \, df \tag{10}$$

so that the transfer function of the optimal interpolation filter, that is, that which minimizes η, is the (nonrealizable) Wiener filter,[9,10]

† It is assumed here that $H(f)$ processes a sequence of ideal impulses. In Fig. 1 the filter input is a sequence of flat pulses of τ second duration so that when a filter described in this analysis is to be included in a real system, its transfer function should be weighted to compensate for the aperture effect.[3]

$$H_{\text{opt}}(f) = \frac{S^*_{xy}(f)}{S^*_{xx}(f)}, \quad \text{for} \quad |f| \leqq f_s/2 \qquad\qquad (11)$$
$$= 0, \qquad \text{for} \quad |f| > f_s/2.$$

The associated minimal quantizing noise power is

$$\eta_{\min} = 2\tau \int_0^{f_s/2} \left\{ S^*_{vv}(f) - \frac{[S^*_{xy}(f)]^2}{S^*_{xx}(f)} \right\} df. \qquad (12)$$

In previous ΔM studies it was assumed that $y(t)$ is bandlimited to W Hz and that the interpolation is performed by a perfect low pass filter with transfer function

$$H_{lpf}(f) = 1, \quad \text{for} \quad |f| \leqq W$$
$$= 0, \quad \text{for} \quad |f| > W.$$

Equation (10) indicates that the quantizing noise power associated with this filter is

$$\eta_{lpf} = 2\tau \int_0^W [S^*_{vv}(f) + S^*_{xx}(f) - 2S^*_{xy}(f)]\, df$$
$$= 2\tau \int_0^W S^*_{ee}(f)\, df. \qquad (13)$$

Thus the quantizing noise power associated with the low-pass filter is the portion of the power of the error signal that lies within the band of the analog input. By substituting the Fourier series with coefficients Q_μ into equation (13) we arrive at the formula for the low pass filter quantizing noise in terms of the error covariance coefficients:

$$\eta_{lpf} = \frac{1}{R}\left[Q_0 + 2 \sum_{n=1}^{\infty} Q_n \frac{\sin\left(\frac{\pi n}{R}\right)}{\left(\frac{\pi n}{R}\right)} \right] \qquad (14)$$

in which $R = f_s/2W$ is the bandwidth expansion ratio of the ΔM system. It is the ratio of the ΔM sampling rate to the Nyquist sampling rate of the input signal. The ratio, F/R, of the two normalized sampling rates is $2W/f_c$, twice the ratio of the highest frequency spectral component of $y(t)$ to the effective bandwidth.

IV. PRINCIPAL RESULTS

4.1 *Covariance Coefficients*

By means of the formulas of the preceding sections, the characteristics of granular quantizing noise may be expressed in terms of the correlation statistics ρ_μ, r_μ, and ϕ_μ. These quantities depend on the nature of the analog input and on the normalized sampling rate, F. Details of the derivations of r_μ and ϕ_μ, when the input is drawn from a stationary gaussian process, are given in the subsequent sections of this paper. Here we present the covariance formulas and use them to investigate the quantizing noise properties.

As multiples of the mean square input, the autocovariance coefficients of the integrated ΔM signal are

$$\frac{r_0}{\sigma^2} = 1 + 4\sum_{k=1}^{\infty} \exp\left[-\frac{F^2k^2}{32}\right] + \frac{64\pi^2}{F^2}\left\{\frac{1}{3} + \sum_{k=1}^{\infty}\frac{1}{\pi^2k^2}\exp\left[-\frac{F^2k^2}{32}\right]\right\}$$

$$\frac{r_\mu}{\sigma^2} = \rho_\mu\left\{1 + 4\sum_{k=1}^{\infty}\exp\left[-\frac{F^2k^2}{32}\right]\right\} + \frac{64}{F^2}\sum_{m=1}^{\infty}\sum_{k=1}^{\infty}\frac{1}{mk}\left[1 + (-1)^{m+k}\right]$$

$$\cdot\left\{\exp\left[-\frac{F^2(k^2 + m^2 - 2mk\rho_\mu)}{128}\right] - \exp\left[-\frac{F^2(k^2 + m^2 + 2mk\rho_\mu)}{128}\right]\right\}$$

for μ even,

$$\frac{r_\mu}{\sigma^2} = \rho_\mu\left\{1 + 4\sum_{k=1}^{\infty}\exp\left[-\frac{F^2k^2}{32}\right]\right\} + \frac{128}{F^2}\sum_{m=1}^{\infty}\sum_{k=1}^{\infty}\frac{1}{mk}(-1)^m$$

$$\cdot\left\{\exp\left[-\frac{F^2(k^2 + m^2 - 2mk\rho_\mu)}{128}\right] - \exp\left[-\frac{F^2(k^2 + m^2 + 2mk\rho_\mu)}{128}\right]\right\}$$

for μ odd. (15)

The cross-covariance function of $\{x_n\}$ and $\{y_n\}$ is proportional to $\sigma^2\rho_\mu$, the autocovariance function of $\{y_n\}$. Thus

$$\frac{\phi_\mu}{\sigma^2} = c\rho_\mu \tag{16}$$

where

$$c = 1 + 2\sum_{k=1}^{\infty}\exp\left[-\frac{F^2k^2}{32}\right]. \tag{17}$$

Q_μ, the autocovariance of the error signal, is related to ρ_μ, r_μ, and ϕ_μ through equation (5). Therefore

$$\frac{Q_0}{\sigma^2} = \frac{64\pi^2}{F^2}\left\{\frac{1}{3} + \sum_{k=1}^{\infty}\frac{1}{\pi^2k^2}\exp\left[-\frac{F^2k^2}{32}\right]\right\}, \tag{18}$$

$$\frac{Q_\mu}{\sigma^2} = \frac{64}{F^2}\sum_{k=1}^{\infty}\sum_{m=1}^{\infty}\frac{1}{mk}\left[1 + (-1)^{m+k}\right]\left\{\exp\left[-\frac{F^2(k^2 + m^2 - 2mk\rho_\mu)}{128}\right]\right.$$

$$- \exp\left[-\frac{F^2(k^2 + m^2 + 2mk\rho_\mu)}{128}\right]\Bigg\}, \quad \text{for} \quad \mu \text{ even},$$

$$\frac{Q_\mu}{\sigma^2} = \frac{128}{F^2}\sum_{m=1}^{\infty}\sum_{k=1}^{\infty}\frac{1}{mk}(-1)^m\Bigg\{\exp\left[-\frac{F^2(k^2 + m^2 - 2mk\rho_\mu)}{128}\right]$$

$$- \exp\left[-\frac{F^2(k^2 + m^2 + 2mk\rho_\mu)}{128}\right]\Bigg\}, \quad \text{for} \quad \mu \text{ odd}. \tag{19}$$

4.2 *The Minimal Output Quantizing Noise Power*

The proportionality of the autocovariance of the input signal and the cross-covariance of the input and the integrated ΔM signal implies that the related spectra are also proportional: $S^*_{xy}(f) = cS^*_{yy}(f)$. When this relationship is substituted into equation (12), the formula for the quantizing noise power at the output of an optimal interpolation filter, the result is

$$\eta_{\min} = 2\tau\int_0^{f_s/2} S^*_{yy}(f)\left[1 - \frac{c^2S^*_{yy}(f)}{(2c-1)S^*_{yy}(f) + S^*_{ee}(f)}\right]df \tag{20}$$

in which equation (9) has been used to substitute for $S^*_{xx}(f)$. By algebraic manipulation equation (20) may be shown to be identical to

$$\eta_{\min} = \frac{c^2}{(2c-1)^2}\left\{2\tau\int S^*_{ee}(f)\,df - 2\tau\int\frac{[S^*_{ee}(f)]^2}{(2c-1)S^*_{yy}(f) + S^*_{ee}(f)}\,df\right\}$$

$$- \frac{(c-1)^2}{2c-1}\,2\tau\int S^*_{yy}(f)\,df \tag{21}$$

in which the integrals are taken over the set of f in $0 \leqq f \leqq f_s/2$ for which $S^*_{yy}(f) \neq 0$. The third integral in equation (21) is $\sigma^2/2\tau$; if the input is bandlimited to W Hz [with $S^*_{yy}(f) \neq 0$ for $|f| < W$], the first integral is that given in (13), $\eta_{lpf}/2\tau$. It follows that equation (21) may be rewritten as

$$\eta_{\min} = \frac{c^2}{(2c-1)^2}\left[\eta_{lpf} - 2\tau\int_0^W\frac{[S^*_{ee}(f)]^2}{(2c-1)S^*_{yy}(f) + S^*_{ee}(f)}\,df\right]$$

$$- \frac{(c-1)^2}{2c-1}\,\sigma^2. \tag{22}$$

Thus for a bandlimited signal, equation (22) relates the quantizing noise power at the output of a low pass interpolation filter to the noise at the output of an optimal filter. As the sampling rate increases, $c \to 1$ and the integral in equation (22), of a quadratic form of the coefficients, Q_μ, becomes negligible relative to η_{lpf} which is the integral of a linear form.

Thus for high sampling rates, $\eta_{\min} \approx \eta_{lpf}$, indicating that the transfer function of the optimal filter is nearly flat over frequencies at which $S_{\nu\nu}^{*}(f) \neq 0$ and is zero where $S_{\nu\nu}^{*}(f) = 0$.

4.3 *Approximations*

The infinite series in the formulas for the covariance coefficients converge rapidly, and in many cases of practical interest, entire series contribute negligibly to the values of the coefficients. For example, if the input possesses a flat power spectrum, cutoff at W Hz, the effective bandwidth is $W/(3)^{\frac{1}{2}}$ and the normalized sampling rate is related to the bandwidth expansion ratio by $F = 2(3)^{\frac{1}{2}} R$. Thus for $R \geq 12$ ΔM samples per Nyquist interval, the single summations in equations (15), (17), (18) and (19) consist of powers of a e^{-54} or less. These summations are added to 0.25 or to $\pi^2/3$ and thus have negligible effect on the values of the covariance coefficients. In the double summations, only the terms obtained with the two indices equal contribute significantly to the total when F is high. These double summations may, therefore, be replaced by single sums and we have the following approximations:

$$\frac{r_0}{\sigma^2} \approx 1 + \frac{64\pi^2}{3F^2}$$

$$\frac{r_\mu}{\sigma^2} \approx \rho_\mu + \frac{256}{F^2} \sum_{k=1}^{\infty} \frac{(-1)^{\mu k}}{k^2} \exp\left(-\frac{F^2k^2}{64}\right) \sinh\left(\frac{F^2k^2\rho_\mu}{64}\right) \tag{23}$$

$$c \approx 1, \qquad \frac{\dot{\Phi}_\mu}{\sigma^2} \approx \rho_\mu \tag{24}$$

$$\frac{Q_0}{\sigma^2} \approx \frac{64\pi^2}{3F^2} = \beta^2/3$$

$$\frac{Q_\mu}{\sigma^2} \approx \frac{256}{F^2} \sum_{k=1}^{\infty} \frac{(-1)^{\mu k}}{k^2} \exp\left(-\frac{F^2k^2}{64}\right) \sinh\left(\frac{F^2k^2\rho_\mu}{64}\right). \tag{25}$$

If in equation (25) we approximate $\sinh x$ by $e^x/2$† and substitute the result in equation (14) for η_{lpf}, we obtain van de Weg's formula for the granular noise power. Van de Weg claims its validity for $R \geqq 2$ samples per Nyquist interval. Our precise formula for Q_μ, equation (19), leads to noise power characteristics that are valid for all sampling rates.

† This leads to a small but nonzero value of Q_μ as $\mu \to \infty$ and $\rho_\mu \to 0$. Retention of the e^{-x} term in the approximate formula for Q_μ results in $Q_\infty = 0$ and thus avoids an anomaly and a source of numerical error in van de Weg's noise power formula.

4.4 *Signal-to-Noise Ratio Characteristics*

In this section we demonstrate the nature of the derived quantizing noise characteristics by illustrating the effect of the ΔM sampling rate on the quantizing noise powers, η_{lpf} and $\eta_{\min}$, and on Q_0, the mean square error at the input to the interpolation filter. In particular, Fig. 2 shows on a dB scale, $S_{opt} = \sigma^2/\eta_{\min}$, the output signal-to-noise ratio of an optimal interpolation filter; $S_{lpf} = \sigma^2/\eta_{lpf}$, the signal-to-noise ratio of a low pass filter; and $S_0 = \sigma^2/Q_0$ the signal-to-noise ratio prior to interpolation. The data in Fig. 2 pertain to the case of a zero-mean stationary gaussian input with a flat bandlimited power spectrum. The signal-to-noise ratios are shown as functions of R, the number of ΔM samples per Nyquist interval.

For high sampling rates, equation (25) indicates that Q_0 is approximately $\delta^2/3$, the mean square value of a random variable distributed uniformly over an interval of length 2δ. Thus with increasing R, S_0 rises at the rate of 20 dB per decade. At high sampling rates S_{lpf} and S_{opt} are nearly identical. Their slope is 30 dB per decade as indicated by equation (14) which is a linear combination of the error covariance coefficients (proportional to R^{-2}), weighted by $1/R$.

At low sampling rates, S_0 and S_{lpf} become very low (-15 dB at the Nyquist rate) while S_{opt} tends toward unity, corresponding to a filter that generates zero output (the mean input), and thus has a mean square error of σ^2.

V. DERIVATION OF COVARIANCE STATISTICS

Although the ΔM system considered in this paper is identical to the one studied by van de Weg and the values obtained for granular noise power are virtually the same as his over a wide range of transmission speeds, the method of analysis used in obtaining the present results differs considerably from van de Weg's. Van de Weg formulated the ensemble of integrated ΔM signals as a nonstationary process; he was thus unable to compute spectral characteristics from derived covariance statistics. Instead of considering correlation properties, van de Weg began with the amplitude spectrum of a sample function of the integrated ΔM signal ensemble. He then calculated the power density spectrum as the mean square amplitude spectrum.

In the work reported in this paper, the ensemble of integrated ΔM signals is stationary in the wide sense, so that the power spectra are Fourier transforms of the covariance functions whose derivations are described in the remainder of this paper. The difference between van

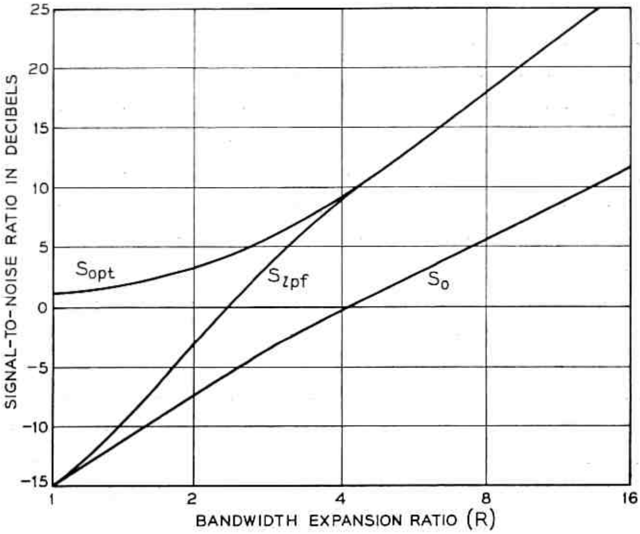


Fig. 2 — Quantizing noise characteristics.

de Weg's signal ensemble and ours lies in the role of the binary phase parameter defined in the Section 5.1.

5.1 *The Integrated ΔM Signal Ensemble*

The integrated ΔM signal, $\{x_n\}$, is a discrete-time discrete-amplitude function. The signal ranges over values $k\delta$ ($k = 0, \pm1, \pm2, \cdots$), and the absence of slope overload implies that x_n takes on the value of the allowed quantization level nearest to y_n. (In overload conditions, x_n and y_n may differ considerably.) At any sampling instant, the set of allowed quantization levels of a given signal is either the odd-parity subset of quantization levels,

$$\pm\delta,\ \pm3\delta,\ \pm5\delta,\ \cdots \tag{26}$$

or the even-parity subset

$$0,\ \pm2\delta,\ \pm4\delta,\ \cdots\ . \tag{27}$$

This restriction to a subset of the $k\delta$ follows from the ΔM mechanism which constrains each sample of $\{x_n\}$ to differ by $\pm\delta$ from its predecessor. Thus if $x_0 = 2k\delta$, $x_1 = (2k \pm 1)\delta$ and any sample that may be written x_{2m} ($m = 0, \pm1, \pm2, \ldots$) is constrained to an even-parity value. Similarly the subsequence $\{x_{2m+1}\}$ ranges over the odd-parity set of quantization levels.

Thus in the absence of slope overload, x_n is the result of processing y_n with a uniform PCM quantizer with quantization intervals of length 2δ. Either x_n is the output of the even-parity quantizer, with levels given by equation (27) or the output of the odd-parity quantizer with levels given in equation (26). The input-output characteristics of the two quantizers are shown in Fig. 3.

In defining the ΔM signal ensemble, van de Weg assumed that the "initial condition," $x_0 = 2k\delta$, applies to all sequences $\{x_n\}$. In van de Weg's analysis, therefore, all samples in $\{x_{2m}\}$ are generated by the even-parity quantizer and all samples in $\{x_{2m+1}\}$ are generated by the odd parity quantizer. Thus the probability functions of x_{2m} and x_{2m+1} differ and the ensemble of sequences $\{x_n\}$ is nonstationary.

We now generalize van de Weg's formulation of the integrated ΔM signal ensemble by observing that the ΔM system may also generate signals with the initial condition, $x_0 = (2k-1)\delta$. In this event $\{x_{2m}\}$ is the output of the odd-parity quantizer of Fig. 3 and $\{x_{2m+1}\}$ is the output of the even-parity quantizer. We shall refer to the initial condition that applies to a given $\{x_n\}$ as the "phase" of the signal. Thus we define the two phase states:

A_1: $\{x_{2m}\}$ generated by the even-parity quantizer
A_2: $\{x_{2m}\}$ generated by the odd-parity quantizer.

A delay of a signal by τ seconds results in a phase reversal from A_1 to A_2 or from A_2 to A_1.

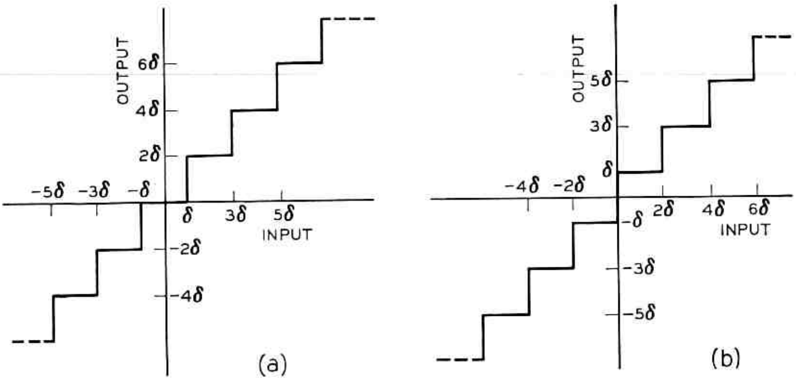


Fig. 3 — Two uniform quantizers: (a) even-parity quantizer, (b) odd-parity quantizer.

If we admit signals with both phases to the ΔM ensemble, the statistics of the ensemble of $\{x_n\}$ depend on the relative frequency of occurrence, that is, on the prior probability of the two phases. Van de Weg's ensemble is a "coherent" one for which the prior probability function is

$$\Pr\{A_1\} = 1, \qquad \Pr\{A_2\} = 0. \tag{28}$$

In this paper we study the statistics of the noncoherent ensemble in which

$$\Pr\{A_1\} = \Pr\{A_2\} = \tfrac{1}{2}. \tag{29}$$

The correlation analysis begins with the derivation of probability functions conditioned on each of the two phases. Marginal probabilities may be calculated on the basis of a prior probability function as

$$\Pr\{x_n = k\delta\} = \Pr\{A_1\}\Pr\{x_n = k\delta \mid A_1\} + \Pr\{A_2\}\Pr\{x_n = k\delta \mid A_2\}. \tag{30}$$

When equation (29) is used in the computation of equation (30), the result is independent of n. Similarly the joint marginal probability of x_n and $x_{n+\mu}$ is independent of n when equation (29) is accepted. When equation (28) is accepted, as it is in van de Weg's analysis, both the single and joint probability functions depend on the parity of n and the covariance statistics are functions of two time variables.

In principle, either equation (28) or (29) may be applicable to the operation of a particular ΔM system. In practice, numerical results based on the two phase conditions are usually quite similar. In analytic work, there is a considerable advantage offered by equation (28), the noncoherence assumption. It admits the techniques of stationary time series analysis to the investigation of questions of interest.

5.2 *The Probability Distribution of* $\mathrm{x_n}$

Here we derive the probability function of a sample, x_n, of the integrated ΔM signal. The probabilities conditioned on A_1 and A_2 depend on whether n is even or odd, but the marginal probability function is independent of n when A_1 and A_2 are equiprobable.

Under the condition A_1, the samples $\{x_{2m}\}$ are outputs of the even-parity quantizer so that $x_{2m} = 2k\delta$ when

$$(2k - 1)\delta \leqq y_{2m} < (2k + 1)\delta.$$

If $\{y_n\}$ is a sample function of a stationary zero-mean gaussian process with variance σ^2, we have

$$\begin{aligned} \Pr\,\{x_{2m} &= 2k\delta \mid A_1\} = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \int_{(2k-1)\delta}^{(2k+1)\delta} \exp\left(-\frac{u^2}{2\sigma^2}\right) du \\ \Pr\,\{x_{2m} &= (2k-1)\delta \mid A_1\} = 0. \end{aligned} \tag{31}$$

The samples $\{x_{2m+1}\}$ are generated by the odd parity quantizer so that

$$\begin{aligned} \Pr\,\{x_{2m+1} &= 2k\delta \mid A_1\} = 0 \\ \Pr\,\{x_{2m+1} &= (2k-1)\delta \mid A_1\} = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \int_{(2k-2)\delta}^{2k\delta} \exp\left(-\frac{u^2}{2\sigma^2}\right) du. \end{aligned} \tag{32}$$

Under the condition A_2, the complementary probability function applies:

$$\begin{aligned} \Pr\,\{x_{2m} = 2k\delta \mid A_2\} &= \Pr\,\{x_{2m+1} = (2k-1)\delta \mid A_2\} = 0, \\ \Pr\,\{x_{2m} = (2k-1)\delta \mid A_2\} &= \Pr\,\{x_{2m+1} = (2k-1)\delta \mid A_1\}, \\ \Pr\,\{x_{2m+1} = 2k\delta \mid A_2\} &= \Pr\,\{x_{2m} = 2k\delta \mid A_1\}. \end{aligned} \tag{33}$$

By combining equations (30) to (33), one may demonstrate that $\Pr\{x_n = k\delta\}$ depends on n (in particular on whether n is even or odd) for all prior probabilities of A_1 and A_2 except the equiprobable pair given in equation (29). Thus equation (29) is a necessary condition for stationarity. When this condition is imposed and $\beta = \delta/\sigma$ incorporated, the formula for the marginal probability of x_n becomes

$$\Pr\,\{x_n = k\delta\} = \frac{1}{2(2\pi)^{\frac{1}{2}}} \int_{(k-1)\beta}^{(k+1)\beta} \exp\left(-\frac{u^2}{2}\right) du. \tag{34}$$

From equation (34), the moments of x_n may be calculated. We have

$$E\{x_n\} = \frac{\delta}{2(2\pi)^{\frac{1}{2}}} \sum_{k=-\infty}^{\infty} k \int_{(k-1)\beta}^{(k+1)\beta} \exp\left(-\frac{u^2}{2}\right) du = 0 \tag{35}$$

and

$$E\{x_n^2\} = r_0 = \frac{\delta^2\beta}{(2\pi)^{\frac{1}{2}}} \sum_{k=-\infty}^{\infty} k^2 \int_0^1 \exp\left[-\frac{\beta^2}{2}(v+k)^2\right] dv, \tag{36}$$

which is equivalent to the form of r_0 given in equation (15). The derivation of equation (15) from (36) is demonstrated in Section A.2 of the appendix.

5.3 *The Joint Probability of* x_n *and* $\mathrm{x}_{n+\mu}$

For each phase condition, the expression of the joint conditional probability of x_n and $x_{n+\mu}$ depends on the parity of n and the parity of μ. For phase A_1, x_n and $x_{n+\mu}$ are both generated by the even-parity quantizer when n and μ are even numbers. Thus the conditional probability that $x_n = 2k\delta$ and $x_{n+\mu} = 2l\delta$ is the probability that

$$(2k - 1)\delta \leqq y_n < (2k + 1)\delta \quad \text{and} \quad (2l - 1)\delta \leqq y_{n+\mu} < (2l + 1)\delta.$$

Thus for n and μ both even,

$$\begin{aligned} \Pr \{x_n &= 2k\delta,\, x_{n+\mu} = 2l\delta \mid A_1\} \\ &= \frac{1}{2\pi\sigma^2(1 - \rho_\mu^2)^{\frac{1}{2}}} \int_{(2k-1)\delta}^{(2k+1)\delta} \int_{(2l-1)\delta}^{(2l+1)\delta} \exp\left[- \frac{u^2 + v^2 - 2\rho_\mu uv}{2\sigma^2(1 - \rho_\mu^2)} \right] du\, dv \\ \Pr \{x_n &= (2k - 1)\delta,\, x_{n+\mu} = l\delta \mid A_1\} \\ &= \Pr \{x_n = k\delta,\, x_{n+\mu} = (2l - 1)\delta \mid A_1\} = 0. \end{aligned} \tag{37}$$

Similarly we derive conditional probability expressions for the eight cases listed under step 1 in Table I. The four marginal probabilities indicated under step 2 are calculated as

$$\begin{aligned} \Pr \{x_n = k\delta,\, x_{n+\mu} = l\delta\} &= \tfrac{1}{2} \Pr \{x_n = k\delta,\, x_{n+\mu} = l\delta \mid A_1\} \\ &\quad + \tfrac{1}{2} \Pr \{x_n = k\delta,\, x_{n+\mu} = l\delta \mid A_2\}. \end{aligned} \tag{38}$$

Among the four cases there are only two different formulas. One is applicable to even values of μ and the other to odd values of μ. When μ is even, x_n and $x_{n+\mu}$ are generated by the same quantizer and when μ is odd they are generated by different quantizers. The marginal joint probability function is independent of n. It may be expressed in terms of the double integral expression

$$\begin{aligned} p(k, l, \mu) = \frac{1}{4\pi(1 - \rho_\mu^2)^{\frac{1}{2}}} &\int_{(k-1)\beta}^{(k+1)\beta} \exp\left(- \frac{v^2}{2} \right) \\ &\cdot \int_{(l-1)\beta}^{(l+1)\beta} \exp\left[- \frac{(u - v\rho_\mu)^2}{2(1 - \rho_\mu^2)} \right] du\, dv \end{aligned} \tag{39}$$

as

$$\begin{aligned} \Pr \{x_n = k\delta,\, x_{n+\mu} = l\delta\} &= p(k, l, \mu) \quad \text{for} \quad k + l + \mu \text{ even} \\ &= 0 \qquad\qquad\; \text{for} \quad k + l + \mu \text{ odd}. \end{aligned} \tag{40}$$

TABLE I—STEPS IN DERIVING $\Pr\{x_n = k\delta,\ x_{n+\mu} = l\delta\}$

Step 1 Conditional probabilities obtained for cases	Step 2 Marginal probabilities	Step 3 Identical expressions except for cases
n even, μ even A_1	n even, μ even	μ even
n even, μ even A_2		
n odd, μ even A_1	n odd, μ even	
n odd, μ even A_2		
n even, μ odd A_1	n even, μ odd	μ odd
n even, μ odd A_2		
n odd, μ odd A_1	n odd, μ odd	
n odd, μ odd A_2		

The autocovariance coefficient, r_μ, is the expected product of x_n and $x_{n+\mu}$:

$$r_\mu = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} (k\delta)(l\delta) \Pr\{x_n = k\delta,\ x_{n+\mu} = l\delta\}. \tag{41}$$

Substitution of equation (40) into (41) results in

$$\begin{aligned} r_\mu &= \delta^2 \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} (2k)(2l)p(2k,\ 2l,\ \mu) \\ &\quad + \delta^2 \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} (2k-1)(2l-1)p(2k-1,\ 2l-1,\ \mu) \quad \text{for } \mu \text{ even} \\ r_\mu &= 2\delta^2 \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} (2k)(2l-1)p(2k,\ 2l-1,\ \mu) \qquad \text{for } \mu \text{ odd.} \end{aligned} \tag{42}$$

Section A.3 of the appendix outlines the derivation of equation (15) from (42) and (39).

5.4 *The Joint Distribution of* y_n *and* $x_{n+\mu}$

Here we consider the joint probability function of a discrete random variable, $x_{n+\mu}$ and a continuous random variable y_n. Once again the marginal distributions are independent of n when the two phases are equiprobable. For $\mu = 0$, the marginal probability function is

$$\begin{aligned}\Pr\,\{y_n = u,\, x_n = k\delta\} &= \frac{1}{2\sigma(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{u^2}{2\sigma^2}\right) du \\ &\qquad \text{for} \quad (k-1)\delta \leqq u < (k+1)\delta \\ &= 0 \quad \text{for other values of } u.\end{aligned} \tag{43}$$

The expected value of $x_n y_n$ may be computed as

$$E\{x_n y_n\} = \phi_0 = \frac{\delta\sigma}{2(2\pi)^{\frac{1}{2}}} \sum_{k=-\infty}^{\infty} k \int_{(k-1)\beta}^{(k+1)\beta} u \exp\left(-\frac{u^2}{2}\right) du \tag{44}$$

which is shown in Section A.4 of appendix to be $c\sigma^2$, with c given by equation (17).

For other values of μ, the conditional probability function of u and $k\delta$ is the probability that $y_n = u$ and $(k-1)\delta \leqq y_{n+\mu} < (k+1)\delta$, provided $k\delta$ is an output level of the quantizer that processes $y_{n+\mu}$. The marginal probability function may be written as

$$\begin{aligned}&\Pr\,\{y_n = u,\, x_{n+\mu} = k\delta\} \\ &= \frac{1}{4\pi\sigma^2(1-\rho_\mu^2)^{\frac{1}{2}}} \int_{(k-1)\delta}^{(k+1)\delta} \exp\left[-\frac{u^2 + v^2 - 2uv\rho_\mu}{2\sigma^2(1-\rho_\mu^2)}\right] dv\, du\end{aligned} \tag{45}$$

from which cross-covariance coefficient ϕ_μ may be calculated as

$$\phi_\mu = \sum_{k=-\infty}^{\infty} k\delta \int_{-\infty}^{\infty} u \Pr\,\{y_n = u,\, x_{n+\mu} = k\delta\}\,. \tag{46}$$

If equation (45) is substituted into (46) and the integration with respect to u is performed first, the result is

$$\phi_\mu = \frac{\rho_\mu \delta\sigma}{2(2\pi)^{\frac{1}{2}}} \sum_{k=-\infty}^{\infty} k \int_{(k-1)\beta}^{(k+1)\beta} v \exp\left(-\frac{v^2}{2}\right) dv \tag{47}$$

which is equation (44) multiplied by ρ_μ.

APPENDIX

Applications of the Poisson Sum Formula to the Derivation of Covariance Coefficients

A.1 *Basic Formula*[11]

$$f(x,\, t) = \sum_{n=-\infty}^{\infty} \exp\,[-t(x+n)^2]. \tag{48}$$

$$f(x,\, t) = \left(\frac{\pi}{t}\right)^{\frac{1}{2}} \left[1 + 2\sum_{k=1}^{\infty} \exp\left(-\frac{\pi^2 k^2}{t}\right) \cos 2\pi k x\right]. \tag{49}$$

A.1.1 *Even Terms*

$$g(x,\,t) = \sum_{n=-\infty}^{\infty} \exp\,[-t(x+2n)^2] = f\!\left(\frac{x}{2},\,4t\right) \tag{50}$$

$$g(x,\,t) = \frac{1}{2}\left(\frac{\pi}{t}\right)^{\frac{1}{2}}\left[1 + 2\sum_{k=1}^{\infty} \exp\left(-\frac{\pi^2k^2}{4t}\right)\cos \pi kx\right]. \tag{51}$$

A.1.2 *Odd Terms*

$$h(x,\,t) = \sum_{n=-\infty}^{\infty} \exp\,[-t(x+2n-1)^2] = f(x,\,t) - g(x,\,t) \tag{52}$$

$$h(x,\,t) = \frac{1}{2}\left(\frac{\pi}{t}\right)^{\frac{1}{2}}\left[1 + 2\sum_{k=1}^{\infty} (-1)^k \exp\left(-\frac{\pi^2k^2}{4t}\right)\cos \pi kx\right]. \tag{53}$$

A.2 *Mean Square Value of* x_n

Equation (36) may be developed in terms of the partial derivatives of equation (48):

$$f_t(x,\,t) = -\sum_{n=-\infty}^{\infty} (x+n)^2 \exp\,[-t(x+n)^2] \tag{54}$$

and

$$f_x(x,\,t) = -2t\sum_{n=-\infty}^{\infty} (x+n)\exp\,[-t(x+n)^2]. \tag{55}$$

Equations (54) and (55) may be combined to form

$$\sum_{n=-\infty}^{\infty} n^2 \exp\,[-t(x+n)^2] = x^2 f(x,\,t) + \frac{x}{t}\,f_x(x,\,t) - f_t(x,\,t). \tag{56}$$

If the order of summation and integration in equation (36) is reversed, the resulting integrand is identical in form to the left side of equation (56). Thus equation (49) may be substituted into the right side of equation (56) and the three terms integrated over $0 \leqq x \leqq 1$. The result is

$$\begin{aligned}\int_0^1 \sum_{n=-\infty}^{\infty} n^2 \exp\,[-t(x+n)^2]\,dx \\ = \frac{1}{2t}\left(\frac{\pi}{t}\right)^{\frac{1}{2}}\left[1 + 4\sum_{k=1}^{\infty}\exp\left(-\frac{\pi^2k^2}{t}\right)\right] \\ + \left(\frac{\pi}{t}\right)^{\frac{1}{2}}\left[\frac{1}{3} + \sum_{k=1}^{\infty}\left(\frac{1}{\pi k}\right)^2 \exp\left(-\frac{\pi^2k^2}{t}\right)\right].\end{aligned} \tag{57}$$

The variable, t, in equation (57) is related to equation (36) by $t =$

$\beta^2/2 = 32\pi^2/F^2$; when this latter form is substituted for t, the form of r_0 given in equation (15) results.

A.3 *Autocovariance Coefficients*

In order to illustrate the derivation of equation (15) for r_μ from equation (42), we consider odd values of μ. By substituting into equation (42) the form of $p(k, l, \mu)$ given in equation (39) we write

$$r_\mu = \frac{2\delta^2}{4\pi(1 - \rho_\mu^2)^{\frac{1}{2}}} \sum_{k=-\infty}^{\infty} 2k \int_{(2k-1)\beta}^{(2k+1)\beta} \exp\left(-\frac{v^2}{2}\right) G(v)\, dv, \tag{58}$$

in which we have defined

$$G(v) = \beta \int_{-1-v\rho_\mu/\beta}^{1-v\rho_\mu/\beta} \sum_{l=-\infty}^{\infty} (2l - 1) \exp\left[-\frac{\beta^2(x + 2l - 1)^2}{2(1 - \rho_\mu^2)}\right] dx. \tag{59}$$

The integrand in equation (59) is related to the infinite series in equation (52) and its partial derivative with respect to x by

$$\sum_{n=-\infty}^{\infty} (2n - 1) \exp\left[-t(x + 2n - 1)^2\right] = -xh(x, t) - \frac{1}{2t} h_x(x, t), \tag{60}$$

in which the variable $t = \beta^2/2(1 - \rho_\mu^2)$. Into equation (60) we substitute the form of $h(x, t)$ given in equation (53) and perform the integration required in equation (59). The integral of the second term is zero so that $G(v)$ is β times the integral of the first term of equation (60). Thus equation (58) may be written in the form

$$G(v) = [2\pi(1 - \rho_\mu^2)]^{\frac{1}{2}} \cdot \left\{\frac{\rho_\mu v}{\beta} + 2 \sum_{m=1}^{\infty} \frac{1}{\pi m} \exp\left[-\frac{\pi^2 m^2(1 - \rho_\mu^2)}{2\beta^2}\right] \sin \frac{\pi m \rho_\mu v}{\beta}\right\}, \tag{61}$$

which must be weighted by $\exp(-v^2/2)$ and integrated according to equation (58).

Equations (58) and (61) thus show r_μ to be the sum of two terms. The first term consists of a constant, $\rho_\mu\sigma\delta/(2\pi)^{\frac{1}{2}}$, multiplying the sum

$$\sum_{k=-\infty}^{\infty} 2k \int_{(2k-1)\beta}^{(2k+1)\beta} v \exp\left(-\frac{v^2}{2}\right) dv = 2 \sum_{k=-\infty}^{\infty} \exp\left[-\frac{\beta^2}{2}(2k - 1)^2\right]. \tag{62}$$

This latter summation is in the form of equation (52) with $x = 0$, $t = \beta^2/2$ so that with the application of equation (53), (62) becomes

$$h\left(0, \frac{\beta^2}{2}\right) = \frac{(2\pi)^{\frac{1}{2}}}{\beta}\left[1 + \sum_{k=1}^{\infty} (-1)^k \exp\left(-\frac{\pi^2k^2}{2\beta^2}\right)\right]. \tag{63}$$

The second term in the expression for r_μ may be written in the form,

$$\delta^2\left(\frac{2}{\pi}\right)^{\frac{1}{2}} \sum_{m=1}^{\infty} \frac{1}{\pi m} \exp\left[-\frac{\pi^2m^2(1-\rho_\mu^2)}{2\beta^2}\right] \cdot \sum_{k=-\infty}^{\infty} 2k \int_{(2k-1)\beta}^{(2k+1)\beta} \exp\left(-\frac{v^2}{2}\right) \sin\frac{\pi m\rho_\mu v}{\beta}\, dv. \tag{64}$$

If the sine in this expression is developed in exponential form, the summation, ranging over k, in the above expression, has a form similar to the integral and sum in equation (59). If it is analyzed in the manner that $G(v)$ was reduced the following identity may be demonstrated:

$$\sum_{k=-\infty}^{\infty} k \int_{(2k-1)\beta}^{(2k+1)\beta} \exp\left(-\frac{v^2}{2}\right) \sin\frac{\pi m\rho_\mu v}{\beta}\, dv = \left(\frac{\pi}{2}\right)^{\frac{1}{2}} \exp\left(-\frac{\pi^2m^2\rho_\mu}{2\beta^2}\right)\left[\frac{\pi m\rho_\mu}{\beta^2} + 2\sum_{k=1}^{\infty} \frac{(-1)^k}{\pi k} \times \exp\left(-\frac{\pi^2k^2}{2\beta^2}\right) \sinh\left(\frac{\pi^2km\rho_\mu}{\beta^2}\right)\right]. \tag{65}$$

Thus equation (64) becomes

$$2\rho_\mu\sigma^2 \sum_{m=1}^{\infty} \exp\left(-\frac{\pi^2m^2}{2\beta^2}\right) + 2\delta^2 \sum_{k=1}^{\infty}\sum_{m=1}^{\infty} \frac{(-1)}{\pi^2km} \cdot \exp\left[-\frac{\pi^2(k^2+m^2)}{2\beta^2}\right] \sinh\left(\frac{\pi^2km\rho_\mu}{\beta^2}\right) \tag{66}$$

so that r_μ for μ odd, the sum of (66) and $\rho_\mu\sigma\delta/(2\pi)^{1/2}$ times (63), may be expressed as

$$r_\mu = \rho_\mu\sigma^2\left[1 + 4\sum_{k=1}^{\infty} \exp\left(-\frac{2\pi^2k^2}{\beta^2}\right)\right] + 2\delta^2 \sum_{k=1}^{\infty}\sum_{m=1}^{\infty} \frac{(-1)^m}{\pi^2km} \cdot \exp\left[-\frac{\pi^2(k^2+m^2)}{2\beta^2}\right] \sinh\left(\frac{\pi^2km\rho_\mu}{\beta^2}\right). \tag{67}$$

If $8\pi/F = \beta^2$ is substituted in equation (67) the result is equation (15).

Similarly the formula given in equation (42) for r_μ when μ is even may be developed to demonstrate its identity to the formula in equation (15).

A.4 *Cross-Covariance*

Performing the integration indicated in equation (44) we have

$$\begin{aligned}\phi_0 &= \frac{\delta\sigma}{2(2\pi)^{\frac{1}{2}}} \sum_{k=-\infty}^{\infty} k\left\{\exp\left[-\frac{\beta^2}{2}(k-1)^2\right] - \exp\left[-\frac{\beta^2}{2}(k+1)^2\right]\right\} \\ &= \frac{\delta\sigma}{(2\pi)^{\frac{1}{2}}} \sum_{k=-\infty}^{\infty} \exp\left[-\frac{\beta^2}{2}k^2\right], \qquad (68)\end{aligned}$$

which is equivalent to equation (48) with $x = 0$, $t = \beta^2/2 = [32\,\pi^2/F^2]$. Thus equation (49) may be substituted with the result given in equation (16):

$$\phi_0 = \sigma^2\left[1 + 2\sum_{k=1}^{\infty} \exp\left(-\frac{2\pi^2 k^2}{\beta^2}\right)\right].$$

REFERENCES

1. de Jager, F., "Delta Modulation, A method of PCM Transmission Using the 1-Unit Code," Philips Res. Rep., *7,* No. 6 (December 1962), pp. 442–466.
2. van de Weg, H., "Quantizing Noise of a Single Integration Delta Modulation System with an *N*-Digit Code," Philips Res. Rep., *8,* No. 5 (October 1953), pp. 367–385.
3. Bennett, W. R., "Spectra of Quantized Signals," B.S.T.J. *27,* No. 3 (July 1948), pp. 446–472.
4. Zetterberg, L. H., "A Comparison Between Delta and Pulse Code Modulation," Ericsson Technics, *11,* No. 1 (1955), pp. 95–154.
5. Protonotarios, E. N., "Slope Overload Noise in Differential Pulse Code Modulation," B.S.T.J., *46,* No. 9 (November 1967), pp. 2119–2162.
6. O'Neal, J. B., "Delta Modulation Quantizing Noise Analytical and Computer Simulation Results for Gaussian and Television Input Signals," B.S.T.J., *45,* No. 1 (January 1966), pp. 117–141.
7. Goodman, D. J., "The Application of Delta Modulation to Analog-to-PCM Encoding," *48,* No. 2 (February 1969), pp. 321–343.
8. Gabor, D., "Theory of Communications," J. IEE, *93,* No. 26, part III (November 1946), pp. 429–457.
9. Ruchkin, D. S., "Linear Reconstruction of Quantized and Sampled Random Signals," IRE Trans. Commun. Syst., *CS-9,* No. 4 (December 1961), pp. 350–355.
10. Cox, D. R. and Miller, H. D., *The Theory of Stochastic Processes,* New York: Wiley, 1965, pp. 325–330.
11. Bellman, R., *A Brief Introduction to Theta Functions,* New York: Holt, Rinehart and Winston, 1961, pp. 7–12.

A System Approach to Quantization and Transmission Error

By M. M. BUCHNER, JR.

(Manuscript received October 10, 1968)

In a system designed to quantize the output of an analog data source and to transmit this information over a digital channel, errors are introduced by the quantization and transmission processes. Quantization resolution can be improved by using all positions available in a data stream to carry information, or transmission accuracy can be improved if some of the positions are used for redundancy with error-correcting codes. The problem is to determine, from a system viewpoint, the proper allocation of the available positions in order to reduce the average system error rather than concentrate exclusively on either the quantization problem or the transmission problem.

We develop a criterion for the performance of data transmission systems based upon the numerical error that occurs between the analog source and the destination. The criterion, termed the average system error, is used to evaluate and compare possible system configurations. Significant-bit packed codes are defined. These codes are useful because their protection can be matched to the numerical significance of the data and their redundancy can be sufficiently small to maintain good quantization resolution. The average system error resulting from representative system designs is numerically evaluated and compared.

I. INTRODUCTION

When designing a system to sample the output of an analog data source and to transmit the samples over a digital channel, the usual approach is to consider the errors introduced by quantization and transmission as separate problems. However, from a system viewpoint, a conflict arises. On the one hand, the quantization resolution can be improved by using all of the available positions in a data stream to carry information. Alternatively, the transmission accuracy can be improved if redundancy and error-correcting codes are introduced by converting some of the information positions into parity check positions. The problem then is to determine the proper allocation of the available symbols in order to reduce the average system error rather than concentrate exclusively on either the quantization problem or the transmission problem.

We consider a data transmission system with uniform quantization. The average absolute error that occurs between the analog source and the destination is used as the criterion of system performance. The criterion, termed the average system error (ASE), is used to evaluate and compare the effectiveness of various systems.

Some work has been done on the design of error-correcting codes which provide different amounts of protection for different positions within a code word. In Ref. 1, the general algebraic properties of these codes, referred to as unequal error protection codes, were investigated. In Ref. 2, significant-bit codes (which turn out to be a subclass of unequal error protection codes) and a criterion for evaluating the performance of codes for the transmission of numerical data were developed.

In this paper, we define packed codes and significant-bit packed codes, we analyze their performance, and we numerically evaluate the average system error resulting from the use of representative quatization resolutions and coding schemes.

II. PRELIMINARIES

We consider a binary symmetric channel in which the errors are independent of the symbols actually transmitted. In the numerical examples, we further assume that the errors occur independently with probability $p = 1 - q$. The error-correcting codes to be discussed are binary block codes in which the code vectors form a group under component by component modulo 2 addition. Let n denote the block length and k denote the number of information positions per code vector. The notation (n, k) is used to denote such a code. A complete discussion of these codes is contained in Ref. 3.

The encoder receives k binary information symbols [called a message and denoted by $(v_k, v_{k-1}, \cdots, v_1)$] as an input and determines from the message $(n - k)$ binary parity check symbols. The decoder operates upon the blocks of n binary symbols coming from the channel in an attempt to correct transmission errors and provides k binary symbols at its output.

Let H denote the parity check matrix for such a code. An n-tuple u is a code vector if and only if

$$u\tilde{H} = 0 \tag{1}$$

where $\tilde{H}$ is the transpose of H. The matrix H can be written in the form

$$H = (C_k, C_{k-1}, \cdots, C_1 I_{n-k})$$

where $C_i (1 \leqq i \leqq k)$ is the column of H in the position corresponding to information position v_i in a code vector and I_{n-k} is the $(n - k) \times (n - k)$ identity matrix.

When the integer s is to be sent, the message used is $B_k(s)$ such that*

$$B_k(s) = (v_k, v_{k-1}, \cdots, v_1)$$

where

$$s = \sum_{i=1}^{k} 2^{i-1} v_i .$$

The parity check symbols $E(s)$ are chosen so that the code vector $C(s) = B_k(s) \mid E(s)$ satisfies equation (1) where the symbol $\mid$ indicates that $C(s)$ can be partitioned into $B_k(s)$ and $E(s)$.

III. PACKED CODES

A model of the data transmission system is shown in Fig. 1. Let us assume that each quantization step is of equal size and that there are 2^l quantization levels. For many applications, the quantizer uses a relatively small l (perhaps 15 or less). In addition, coding schemes must have low redundancy; otherwise so many information positions are converted into check positions that the quantization error becomes too large. These requirements lead us to define "packed" codes in the follow-

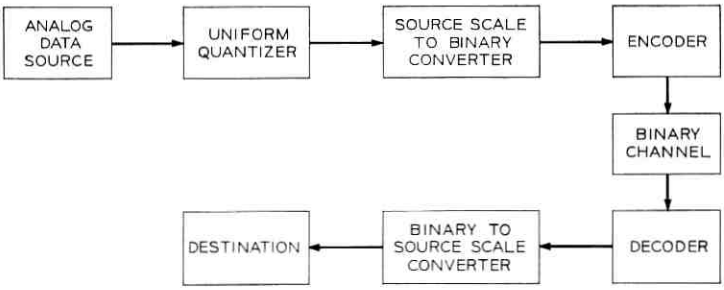


Fig. 1 — System model.

[*] $B_i(j)$ denotes the i-bit binary representation of the integer j where $0 \leqq j \leqq 2^i - 1$.

ing manner. Consider an (n, k) binary group code in which α samples are packed into each code vector. If each sample consists of l bits, then $k = \alpha l$. Let s_m denote the integer that is transmitted for the mth sample in a code vector where $0 \leqq s_m \leqq 2^l - 1$ and $1 \leqq m \leqq \alpha$. Accordingly, the code vector actually transmitted is

$$C(s) = B_l(s_\alpha) \mid B_l(s_{\alpha-1}) \mid \cdots \mid B_l(s_1) \mid E(s)$$

where

$$s = \sum_{m=1}^{\alpha} 2^{(m-1)l} s_m . \tag{2}$$

A packed code vector is shown schematically in Fig. 2.

Two examples are in order. In the first, a (7, 4) perfect single error-correcting code is used to form a packed code with $\alpha = 2$ and $l = 2$.

$$H = \begin{bmatrix} 1\;1 & 1\;0 & \\ 1\;1 & 0\;1 & I_3 \\ \underbrace{1\;0}_{s_2\text{ positions}} & \underbrace{1\;1}_{s_1\text{ positions}} & \end{bmatrix} .$$

In the second example, the idea behind significant-bit codes is applied to packed codes and results in what will be referred to as a significant-bit packed code.[2] Specifically, the basic (7, 4) code can have its protection capabilities arranged to match the numerical significance of the bit positions; that is, to protect the most significant bit of each of four samples ($\alpha = 4$ and $l = 2$).

$$H = \begin{bmatrix} 1\;0 & 1\;0 & 1\;0 & 0\;0 & \\ 1\;0 & 1\;0 & 0\;0 & 1\;0 & I_3 \\ \underbrace{1\;0}_{s_4\text{ positions}} & \underbrace{0\;0}_{s_3\text{ positions}} & \underbrace{1\;0}_{s_2\text{ positions}} & \underbrace{1\;0}_{s_1\text{ positions}} & \end{bmatrix} .$$

Notice that the significant-bit packed code requires only half as many parity check positions per sample as the packed code.

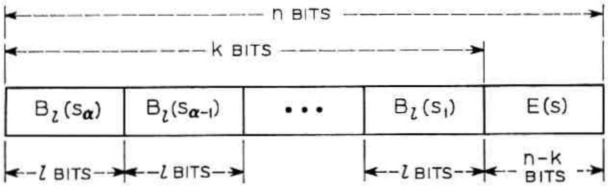


Fig. 2 — Packed code vector.

Many packed codes can be designed to provide desired levels of protection and redundancy. Numerical data concerning the effectiveness of representative packed codes are presented in Sections VI and VII.

IV. FORMULATION OF A CRITERION OF SYSTEM FIDELITY

In this section, we develop a criterion of system fidelity as a function of the number of quantization levels and the capability of the error-correcting code. This is done for packed codes because of their generality.

Let x_m denote the output of the analog source that results in s_m being transmitted. It is assumed that x_m is a random variable that is uniformly distributed on the interval (X_1, X_2). The probability density function for x_m is

$$\begin{aligned} f(x_m) &= \frac{1}{X_2 - X_1} \quad \text{for} \quad X_1 \leqq x_m \leqq X_2 \\ &= 0 \qquad\qquad \text{for} \quad x_m < X_1 \quad \text{or} \quad x_m > X_2 \,. \end{aligned} \tag{3}$$

If

$$X_1 + s_m\left(\frac{X_2 - X_1}{2^l}\right) < x_m < X_1 + (s_m + 1)\left(\frac{X_2 - X_1}{2^l}\right),$$

then the output of the quantizer is

$$X_1 + (s_m + \tfrac{1}{2})\left(\frac{X_2 - X_1}{2^l}\right).$$

The "source scale to binary converter" receives

$$X_1 + (s_m + \tfrac{1}{2})\left(\frac{X_2 - X_1}{2^l}\right)$$

from the quantizer and delivers $B_l(s_m)$ to the encoder. After α samples are received by the encoder, the message

$$B_k(s) = B_l(s_\alpha) \mid B_l(s_{\alpha-1}) \mid \cdots \mid B_l(s_1)$$

is encoded to form the code vector $C(s) = B_k(s) \mid E(s)$ where the value of s is determined from equation (2). At the destination, the decoder attempts to correct errors and provides the message

$$B_k(r) = B_l(r_\alpha) \mid B_l(r_{\alpha-1}) \mid \cdots \mid B_l(r_1)$$

at its output where $0 \leqq r_m \leqq 2^l - 1$ for $1 \leqq m \leqq \alpha$ and

$$r = \sum_{m=1}^{\alpha} 2^{(m-1)l} r_m . \tag{4}$$

The "binary to source scale converter" receives $B_l(r_m)$ and delivers

$$X_1 + (r_m + \tfrac{1}{2})\left(\frac{X_2 - X_1}{2^l}\right)$$

to the destination. Because uniform quantization is used, a useful measure of the numerical error that occurs as a result of the quantization and transmission of x_m is

$$\left| x_m - \left[X_1 + (r_m + \tfrac{1}{2})\left(\frac{X_2 - X_1}{2^l}\right)\right] \right|^\gamma$$

where $\gamma > 0$. The appropriate value of γ will depend upon the nature and use of the signal. For this paper, let $\gamma = 1$.

For the mth sample position in a packed code, let $\Pr_m\{r_m \mid s_m\}$ denote the probability that r_m is received when s_m is sent. Accordingly, the average system error for the mth sample (ASE_m) is

$$\mathrm{ASE}_m = \sum_{r_m=0}^{2^l-1} \sum_{s_m=0}^{2^l-1} \int_{X_1+s_m(X_2-X_1)/2^l}^{X_1+(s_m+1)(X_2-X_1)/2^l} \left| x_m - X_1 - (r_m + \tfrac{1}{2})\left(\frac{X_2 - X_1}{2^l}\right)\right| \cdot \Pr_m\,\{r_m \mid s_m\} f(x_m)\, dx_m . \tag{5}$$

It is desirable to express $\Pr_m\{r_m \mid s_m\}$ in terms of the properties of the error-correcting code. Let $\Pr\{r \mid s\}$ denote the probability that r occurs at the output of the decoder when s is the input to the encoder. As shown in Appendix A, for a channel in which the errors are independent of the symbols actually transmitted,

$$\Pr_m\,\{r_m \mid s_m\} = \underbrace{\sum_{t_\alpha=0}^{2^l-1} \cdots \sum_{t_1=0}^{2^l-1}}_{\text{excluding } t_m} \Pr\left\{ \sum_{m'=1}^{\alpha} 2^{(m'-1)l} t_{m'} \,\middle|\, 0 \right\}$$

where $B_l(t_m) = B_l(r_m) \oplus B_l(s_m)$.* This expression is interesting because it permits us to compute $\Pr_m\{r_m \mid s_m\}$ from the properties of the code. Specifically, it is necessary to determine the probability that each possible sequence of α samples, in which the mth position equals t_m, is received, given that zero is transmitted for each sample, and then to sum these probabilities.

For the case in which one sample is transmitted per code word (that is, $\alpha = 1$ and $l = k$) and all samples are equally likely to be transmitted, the average numerical error (ANE) that occurs during transmission has been defined as[2]

$$\text{ANE} = \frac{1}{2^k} \sum_{r=0}^{2^k-1} \sum_{s=0}^{2^k-1} | r - s | \Pr \{r \mid s\}.$$

The average numerical error is the average magnitude by which the output of the decoder differs numerically from the input to the encoder and thus provides a measure of the performance of the channel and the code. This concept can be generalized by defining the average numerical error for the mth sample as

$$\text{ANE}_m = \frac{1}{2^l} \sum_{r_m=0}^{2^l-1} \sum_{s_m=0}^{2^l-1} | r_m - s_m | \Pr_m \{r_m \mid s_m\}. \tag{6}$$

By reasoning analogous to that in Theorem 1 of Ref. 2, for a binary group code used with a binary symmetric channel, equation (6) can be reduced to

$$\text{ANE}_m = \sum_{i=1}^{l} 2^{i-1} \sum_{r_m=2^{i-1}}^{2^i-1} \Pr_m \{r_m \mid 0\}.$$

With this definition of ANE_m, the probability density function in equation (3), and the steps shown in Appendix B, the average system error for the mth sample, as given in equation (5), can be expressed as

$$\text{ASE}_m = \left(\frac{X_2 - X_1}{2^l}\right)(\text{ANE}_m + \tfrac{1}{4} \Pr_m \{0 \mid 0\}).$$

One feature of packed codes is that the protection afforded various samples against transmission errors can be unequal. If this occurs, different positions will have different system error. Therefore, in general, the average system error per sample (ASE) is

$$\text{ASE} = \frac{1}{\alpha} \sum_{m=1}^{\alpha} \text{ASE}_m .$$

* The symbol $\oplus$ denotes component by component modulo 2 addition of vectors.

The range of the analog source is specified by X_1 and X_2. When considering system design, it is convenient to let $X_2 - X_1 = 1$ (or to consider a normalized average system error). Accordingly, in the remainder of this paper, we shall be concerned with the expression in equation (7).

$$\text{ASE} = \frac{1}{\alpha} \sum_{m=1}^{\alpha} \left[\frac{1}{2^l} (\text{ANE}_m + \tfrac{1}{4} \Pr_m \{0 \mid 0\}) \right]. \tag{7}$$

For a system in which one sample is transmitted per code word (that is, $\alpha = 1$ and $l = k$),

$$\text{ASE} = \frac{1}{2^l} (\text{ANE} + \tfrac{1}{4} \Pr \{0 \mid 0\}) \tag{8}$$

where ANE and $\Pr\{0 \mid 0\}$ are for the entire code.

For error-free transmission, $\Pr_m\{0 \mid 0\} = 1$ and $\text{ANE}_m = 0$ for all coding schemes including uncoded transmission. In this case, $\text{ASE} = 2^{-(l+2)}$. Thus, the system error is independent of the particular code, is minimized by maximizing l, and cannot be reduced to zero but is bounded by the quantization error.

V. THE AVERAGE SYSTEM ERROR FOR UNCODED TRANSMISSION

Before examining the role that error-correcting codes can play in reducing the average system error, it is advantageous to consider system effectiveness when uncoded transmission is used with a memoryless channel. In the system model, uncoded transmission is characterized by $\alpha = 1$ and $l = k = n$. Let ASN_{UC} denote the average system error for uncoded transmission. From Theorem 2 and the comment following the proof of the theorem in Ref. 2 (these are summarized in Appendix C), the average numerical error for uncoded transmission is

$$\text{ANE}_{\text{UC}} = p \sum_{i=1}^{l} 2^{i-1} q^{l-i} = 2^{l-1} p \frac{1 - \left(\frac{q}{2}\right)^l}{1 - \frac{q}{2}}.$$

The probability of correct transmission is q^l. Therefore, from equation (8)

$$\text{ASE}_{\text{UC}} = \frac{1}{2^l} \left(p \sum_{i=1}^{l} 2^{i-1} q^{l-i} + \frac{q^l}{4} \right). \tag{9}$$

Figures 3 and 4 present the average system error for uncoded transmission for representative values of l and p.

For each value of l, notice that as $p \to 0$, $\text{ASE}_{\text{UC}} \to 2^{-(l+2)}$ which is the limitation imposed by the quantization error. Also, ASE_{UC} increases monotonically with p for $0 < p < \frac{1}{2}$ (see Appendix D). For a given value of l, how large must p become so that ASE_{UC} deviates appreciably from $2^{-(l+2)}$ (that is, for what values of p does the transmission error make a significant contribution to the system error?)

For small p, equation (9) yields

$$\text{ASE}_{\text{UC}} \cong \frac{1}{2^l}\left[\left(2^l - 1 - \frac{l}{4}\right)p + \frac{1}{4}\right]. \qquad (10)$$

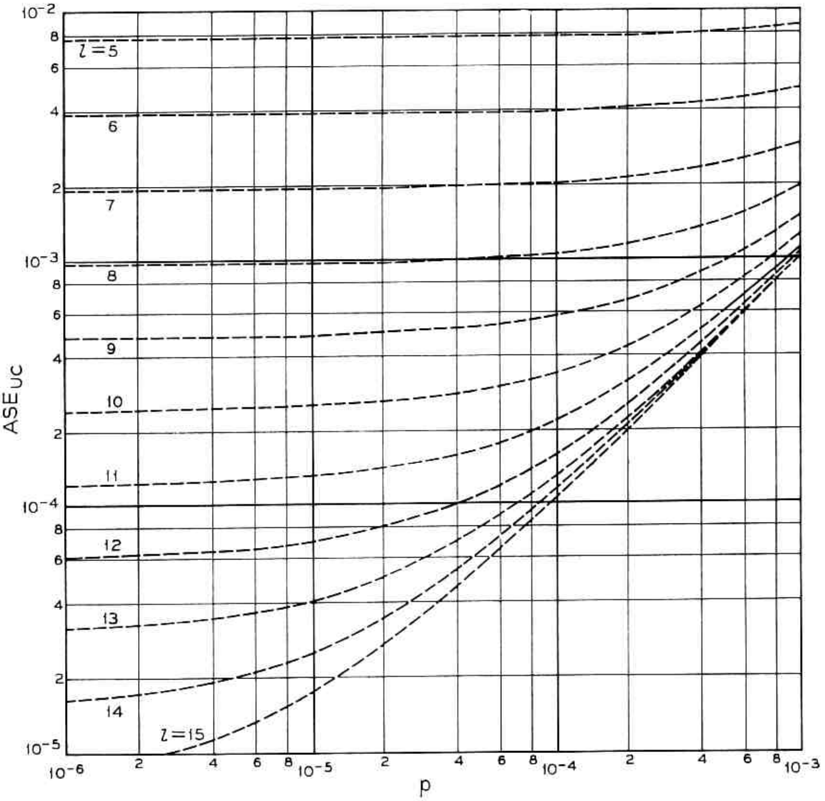


Fig. 3 — Average system error for uncoded transmission (ASE_{UC}) for various l.

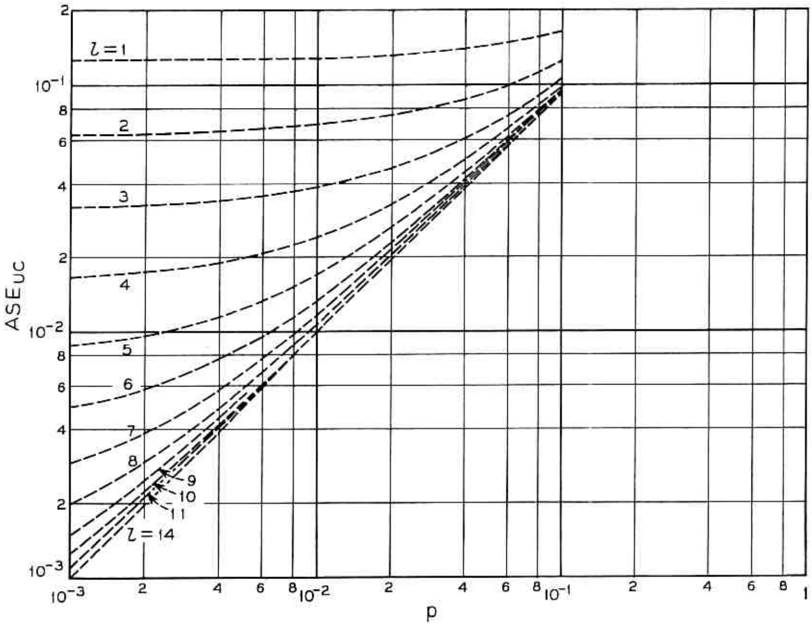


Fig. 4 — Average system error for uncoded transmission (ASE_{UC}) for various l.

This expression can be broken into two components; the term

$$\left[\frac{2^l - 1 - \frac{l}{4}}{2^l}\right]p$$

and the term $2^{-(l+2)}$. These components are shown in Fig. 5 for $l = 15$. In Fig. 5, the terms intersect at a probability of error denoted by p_c where

$$p_e = \frac{1}{4\left(2^l - 1 - \frac{l}{4}\right)}.$$

Notice that p_e is the value of p for which the transmission error equals the quantization error [within the approximations leading to equation (10)]. Accordingly, for $p = p_c$, $\text{ASE}_{\text{UC}} \cong 2^{-(l+1)}$. In Fig. 6, p_e is given for various l. From p_c, it is possible to obtain an estimate of the general region in which ASE_{UC} begins to deviate from $2^{-(l+2)}$ because of transmission errors.

An additional feature of Figs. 3 and 4 is that, for a given value of l and for p greater than the appropriate p_c, ASE_{UC} is approximately equal to p. This causes the converging of the curves as p increases and implies that systems with different l will have essentially the same performance. Let us consider qualitatively the cause of this phenomenon.

For $p > p_c$, the transmission error is significantly greater than the quantization error and, thus, the average system error is largely determined by the transmission error. If a single error occurs in a sample and if it occurs in the most significant position, on the average, a numerical error of ½ will result for any l. For values of p that are of practical interest, the probability that this occurs is essentially independent of l and equal to p. Similar reasoning can be applied to the less significant positions although the numerical error that results will, of course, be less than ½. The point is that the probability that

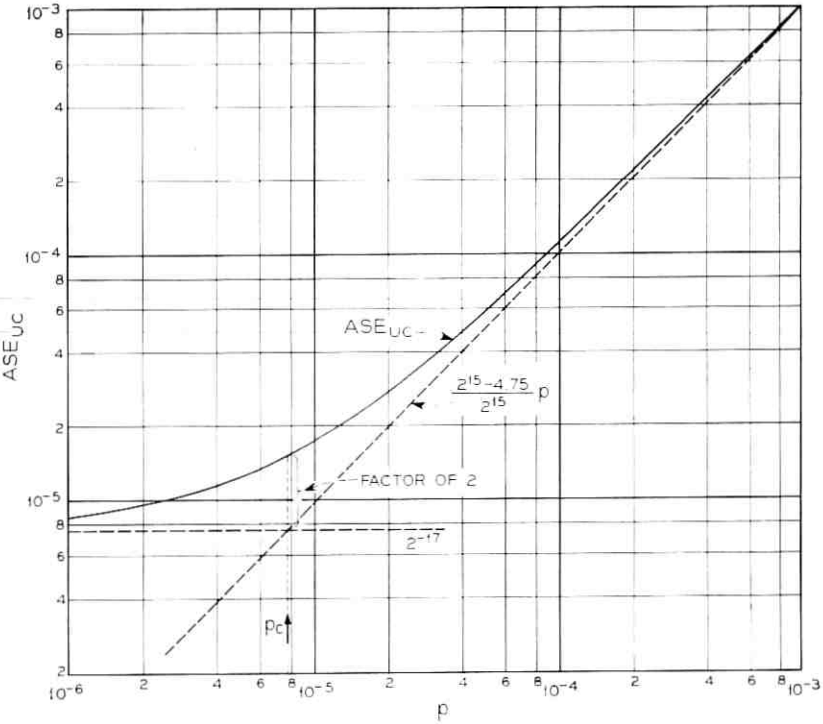


Fig. 5 — Average system error for uncoded transmission (ASE_{UC}) for $l = 15$.

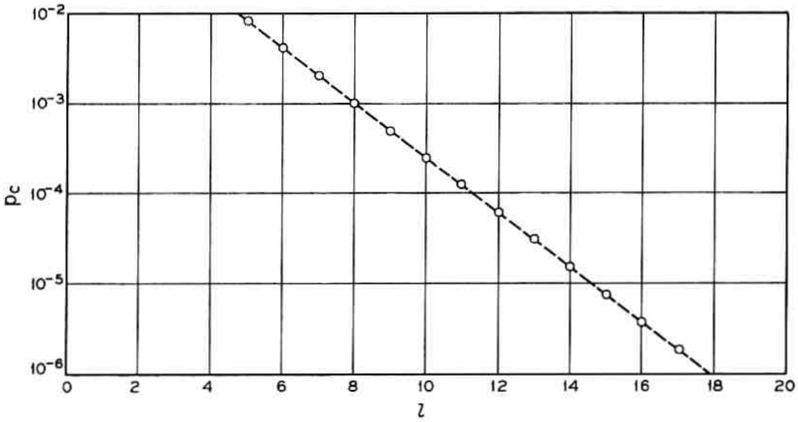


Fig. 6 — p_c for various l.

these single errors occur and the numerical errors that result are essentially independent of l. This implies that the transmission error (and thus the average system error) will be relatively insensitive to l.

Notice that p_c decreases as l increases. The reason is that the quantization error decreases as l increases whereas the transmission error is approximately independent of l. Thus, the value of p where the transmission error becomes a significant portion of the system error decreases.

From equation (10), no system using uncoded transmission can have an average system error significantly less than p no matter how large l becomes. This leads to the problem of how to make the average system error less than p.

Suppose that the σ most significant positions per sample are protected by coding and that the remaining $(l - \sigma)$ positions are not protected. Further, assume that sufficient protection is provided so that the probability of error in the protected positions is substantially less than p. Under these conditions, the transmission error is determined primarily by errors in the least significant positions and we can consider the protected positions to be free of errors. Then, from Theorem 2 of Ref. 2 (summarized in Appendix C),

$$\text{ASE} = \frac{1}{2^l}\left(p \sum_{i=1}^{l-\sigma} 2^{i-1} q^{l-\sigma-i} + \tfrac{1}{4} q^{l-\sigma}\right).$$

For values of p that are of practical interest,

$$\text{ASE} \cong \frac{1}{2^l}\left[\left(2^{l-\sigma} - 1 - \frac{l-\sigma}{4}\right)p + \frac{1}{4}\right]. \tag{11}$$

Accordingly, for p in the range where transmission is the major source of system error, the average system error can be reduced by a factor of approximately $2^{-\sigma}$ from the average system error for uncoded transmission. This implies that we should seek codes that can both protect the significant positions of each sample and maintain quantization resolution by requiring small redundancy. The above requirements provide the motivation for significant-bit packed codes.

VI. SOME EXAMPLES OF SYSTEM PERFORMANCE WITH CODING

In this section we assume that a predetermined number of positions (denoted by ξ) are available to transmit each sample. By numerical evaluation, the average system error that results from the use of representative coding schemes (for $\xi = 7$ and $\xi = 15$*) is determined for various values of p. The examples illustrate that system performance depends upon p and upon the manner in which the ξ positions are allocated between information bits and redundancy for error control.

Let ASE_{UC} denote uncoded transmission. First, consider codes in which one code vector is used per sample ($\alpha = 1$). Listed below is a brief description of each code. The codes are indexed by the notation used for their average system error in Fig. 7 ($\xi = 7$) and Fig. 8 ($\xi = 15$).

$\text{ASE}_{(3,1)}$: A (3, 1) perfect single error-correcting code is used to protect the most significant position.

$\xi = 7$: $\alpha = 1$ $l = 5$

$\xi = 15$: $\alpha = 1$ $l = 13$

$\text{ASE}_{(3,1),(3,1)}$: Independent (3, 1) perfect single error-correcting codes are used to protect the two most significant positions.

$\xi = 7$: $\alpha = 1$ $l = 3$

$\xi = 15$: $\alpha = 1$ $l = 11$

* These values were selected because in each case it is possible to construct a perfect single error-correcting code and thus to compare uniform protection with protection that is heavily weighted in favor of the most significant bit per sample.

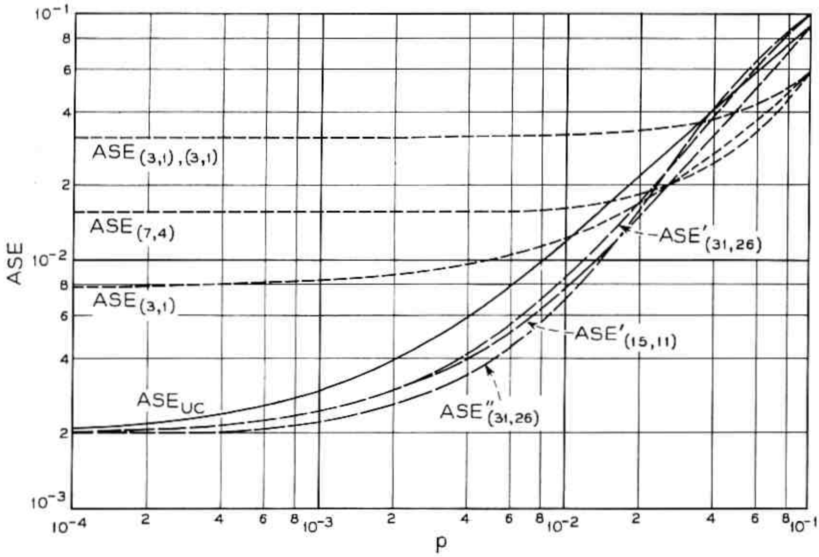


Fig. 7 — Average system error (ASE) with representative codes; 7 positions per sample ($\xi = 7$).

$\text{ASE}_{(7,4)}$: A (7, 4) perfect single error-correcting code is used to protect the four most significant positions.

$$\xi = 7: \qquad \alpha = 1 \qquad l = 4$$

$$\xi = 15: \qquad \alpha = 1 \qquad l = 12$$

$\text{ASE}_{(15,11)}$: A (15, 11) perfect single error-correcting code is used to protect all 11 positions.

$$\xi = 15: \qquad \alpha = 1 \qquad l = 11$$

Although many significant-bit packed codes can be constructed, we consider only three examples. They were selected because the codes should protect the most significant positions of each sample and because a small number of parity check positions per sample should be used so that we can reasonably consider 2^l quantization levels. The codes illustrate the general capabilities of significant-bit packed codes and are easy to implement. One prime is used in the average system error notation to indicate that the most significant position of each sample is protected and two primes to indicate that the two most significant positions of each sample are protected. Let ρ denote the number of parity check positions per sample where $\rho = (n-k)/\alpha$. Let R denote the code rate where $R = k/n$.

$\text{ASE}'_{(15,11)}$: A (15, 11) perfect single error-correcting code is used in a significant-bit packed code to protect the most significant position of each sample.

$\xi = 7$:	$\alpha = 11$	$l = 7$	$\rho = 0.36$	$R = 0.950$
$\xi = 15$:	$\alpha = 11$	$l = 15$	$\rho = 0.36$	$R = 0.976$

$\text{ASE}'_{(31,26)}$: A (31, 26) perfect single error-correcting code is used in a significant-bit packed code to protect the most significant position of each sample.

$\xi = 7$:	$\alpha = 26$	$l = 7$	$\rho = 0.19$	$R = 0.974$
$\xi = 15$:	$\alpha = 26$	$l = 15$	$\rho = 0.19$	$R = 0.987$

$\text{ASE}''_{(31,26)}$: A (31, 26) perfect single error-correcting code is used in a significant-bit packed code to protect the two most significant

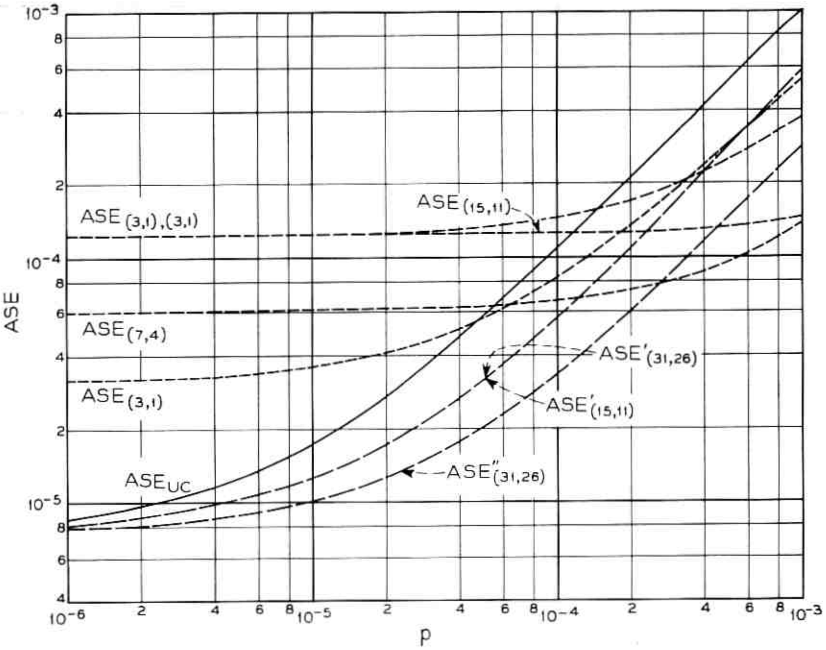


Fig. 8 — Average system error (ASE) with representative codes; 15 positions per sample ($\xi = 15$).

positions of each sample.

$\xi = 7$: $\alpha = 13$ $l = 7$ $\rho = 0.38$ $R = 0.948$

$\xi = 15$: $\alpha = 13$ $l = 15$ $\rho = 0.38$ $R = 0.975$

We can make the following observations concerning system performance when codes are used. In all cases, as $p \rightarrow 0$, ASE $\rightarrow 2^{-(l+2)}$ which is the limitation on system performance because of quantization. As l increases, the quantization error decreases. Thus, the value of p for which the transmission error becomes a significant portion of the system error decreases. In other words, if you design for good quantization resolution, then you need a good channel. This implies that, as the number of positions per sample increases, codes are useful for smaller values of p in order to bring the channel up to the required quality.

Because all $\alpha = 1$ codes necessitate a sizable reduction in l to allow for redundancy, they are only attractive for larger p where considerable coding capability is required. For these p, we have demonstrated that system performance can be improved (by an appreciable amount in some cases) by sacrificing quantization resolution for an improvement in transmission fidelity. However, because significant-bit packed codes provide protection for the most significant positions without the large penalty in quantization resolution required by the $\alpha = 1$ codes, significant-bit packed codes are effective for considerably smaller values of p than are the $\alpha = 1$ codes.

Notice that $\text{ASE}'_{(31,26)}$ and $\text{ASE}'_{(15,11)}$ are nearly equal. The reason is that although the significant-bit packed code using the (31, 26) code provides less error protection than the significant-bit packed code based on the (15, 11) code, in each case the protection provided for the most significant position is "sufficient" and, thus, the errors that hurt are coming in the less significant positions.

On the other hand, $\text{ASE}''_{(31,26)}$ is less than either $\text{ASE}'_{(31,26)}$ or $\text{ASE}'_{(15,11)}$ for the values of p where significant-bit packed codes are preferable. The reason is that errors are now nearly eliminated in the two most significant positions in each sample. Further reductions in system error could be achieved by using significant-bit packed codes which protect three or more positions per sample. However, we must be careful not to go too far or we should begin to charge the redundancy against quantization resolution.

Significant-bit packed codes achieve an effect similar to interleaving. Thus, although the computations herein have been for independent errors, significant-bit packed codes could prove useful for a channel with clustered errors.

VII. SIGNIFICANT-BIT PACKED CODES FOR DIFFERENT l

Several interesting points are illustrated in Fig. 9. Indexed on the left are the four values of l considered. For $l = 15$, ASE_{UC} is shown. For $l = 15$, 14, 13, and 12, $\text{ASE}'_{(31,26)}$ and $\text{ASE}''_{(31,26)}$ are given.

The following observations concerning Fig. 9 can be made. For small p, the $l = 15$ schemes are best. This is to be expected because quantization is the major source of system error for small p.

However, for larger p, the significant-bit packed codes with $l < 15$ have less system error than uncoded transmission for $l = 15$. This is particularly interesting because, in these significant-bit packed codes, more positions are saved by reducing l than are added by the parity check positions. For example, in the $l = 13$ system that results in $\text{ASE}'_{(31,26)}$, $\alpha = 26$ and $n = 343$. If uncoded transmission with $l = 15$ is

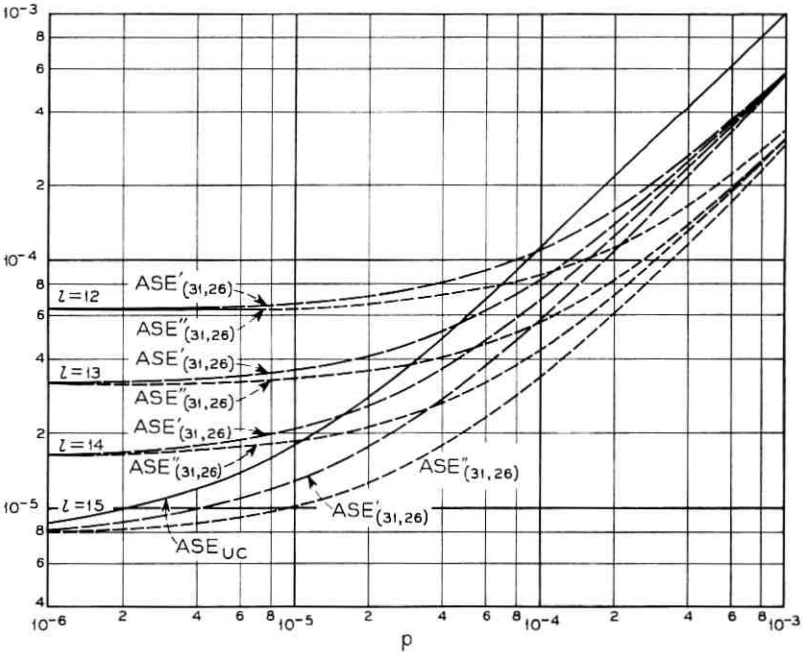


Fig. 9 — Average system error (ASE) with significant-bit packed codes.

used to send these 26 samples, 390 positions are required. Thus, for $p > 4.5 \cdot 10^{-5}$, this significant-bit packed code reduces system error and saves 47 positions every 26 samples. Similar behavior can be noted for other significant-bit packed codes considered in Fig. 9.

For $p = 10^{-3}$, the three systems with $\sigma = 1$ converge to approximately 2^{-1} ASE_{UC} and the three systems with $\sigma = 2$ converge to approximately 2^{-2} ASE_{UC} even though the systems use different quantization resolutions. However, for $p = 10^{-6}$, the convergence is determined by l. This clearly demonstrates the two extreme cases in system behavior: limitation by transmission error and limitation by quantization error.

VIII. THE SYNTHESIS PROBLEM—AN EXAMPLE

Suppose that the probability of error and the maximum allowable average system error are specified. Let these be denoted by p_s and ASE_s respectively. From equation (11), σ and l should be chosen to satisfy the relation

$$\text{ASE}_s > 2^{-\sigma} p_s + 2^{-(l+2)} \qquad (12)$$

where σ represents the number of protected positions per sample. Because equation (11) is an approximation, values of l and σ that satisfy equation (12) cannot be guaranteed to provide a system with an $\text{ASE} \leqq \text{ASE}_s$. However, as σ decreases compared with l, equation (12) becomes increasingly reliable.*

Notice that l and σ appear as negative exponents in equation (12). Therefore, for a given p_s, a wide range of values for the ASE_s can be achieved by varying l and σ. Also, equation (12) frequently can be satisfied by several pairs of values for l and σ. For each pair, there may be several possible coding schemes. The system designer must then choose the final system configuration from these candidates on the basis of such items as the cost of implementation or the number of positions in the data stream per sample.

As an example of system design, consider a telemetry channel in planetary space missions. This channel can often be modeled satisfactorily by the memoryless binary symmetric channel and typically

* A major assumption leading to equation (11) is that all of the average numerical error comes from the unprotected positions. However, if σ is large, then errors in the protected positions result in a much larger numerical error than errors in the unprotected positions. Therefore, even though errors in the protected positions are less likely, a significant portion of the average numerical error can come from these positions.

has a bit error rate of $5\cdot10^{-3}$. Thus, equation (12) becomes

$$\text{ASE}_s > 5\cdot10^{-3}\cdot2^{-\sigma}+2^{-(l+2)}. \tag{13}$$

If uncoded transmission is a system requirement, then $\sigma = 0$ and

$$\text{ASE}_s > 5\cdot10^{-3}+2^{-(l+2)}.$$

Notice that successive increases in l result in successively smaller reductions in the average system error and that the average system error can never be less than $5\cdot10^{-3}$. From Fig. 4, all systems with $l \geqq 8$ have essentially the same average system error and, thus, little is gained by using $l > 8$.

A more interesting situation exists if the system designer is permitted to choose l and the coding scheme. If $\text{ASE}_s > 5\cdot10^{-3}$, it is possible to design a system using uncoded transmission although coding could prove effective as ASE_s approaches $5\cdot10^{-3}$. However, if $\text{ASE}_s <$ $5\cdot10^{-3}$, some form of coding is mandatory. Conversely, from equation (13), if coding is used, the system error can be made small by choosing appropriate values of l and σ. In Table I, the approximate average system error is given for representative l and σ. The information in Table I was computed by using equation (11) and, thus, is subject to the assumptions and approximations leading to equation (11). However, from Table I, the improvements in system performance that can be achieved by coding are evident .

Table I—Approximate Average System Error (ASE) for Representative l and σ; $p = 5\cdot10^{-3}$

l	σ	Approximate ASE
7	1	$4.4\cdot10^{-3}$
	2	$3.2\cdot10^{-3}$
	3	$2.5\cdot10^{-3}$
8	1	$3.5\cdot10^{-3}$
	2	$2.2\cdot10^{-3}$
	3	$1.6\cdot10^{-3}$
9	1	$3.0\cdot10^{-3}$
	2	$1.7\cdot10^{-3}$
	3	$1.1\cdot10^{-3}$
10	1	$2.7\cdot10^{-3}$
	2	$1.5\cdot10^{-3}$
	3	$8.7\cdot10^{-4}$

Consider the following specific example which illustrates certain alternatives in code selection without requiring extensive computational effort. Suppose $\text{ASE}_s = 4 \cdot 10^{-3}$. From equation (13) or Table I, we can use $\sigma = 1$ and $l \geqq 8$ or $\sigma = 2$ and $l \geqq 7$. The minimum values of l will be used. Several coding schemes are possible in each case. The codes, indexed below by the notation used for their average system error in Fig. 10, follow the ideas in Section VI. Thus, the parity check matrices are not presented.

For $\sigma = 1, \quad l = 8$:

$\text{ASE}_{(3,1)}$: A (3, 1) perfect single error-correcting code is used to protect the most significant position.

$$\alpha = 1 \qquad l = 8$$

$\text{ASE}'_{(15,11)}$: A (15, 11) perfect single error-correcting code is used in a significant-bit packed code to protect the most significant position of each sample.

$$\alpha = 11 \qquad l = 8$$

$\text{ASE}'_{(31,26)}$: A (31, 26) perfect single error-correcting code is used in a significant-bit packed code to protect the most significant position of each sample.

$$\alpha = 26 \qquad l = 8$$

For $\sigma = 2, \qquad l = 7$:

$\text{ASE}_{(3,1),(3,1)}$: Independent (3, 1) perfect single error-correcting codes are used to protect the two most significant positions.

$$\alpha = 1 \qquad l = 7$$

$\text{ASE}''_{(31,26)}$: A (31, 26) perfect single error-correcting code is used in a significant-bit packed code to protect the two most significant positions of each sample.

$$\alpha = 13 \qquad l = 7$$

The design objective, denoted by an asterisk in Fig. 10, is satisfied by each system although the systems vary somewhat in performance for other p. Notice that the systems differ in the coding equipment and quantization resolution required for implementation. Also, notice that the systems vary in the number of positions per sample in the data stream [from a low of 7.4 for $\text{ASE}''_{(31,26)}$ to a high of 11 for $\text{ASE}_{(3,1),(3,1)}$]. Which system would actually be selected would thus depend upon the details of the specific application.

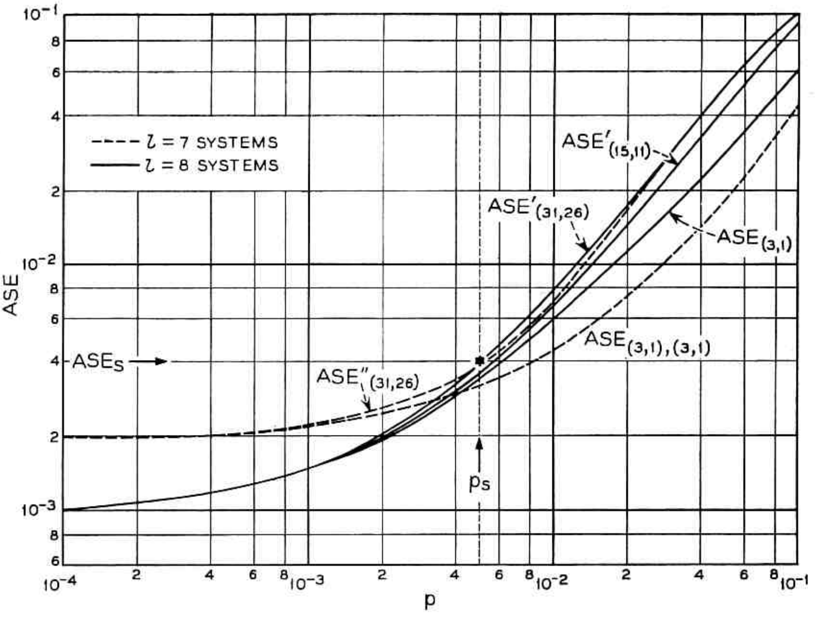


Fig. 10 — Systems for space telemetry channel.

IX. CONCLUSIONS

A general formulation of the error introduced by quantization and transmission has been developed for the data transmission system shown in Fig. 1. It has been shown that system performance is influenced by both the quantization resolution and the channel error characteristics, that certain levels of performance cannot be achieved without the use of coding no matter how fine the quantization, and that performance can, in some cases, be improved by sacrificing quantization for redundancy and error control. In general, when coding is used, it is beneficial to use codes that match their protection to the numerical significance of the information positions. Significant-bit packed codes are particularly useful because they provide protection for the most significant positions without incurring a large penalty in quantization resolution. The problem of determining the coding capability and the number of quantization levels required to achieve a specified average system error has been considered.

The specific results are based upon the choice of $\gamma = 1$ in Section IV. However, varying γ simply changes the "cost" assigned to the numerical errors and, thus, the general ideas presented here are applicable for any $\gamma > 0$: for example, the desirability of the system approach to quantization and transmission error, the possibility of improving system performance by sacrificing quantization resolution for redundancy, and the use of codes that concentrate protection on the numerically most significant positions. Actually, it appears that as γ increases, the desirability of protection for the most significant positions also increases.

Because of the unit distance properties of Gray codes, it is natural to inquire whether Gray codes could prove useful in the system discussed in this paper. It can be shown (for $\gamma = 1$) that a Gray code with 2^l levels gives exactly the same average numerical error and average system error as the natural binary numbering with 2^l levels even when error-correcting codes are used.

APPENDIX A

Derivation of an Expression for $Pr_{\mathrm{m}}\{\mathrm{r_m} \mid \mathrm{s_m}\}$

Let $\Pr_m \{r_m \mid s_m\}$ denote the probability of receiving r_m when s_m is transmitted using a packed code. Let $\Pr\{s_i\}$ $(1 \leqq i \leqq \alpha)$ denote the probability that s_i is transmitted. Then

$$\Pr_m \{r_m \mid s_m\} = \underbrace{\sum_{r_\alpha=0}^{2^l-1} \cdots \sum_{r_1=0}^{2^l-1} \sum_{s_\alpha=0}^{2^l-1} \cdots \sum_{s_1=0}^{2^l-1}}_{\text{excluding } r_m \text{ and } s_m} \Pr\{r \mid s\} \underbrace{\Pr\{s_\alpha\} \cdots \Pr\{s_1\}}_{\text{excluding } \Pr\{s_m\}}$$

where the values of r and s are determined from equations (4) and (2), respectively. However, $\Pr\{s_i\} = 2^{-l}$ for $1 \leqq i \leqq \alpha$, $i \neq m$. Thus,

$$\Pr_m \{r_m \mid s_m\} = \frac{1}{2^{(\alpha-1)l}} \underbrace{\sum_{r_\alpha=0}^{2^l-1} \cdots \sum_{r_1=0}^{2^l-1} \sum_{s_\alpha=0}^{2^l-1} \cdots \sum_{s_1=0}^{2^l-1}}_{\text{excluding } r_m \text{ and } s_m} \Pr\{r \mid s\}. \tag{14}$$

The expression in equation (14) can be simplified. From equations (2) and (4), equation (14) can be written as

$$\Pr_m \{r_m \mid s_m\} = \frac{1}{2^{(\alpha-1)l}} \underbrace{\sum_{s_\alpha=0}^{2^l-1} \cdots \sum_{s_1=0}^{2^l-1} \sum_{r_\alpha=0}^{2^l-1} \cdots \sum_{r_1=0}^{2^l-1}}_{\text{excluding } r_m \text{ and } s_m}$$

$$\cdot \Pr\left\{ \sum_{m'=1}^{\alpha} 2^{(m'-1)l} r_{m'} \;\middle|\; \sum_{m'=1}^{\alpha} 2^{(m'-1)l} s_{m'} \right\}. \tag{15}$$

By Lemma 1 of Ref. 2, for a binary group code used with a binary symmetric channel in which the errors are independent of the symbols actually transmitted,

$$\Pr\left\{\sum_{m'=1}^{\alpha} 2^{(m'-1)l}t_{m'} \,\middle|\, 0\right\} = \Pr\left\{\sum_{m'=1}^{\alpha} 2^{(m'-1)l}r_{m'} \,\middle|\, \sum_{m'=1}^{\alpha} 2^{(m'-1)l}s_{m'}\right\}$$

where $B_l(t_{m'}) = B_l(r_{m'}) \oplus B_l(s_{m'})$. By Lemma 2 of Ref. 2, equation (15) can be written as

$$\Pr_m \{r_m \mid s_m\} = \frac{1}{2^{(\alpha-1)l}} \underbrace{\sum_{s_\alpha=0}^{2^l-1} \cdots \sum_{s_1=0}^{2^l-1} \sum_{t_\alpha=0}^{2^l-1} \cdots \sum_{t_1=0}^{2^l-1}}_{\text{excluding } s_m \text{ and } t_m} \Pr\left\{\sum_{m'=1}^{\alpha} 2^{(m'-1)l}t_{m'} \,\middle|\, 0\right\}$$

which reduces to

$$\Pr_m \{r_m \mid s_m\} = \underbrace{\sum_{t_\alpha=0}^{2^l-1} \cdots \sum_{t_1=0}^{2^l-1}}_{\text{excluding } t_m} \Pr\left\{\sum_{m'=1}^{\alpha} 2^{(m'-1)l}t_{m'} \,\middle|\, 0\right\}.$$

APPENDIX B

Reduction of the Expression for the Average System Error

By substituting equation (3) into (5) and rewriting,

$$\text{ASE}_m = \frac{1}{(X_2 - X_1)} \sum_{r_m=0}^{2^l-1} \sum_{s_m=0}^{2^l-1} \Pr_m \{r_m \mid s_m\} \cdot \int_{X_1+s_m(X_2-X_1)/2^l}^{X_1+(s_m+1)(X_2-X_1)/2^l} \left| x_m - X_1 - (r_m + \tfrac{1}{2})\left(\frac{X_2 - X_1}{2^l}\right) \right| dx_m .$$

However,

$$\int_{X_1+s_m(X_2-X_1)/2^l}^{X_1+(s_m+1)(X_2-X_1)/2^l} \left| x_m - X_1 - (r_m + \tfrac{1}{2})\left(\frac{X_2 - X_1}{2^l}\right) \right| dx_m = \left(\frac{X_2 - X_1}{2^l}\right)^2 (\mid r_m - s_m \mid + \tfrac{1}{4}\, \delta_{r_m s_m})$$

where

$$\begin{aligned} \delta_{r_m s_m} &= 1 \quad \text{for} \quad r_m = s_m \\ &= 0 \quad \text{for} \quad r_m \neq s_m . \end{aligned}$$

Thus,

$$\text{ASE}_m = \frac{X_2 - X_1}{2^{2l}} \sum_{r_m=0}^{2^l-1} \sum_{s_m=0}^{2^l-1} \mid r_m - s_m \mid \Pr_m \{r_m \mid s_m\} + \frac{X_2 - X_1}{4 \cdot 2^{2l}} \sum_{r_m=0}^{2^l-1} \sum_{s_m=0}^{2^l-1} \Pr_m \{r_m \mid s_m\}\, \delta_{r_m s_m} .$$

The average numerical error for the *m*th sample was defined in equation (6) as

$$\text{ANE}_m = \frac{1}{2^l} \sum_{r_m=0}^{2^l-1} \sum_{s_m=0}^{2^l-1} | r_m - s_m | \Pr_m \{r_m \mid s_m\}.$$

In addition, it can be shown that for a channel in which the errors are independent of the symbols actually transmitted,

$$\sum_{r_m=0}^{2^l-1} \sum_{s_m=0}^{2^l-1} \Pr_m \{r_m \mid s_m\} \ \delta_{r_m s_m} = 2^l \Pr_m \{0 \mid 0\}.$$

Therefore,

$$\text{ASE}_m = \left(\frac{X_2 - X_1}{2^l}\right)(\text{ANE}_m + \tfrac{1}{4} \Pr_m \{0 \mid 0\}).$$

APPENDIX C

Theorem 2 of Reference 2

A significant-bit code is a code in which the $(k - k_0)$ most significant positions are protected by what is referred to as a base code and the remaining k_0 positions are transmitted unprotected. For the base code when used alone, let $\Pr_B\{0 \mid 0\}$ denote the probability that the output of the decoder is the zero message when the input to the encoder is the zero message. Also, let ANE_B denote the average numerical error of the base code. The average numerical error for the significant-bit code is given by Theorem 2 of Ref. 2:

Theorem 2: Let the base code be defined as above. For a binary symmetric channel with independent errors and when all messages are equally likely to be transmitted,

$$\text{ANE}_{\text{SB}} = \Pr_B \{0 \mid 0\} p \sum_{j=1}^{k_0} 2^{j-1} q^{k_0 - j} + 2^{k_0} \text{ANE}_\text{B} \ .$$

Uncoded transmission is the special case where $k = k_0$. Thus, the average numerical error for uncoded transmission can be obtained by letting $\text{ANE}_\text{B} = 0$ and $\Pr_B\{0 \mid 0\} = 1$ when $k = k_0$.

APPENDIX D

Proof that the Average System Error for Uncoded Transmission Increases Monotonically with p

In Section V, equation (9) gives the average system error for uncoded transmission as

$$\text{ASE}_{\text{UC}} = \frac{1}{2^l}\left(p \sum_{i=1}^{l} 2^{i-1}q^{l-i} + \frac{q^l}{4}\right).$$

After differentiating with respect to q and grouping terms,

$$\frac{d\text{ASE}_{\text{UC}}}{dq} = \frac{1}{2^l}\left[-\left(l - \frac{l}{4}\right)q^{l-1} - \sum_{i=1}^{l-1}(l-i)(2^i - 2^{i-1})q^{l-i-1}\right].$$

For $\frac{1}{2} < q < 1$,

$$\frac{d\text{ASE}_{\text{UC}}}{dq} < 0.$$

Thus, ASE_{UC} decreases monotonically as q goes from ½ to 1 or, alternatively, ASE_{UC} increases monotonically as p runs from 0 to ½.

APPENDIX E

Parity Check Matrices for Codes Considered in Section VI

$\text{ASE}_{(3,1)}$: A (3, 1) perfect single error-correcting code to protect the most significant position.

$\xi = 7$:

$$H = \left[\begin{matrix}1&0&0&0&0\\1&0&0&0&0\end{matrix}\;\; I_2\right] \qquad \alpha = 1 \qquad l = 5$$

$\xi = 15$:

$$H = \left[\begin{matrix}1&0&0&0&0&0&0&0&0&0&0&0&0\\1&0&0&0&0&0&0&0&0&0&0&0&0\end{matrix}\;\; I_2\right] \qquad \alpha = 1 \qquad l = 13$$

$\text{ASE}_{(3,1),(3,1)}$: Independent (3, 1) perfect single error-correcting codes to protect the two most significant positions.

$\xi = 7$:

$$H = \left[\begin{matrix}1&0&0\\1&0&0\\0&1&0\\0&1&0\end{matrix}\;\; I_4\right] \qquad \alpha = 1 \qquad l = 3$$

$\xi = 15$:

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_4 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \end{bmatrix} \qquad \alpha = 1 \qquad l = 11$$

ASE$_{(7,4)}$: A (7, 4) perfect single error-correcting code to protect the four most significant positions.

$\xi = 7$:

$$H = \begin{bmatrix} 1 & 1 & 1 & 0 & \\ 1 & 1 & 0 & 1 & I_3 \\ 1 & 0 & 1 & 1 & \end{bmatrix} \qquad \alpha = 1 \qquad l = 4$$

$\xi = 15$:

$$H = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_3 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \end{bmatrix} \qquad \alpha = 1 \qquad l = 12$$

ASE$_{(15,11)}$: A (15, 11) perfect single error-correcting code.

$\xi = 15$:

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & I_4 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & \end{bmatrix} \qquad \alpha = 1 \qquad l = 11$$

ASE$'_{(15,11)}$: *A* (15, 11) perfect single error-correcting code in a significant-bit packed code.

$\xi = 7$:

$$H = \left[\begin{array}{ccccccccccccccccccccccc} 1 & & 1 & & 1 & & 1 & & 1 & & 1 & & 1 & & 0 & & 0 & & 0 & & 0 & & \\ 1 & 0_6 & 1 & 0_6 & 1 & 0_6 & 1 & 0_6 & 0 & 0_6 & 0 & 0_6 & 0 & 0_6 & 1 & 0_6 & 1 & 0_6 & 1 & 0_6 & 0 & 0_6 & I_4 \\ 1 & & 1 & & 0 & & 0 & & 1 & & 1 & & 0 & & 1 & & 1 & & 0 & & 1 & & \\ 1 & & 0 & & 1 & & 0 & & 1 & & 0 & & 1 & & 1 & & 0 & & 1 & & 1 & & \end{array}\right]$$

$$\alpha = 11 \qquad l = 7 \qquad \rho = 0.36 \qquad R = 0.950$$

$\xi = 15$:

$$H = \left[\begin{array}{cccccccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & \\
1\;0_{14} & 1\;0_{14} & 1\;0_{14} & 1\;0_{14} & 0\;0_{14} & 0\;0_{14} & 0\;0_{14} & 1\;0_{14} & 1\;0_{14} & 1\;0_{14} & 0\;0_{14} & I_4 \\
1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & \\
1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 &
\end{array}\right]$$

$\alpha = 11 \qquad l = 15 \qquad \rho = 0.36 \qquad R = 0.976$

ASE$'_{(31,26)}$: A (31, 26) perfect single error-correcting code in a significant-bit packed code.

$\xi = 7$:

$$H = \left[\begin{array}{cccccccccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1\;0_6 & 1\;0_6 & 1\;0_6 & 1\;0_6 & 0\;0_6 & 0\;0_6 & 0\;0_6 & 0\;0_6 & 1\;0_6 & 1\;0_6 & 1\;0_6 & 1\;0_6 & 0\;0_6 & 0\;0_6 \\
1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0
\end{array}\right.$$

$$\left.\begin{array}{ccccccccccccc}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & \\
0\;0_6 & 1\;0_6 & 1\;0_6 & 1\;0_6 & 1\;0_6 & 0\;0_6 & 0\;0_6 & 0\;0_6 & 1\;0_6 & 1\;0_6 & 1\;0_6 & 0\;0_6 & I_5 \\
0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & \\
1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 &
\end{array}\right]$$

$\alpha = 26 \qquad l = 7 \qquad \rho = 0.19 \qquad R = 0.973$

$\xi = 15$:

$$H = \left[\begin{array}{ccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1\;0_{14} & 1\;0_{14} & 1\;0_{14} & 1\;0_{14} & 0\;0_{14} & 0\;0_{14} & 0\;0_{14} \\
1 & 1 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 1
\end{array}\right.$$

$$\begin{matrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0\ 0_{14} & 1\ 0_{14} & 1\ 0_{14} & 1\ 0_{14} & 1\ 0_{14} & 0\ 0_{14} & 0\ 0_{14} & 0\ 0_{14} & 1\ 0_{14} & 1\ 0_{14} \\
0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0
\end{matrix}$$

$$\left.\begin{matrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & \\
1\ 0_{14} & 1\ 0_{14} & 0\ 0_{14} & 0\ 0_{14} & 0\ 0_{14} & 1\ 0_{14} & 1\ 0_{14} & 1\ 0_{14} & 0\ 0_{14} & I_5 \\
0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & \\
1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 &
\end{matrix}\right]$$

$$\alpha = 26 \qquad l = 15 \qquad \rho = 0.19 \qquad R = 0.987$$

$\mathrm{ASE}''_{(31,26)}$: A (31, 26) perfect single error-correcting code in a significant-bit packed code.

$\xi = 7$:

$$H = \left[\begin{matrix}
11 & 11 & 11 & 11 & 11 & 11 \\
11 & 11 & 11 & 11 & 00 & 00 \\
11\ 0_5 & 11\ 0_5 & 00\ 0_5 & 00\ 0_5 & 11\ 0_5 & 11\ 0_5 \\
11 & 00 & 11 & 00 & 11 & 00 \\
10 & 10 & 10 & 10 & 10 & 10
\end{matrix}\right.$$

$$\left.\begin{matrix}
11 & 10 & 00 & 00 & 00 & 00 & 00 & \\
00 & 01 & 11 & 11 & 11 & 00 & 00 & \\
00\ 0_5 & 01\ 0_5 & 11\ 0_5 & 10\ 0_5 & 00\ 0_5 & 11\ 0_5 & 10\ 0_5 & I_5 \\
11 & 01 & 10 & 01 & 10 & 11 & 01 & \\
10 & 11 & 01 & 01 & 01 & 10 & 11 &
\end{matrix}\right]$$

$$\alpha = 13 \qquad l = 7 \qquad \rho = 0.38 \qquad R = 0.948$$

$\xi = 15$:

$$H = \left[\begin{array}{cccccccccccc} 11 & & 11 & & 11 & & 11 & & 11 & & 11 & \\ 11 & & 11 & & 11 & & 11 & & 00 & & 00 & \\ 11 & 0_{13} & 11 & 0_{13} & 00 & 0_{13} & 00 & 0_{13} & 11 & 0_{13} & 11 & 0_{13} \\ 11 & & 00 & & 11 & & 00 & & 11 & & 00 & \\ 10 & & 10 & & 10 & & 10 & & 10 & & 10 & \end{array}\right.$$

$$\left.\begin{array}{ccccccccccccccc} 11 & & 10 & & 00 & & 00 & & 00 & & 00 & & 00 & & \\ 00 & & 01 & & 11 & & 11 & & 11 & & 00 & & 00 & & \\ 00 & 0_{13} & 01 & 0_{13} & 11 & 0_{13} & 10 & 0_{13} & 00 & 0_{13} & 11 & 0_{13} & 10 & 0_{13} & I_5 \\ 11 & & 01 & & 10 & & 01 & & 10 & & 11 & & 01 & & \\ 10 & & 11 & & 01 & & 01 & & 01 & & 10 & & 11 & & \end{array}\right]$$

$$\alpha = 13 \qquad l = 15 \qquad \rho = 0.38 \qquad R = 0.975$$

REFERENCES

1. Masnick, B. and Wolf, J. K., "On Linear Unequal Error Protection Codes," IEEE Trans. Inform. Theory, *IT-13*, No. 4 (October 1967), pp. 600–607.
2. Buchner, M. M., Jr., "Coding for Numerical Data Transmission," B.S.T.J., *46*, No. 5 (May-June 1967), pp. 1025–1041.
3. Peterson, W. W., *Error Correcting Codes*, New York: M.I.T. Press and John Wiley and Sons, 1961.

The Chirp z-Transform Algorithm and Its Application

§ By LAWRENCE R. RABINER, RONALD W. SCHAFER, and CHARLES M. RADER*

(Manuscript received November 21, 1968)

We discuss a computational algorithm for numerically evaluating the z-transform of a sequence of N *samples. This algorithm has been named the chirp z-transform algorithm. Using this algorithm one can efficiently evaluate the z-transform at* M *points in the z-plane which lie on circular or spiral contours beginning at any arbitrary point in the z-plane. The angular spacing of the points is an arbitrary constant;* M *and* N *are arbitrary integers.*

The algorithm is based on the fact that the values of the z-transform on a circular or spiral contour can be expressed as a discrete convolution. Thus one can use well-known high-speed convolution techniques to evaluate the transform efficiently. For M *and* N *moderately large, the computation time is roughly proportional to* $(N + M) \log_2 (N + M)$ *as opposed to being proportional to* $N \cdot M$ *for direct evaluation of the z-transform at* M *points.*

Applications discussed include: enhancement of poles in spectral analysis, high resolution narrow-band frequency analysis, interpolation of band-limited waveforms, and the conversion of a base 2 fast Fourier transform program into an arbitrary radix fast Fourier transform program.

I. INTRODUCTION

In dealing with sampled data the z-transform plays the role which is played by the Laplace transform in continuous time systems. One example of its application is spectrum analysis. The computation of sampled z-transforms, which has been greatly facilitated by the fast Fourier transform algorithm, is further facilitated by the "chirp z-transform" algorithm described in this paper.[1,2]

* Mr. Rader is with Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Massachusetts. Lincoln Laboratory is operated with support from the U. S. Air Force.

The z-transform of a sequence of numbers x_n is defined as

$$X(z) = \sum_{n=-\infty}^{\infty} x_n z^{-n}, \tag{1}$$

a function of the complex variable z. In general both x_n and $X(z)$ could be complex. It is assumed that the sum on the right side of equation (1) converges for at least some values of z. We restrict ourselves to the z-transform of sequences with only a finite number N of nonzero points. Therefore, we can rewrite equation (1) without loss of generality as

$$X(z) = \sum_{n=0}^{N-1} x_n z^{-n} \tag{2}$$

where the sum in equation (2) converges for all z except $z = 0$.

Equations (1) and (2) are similar to the defining expressions for the Laplace transform of a train of equally spaced impulses of magnitudes x_n. Let the spacing of the impulses be T and let the train of impulses be

$$\sum_n x_n \, \delta(t - nT).$$

Then the Laplace transform is

$$\sum_n x_n e^{-snT}$$

which is the same as $X(z)$ if we let

$$z = e^{sT}. \tag{3}$$

For sampled waveforms the relation between the original waveform and the train of impulses is well understood in terms of the phenomenon of aliasing. Thus the z-transform of the sequence of samples of a time waveform is representative of the Laplace transform of the original waveform in a way which is well understood. The Laplace transform of a train of impulses repeats its values taken in a horizontal strip of the s-plane of width $2\pi/T$ in every other strip parallel to it. The z-transform maps each such strip into the entire z-plane or, conversely, the entire z-plane corresponds to any horizontal strip of the s-plane, for example, the region $-\infty < \sigma < \infty$, $-\pi/T \leqq \omega < \pi/T$, where $s = \sigma + j\omega$.

In the same correspondence, the $j\omega$ axis of the s-plane, along which we generally equate the Laplace transform with the Fourier transform, is the unit circle in the z-plane; the origin of the s-plane corresponds to $z = 1$. The interior of the z-plane unit circle corresponds to the left half of the s-plane; the exterior corresponds to the right half plane. Straight lines in the s-plane correspond to circles or spirals in the z-plane. Figure 1 shows the correspondence of a contour in the s-plane to a contour in the z-plane. To evaluate the Laplace transform of the impulse train along the linear contour is to evaluate the z-transform of the sequence along the spiral contour.

Values of the z-transform are usually computed along the path corresponding to the $j\omega$ axis, namely the unit circle. This gives the discrete equivalent of the Fourier transform and has many applications including the estimation of spectra, filtering, interpolation, and correlation. The applications of computing z-transforms off the unit circle are fewer, but one is presented in this paper, namely the enhancement of spectral resonances in systems for which one has some foreknowledge of the locations of poles and zeros.

Just as we can only compute equation (2) for a finite set of samples, so we can only compute equation (2) at a finite number of points, say z_k:

$$X_k = X(z_k) = \sum_{n=0}^{N-1} x_n z_k^{-n}. \qquad (4)$$

The special case which has received the most attention is the set of points equally spaced around the unit circle,

$$z_k = \exp\left(j\,\frac{2\pi}{N}\,k\right), \qquad k = 0, 1, \cdots, N-1 \qquad (5)$$

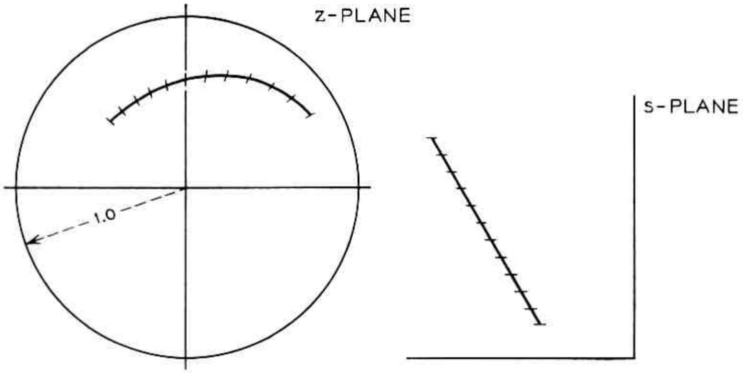


Fig. 1 — The correspondence of a z-plane contour to an s-plane contour through the relation $z = e^{sT}$.

for which

$$X_k = \sum_{n=0}^{N-1} x_n \exp\left(-j\frac{2\pi}{N}nk\right), \qquad k = 0, 1, \cdots, N-1. \tag{6}$$

Equation (6) is called the discrete Fourier transform. The reader may easily verify that, in equation (5), other values of k merely repeat the same N values of z_k, which are the Nth roots of unity. The discrete Fourier transform has assumed considerable importance, partly because of its nice properties but mainly because since 1965 it has become widely known that the computation of equation (6) can be achieved, not in the N^2 complex multiplications and additions called for by direct application of equation (6), but in something of the order of $N \log_2 N$ operations if N is a power of two, or $N \sum_i m_i$ operations if the integers m_i are the prime factors of N. Any algorithm which accomplishes this is called a fast Fourier transform. Much of the importance of the fast Fourier transform is that the discrete Fourier transform may be used as a stepping stone to computing lagged products such as convolutions, autocorrelations, and cross correlations more rapidly than before.[3,4] The discrete Fourier transform has some limitations which can be eliminated using the chirp z-transform algorithm which we describe. We also investigate the computation of the z-transform on a more general contour, of the form

$$z_k = AW^{-k}, \qquad k = 0, 1, \cdots, M-1 \tag{7a}$$

where M is an arbitrary integer and both A and W are arbitrary complex numbers of the form

$$A = A_o \exp(j2\pi\theta_o) \tag{7b}$$

and

$$W = W_o \exp(j2\pi\varphi_o). \tag{7c}$$

(See Fig. 2.) The case $A = 1$, $M = N$, and $W = \exp(-j2\pi/N)$ corresponds to the discrete Fourier transform. The general z-plane contour begins with the point $z = A$ and, depending on the value of W, spirals in or out with respect to the origin. If $W_o = 1$, the contour is an arc of a circle. The angular spacing of the samples is $2\pi\varphi_o$. The equivalent s-plane contour begins with the point

$$s_o = \sigma_o + j\omega_o = \frac{1}{T}\ln A \tag{8}$$

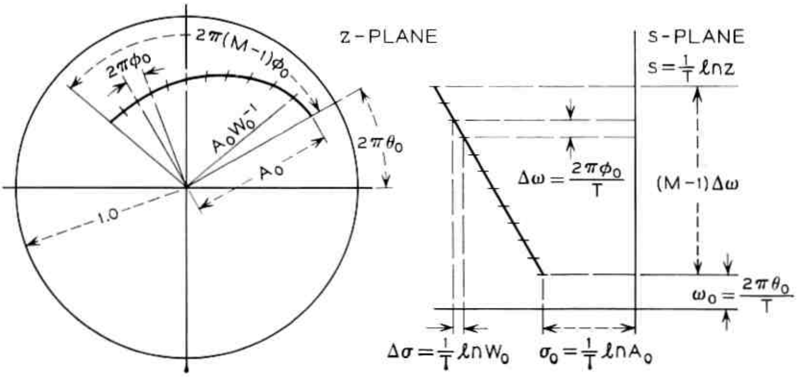

Z-PLANE
S-PLANE
$s = \frac{1}{T} \ln z$
$2\pi(M-1)\phi_0$
$2\pi\phi_0$
$A_0 W_0^{-1}$
A_0
1.0
$2\pi\theta_0$
$\Delta\omega = \frac{2\pi\phi_0}{T}$
$(M-1)\Delta\omega$
$\omega_0 = \frac{2\pi\theta_0}{T}$
$\Delta\sigma = \frac{1}{T} \ln W_0$
$\sigma_0 = \frac{1}{T} \ln A_0$

Fig. 2 — An illustration of the independent parameters of the chirp z-transform algorithm. The upper figure shows how the z-transform is evaluated on a spiral contour starting at the point $z = A$. The lower figure shows the corresponding straight line contour and independent parameters in the s-plane.

and the general point on the s-plane contour is

$$s_k = s_o + k(\Delta\sigma + j\,\Delta\omega) = \frac{1}{T}(\ln A - k \ln W),$$
$$k = 0, 1, \cdots, M - 1. \qquad (9)$$

Since A and W are arbitrary complex numbers we see that the points s_k lie on an arbitrary straight line segment of arbitrary length and sampling density. Clearly the contour indicated in equation (7a) is not the most general contour, but it is considerably more general than that for which the discrete Fourier transform applies. In Fig. 2, an example of this more general contour is shown in both the z-plane and the s-plane.

To compute the z-transform along this more general contour would seem to require NM multiplications and additions since the special symmetries of exp $(j2\pi k/N)$ which are exploited in the derivation of the fast Fourier transform are absent in the more general case. However, we shall see that by using the sequence $W^{n^2/2}$ in various roles we can apply the fast Fourier transform to the computation of the z-transform along the contour of equation (7a). Since for $W_o = 1$, the sequence $W^{n^2/2}$ is a complex sinusoid of linearly increasing frequency, and since a similar waveform used in some radar systems has the picturesque name "chirp", we call the algorithm we are about to present the chirp z-transform. Since this transform permits computing the z-transform on a more general contour than the fast Fourier transform permits, it is more flexible than the fast Fourier transform, although it is also considerably slower. The additional freedoms offered by the chirp z-transform include:

(*i*) the number of time samples does not have to equal the number of samples of the z-transform.

(*ii*) Neither M nor N need be a composite integer.

(*iii*) The angular spacing of the z_k is arbitrary.

(*iv*) The contour need not be a circle but can spiral in or out with respect to the origin. In addition, the point z_o is arbitrary, but this is also the case with the fast Fourier transform if the samples x_n are multiplied by z_o^{-n} before transforming.

II. DERIVATION OF THE CHIRP Z-TRANSFORM

Along the contour of equation (7a), equation (4) becomes

$$X_k = \sum_{n=0}^{N-1} x_n A^{-n} W^{nk}, \qquad k = 0, 1, \cdots, M-1 \tag{10}$$

which, at first appearance, seems to require NM complex multiplications and additions, as we have already observed. But, let us use Bluestein's ingenious substitution[5]

$$nk = \frac{n^2 + k^2 - (k-n)^2}{2} \tag{11}$$

for the exponent of W in equation (10). This produces an apparently more unwieldly expression

$$X_k = \sum_{n=0}^{N-1} x_n A^{-n} W^{(n^2/2)} W^{(k^2/2)} W^{-(k-n)^2/2}, \qquad k = 0, 1, \cdots, M-1 \tag{12}$$

but in fact equation (12) can be thought of as a three step process consisting of: (*i*) forming a new sequence y_n by weighting the x_n according to the equation

$$y_n = x_n A^{-n} W^{n^2/2}, \qquad n = 0, 1, \cdots, N-1, \tag{13}$$

(*ii*) convolving y_n with the sequence v_n defined as

$$v_n = W^{-n^2/2} \tag{14}$$

to give a sequence g_k

$$g_k = \sum_{n=0}^{N-1} y_n v_{k-n}, \qquad k = 0, 1, \cdots, M-1, \tag{15}$$

and (*iii*) multiplying g_k by $W^{k^2/2}$ to give X_k

$$X_k = g_k W^{k^2/2}, \qquad k = 0, 1, \cdots, M - 1. \tag{16}$$

The three step process is illustrated in Fig. 3. Steps *i* and *iii* require N and M multiplications respectively; step *ii* is convolution which may be computed by the high speed technique disclosed by Stockham, based on the use of the fast Fourier transform.[3] Step *ii* is the major part of the computational effort and requires a time roughly proportional to $(N + M) \log (N + M)$.

Bluestein used the substitution of equation (11) to convert a discrete Fourier transform to a convolution as in Fig. 3. The linear system to which the convolution is equivalent can be called a chirp filter which is, in fact, also sometimes used to resolve a spectrum. Bluestein showed that for N a perfect square, the chirp filter could be synthesized recursively with $N^{1/2}$ multipliers and the computation of a discrete Fourier transform could then be proportional to $N^{3/2}$ (see Ref. 5).

The flexibility and speed of the chirp z-transform algorithm are related to the flexibility and speed of the method of high-speed convolution using the fast Fourier transform. Recall that the product of the discrete Fourier transforms of two sequences is the discrete Fourier transform of the circular convolution of the two sequences; therefore, a circular convolution is computable as two discrete Fourier transforms, the multiplication of two arrays of complex numbers, and an inverse discrete Fourier transform, which can also be computed by the fast Fourier transform. Ordinary convolutions can be computed as circular convolutions by appending zeroes to the end of one or both sequences so that the correct numerical answers for the ordinary convolution can result from a circular convolution.

We now summarize the details of the chirp z-transform algorithm on the assumption that an already existing fast Fourier transform

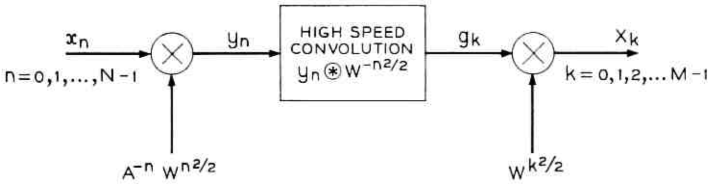


Fig. 3 — An illustration of the steps involved in computing values of the z-transform using the chirp z-transform algorithm.

program (or special purpose machine) is available to compute discrete and inverse discrete Fourier transforms.

We begin with a waveform in the form of N samples x_n and we seek M samples of X_k where A and W have also been chosen:

(*i*) We choose L, the smallest integer greater than or equal to $N + M - 1$ which is also compatible with our high speed fast Fourier transform program. For most users this will mean L is a power of two. Notice that while many fast Fourier transform programs will work for arbitrary L, they are not equally efficient for all L. At the very least, L should be highly composite.

(*ii*) We form an L point sequence y_n from x_n by the equation

$$y_n = \begin{cases} A^{-n}W^{n^2/2}x_n & n = 0, 1, 2, \cdots, N-1 \\ 0 & n = N, N+1, \cdots, L-1 \end{cases} . \tag{17}$$

(*iii*) We compute the L point discrete Fourier transform of y_n by the fast Fourier transform, calling it Y_r, $r = 0, 1, \ldots, L - 1$.

(*iv*) We define an L point sequence v_n by the relation

$$v_n = \begin{cases} W^{-n^2/2} & 0 \leqq n \leqq M-1 \\ W^{-(L-n)^2/2} & L - N + 1 \leqq n < L. \\ \text{arbitrary} & \text{other } n, \text{ if any} \end{cases} \tag{18}$$

If L is exactly equal to $M + N - 1$, the region in which v_n is arbitrary will not exist. If the region does exist an obvious possibility is to increase M, the desired number of points of the z-transform we compute, until the region does not exist.

Notice that v_n could be cut into two with a cut between $n = M - 1$ and $n = L - N + 1$; if the two pieces were abutted differently, the resulting sequence would be a slice out of the indefinite length sequence $W^{-n^2/2}$. This is illustrated in Fig. 4. The sequence v_n is defined the way it is in order to force the circular convolution to give us the desired numerical results of an ordinary convolution.

(*v*) We compute the discrete Fourier transform of v_n and call it V_r, $r = 0, 1, \ldots, L - 1$.

(*vi*) We multiply V_r and Y_r point by point, giving G_r:

$$G_r = V_r Y_r , \qquad r = 0, 1, \cdots, L - 1.$$

(*vii*) We compute the L point inverse discrete Fourier transform g_k, of G_r.

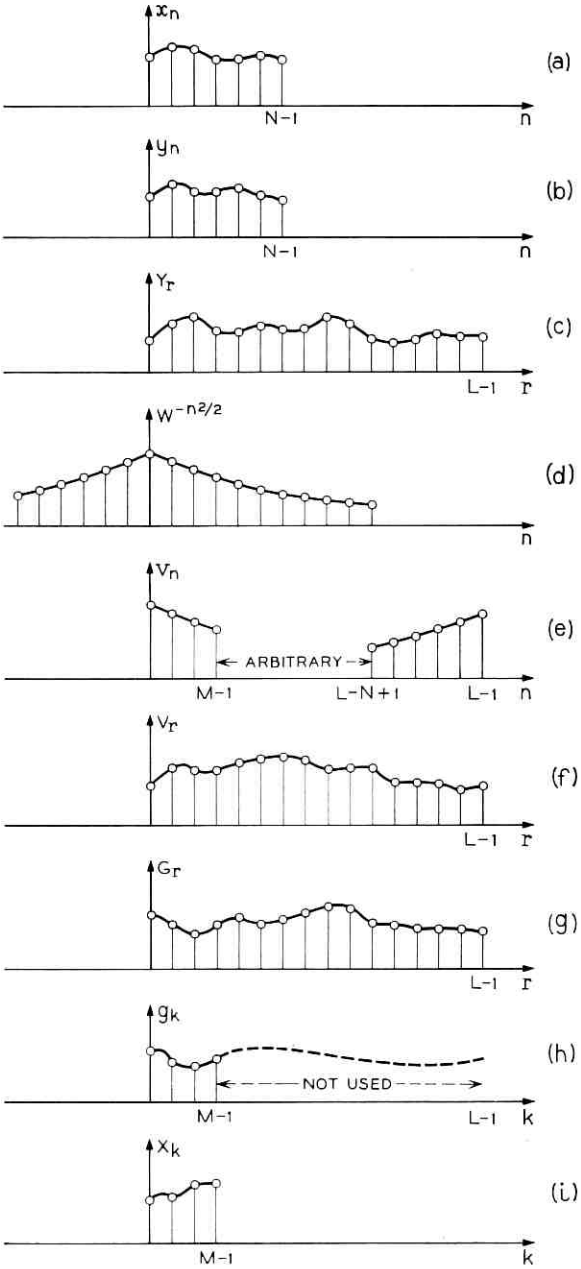


Fig. 4 — Schematic representation of the various sequences involved in the chirp z-transform algorithm: (a) input sequence x_n with N values. (b) weighted input sequence $y_n = A^{-n}W^{n^2/2}x_n$. (c) discrete Fourier transform of y_n. (d) values of the indefinite sequence $W^{-n^2/2}$. (e) sequence v_n formed appropriately from segments of $W^{-n^2/2}$. (f) discrete Fourier transform of v_n. (g) product $G_r = Y_r \cdot V_r$. (h) inverse discrete Fourier transform of G_r. (i) desired M values of the z-transform.

(*viii*) We multiply g_k by $W^{k^2/2}$ to give us the desired X_k:

$$X_k = W^{k^2/2} g_k , \qquad k = 0, 1, 2, \cdots , M - 1.$$

The g_k for $k \geqq M$ are discarded.

Figure 4 represents typical waveforms (magnitudes shown, phase omitted) involved in each step of the process.

III. FINE POINTS OF THE COMPUTATION

3.1 *Operation Count and Timing Considerations*

An operation count can be made, roughly, from the eight steps just presented:

(*i*) We assume that step *i*, that is, choosing L, is a negligible operation.

(*ii*) Forming y_n from x_n requires N complex multiplications, not counting the generation of the constants $A^{-n}W^{n^2/2}$. The constants may be prestored, computed as needed, or generated recursively as needed. The recursive computation would require two complex multiplications per point.

(*iii*) An L point discrete Fourier transform requires a time $k_{\text{FFT}}L \log_2 L$ for L a power of two, and a very simple fast Fourier transform program. More complicated (but faster) programs have more complicated computing time formulas.

(*iv*), (*v*) The value of v_n is computed for either M or N points, whichever is greater. The symmetry in $W^{-n^2/2}$ permits the other values of v_n to be obtained without computation. Again, v_n can be computed recursively. The fast Fourier transform takes the same time as that in step *iii*. If the same contour is used for many sets of data, V_r need only be computed once, and stored.

(*vi*) This step requires L complex multiplications.

(*vii*) This is another fast Fourier transform and requires the same time as step *iii*.

(*viii*) This step requires M complex multiplications.

As the number of samples of x_n or X_k grows large, the computation time for the chirp z-transform grows asymptotically as something proportional to $L \log_2 L$. This is the same sort of asymptotic dependence of the fast Fourier transform, but the constant of proportionality is bigger for the chirp z-transform because two or three fast Fourier transforms are required instead of one, and because L is greater than N or M. Still, the chirp z-transform is faster than the direct computation of equation (10) even for relatively modest values of M and N of the order of 50.

3.2 *Reduction in Storage*

The chirp z-transform can be put into a more useful form for computation by redefining the substitution of equation (11) to read

$$nk = \frac{(n - N_o)^2 + k^2 - (k - n + N_o)^2 + 2N_o k}{2}.$$

Equation (12) can now be rewritten as

$$X_k = W^{k^2/2} W^{N_o k} \sum_{n=0}^{N-1} x_n A^{-n} W^{(n-N_o)^2/2} W^{-(k-n+N_o)^2/2}.$$

The form of the new equation is similar to equation (12) in that the input data x_n are preweighted by a complex sequence $(A^{-n}W^{(n-N_o)^2/2})$, convolved with a second sequence $(W^{-(n-N_o)^2/2})$, and postweighted by a third sequence $(W^{k^2/2}W^{N_o k})$ to compute the output sequence X_k . However, there are differences in the detailed procedures for realizing the chirp z-transform. The input data x_n can be thought of as having been shifted by N_o samples to the left. For example, x_o is weighted by $W^{N_o^2/2}$ instead of W^0. The region over which $W^{-n^2/2}$ must be formed, in order to obtain correct results from the convolution, is

$$-N + 1 + N_o \leqq n \leqq M - 1 + N_o \, .$$

By choosing $N_o = (N - M)/2$ it can be seen that the limits over which $W^{-n^2/2}$ is evaluated are symmetric; that is, $W^{-n^2/2}$ is a symmetric function in both its real and imaginary parts. (Therefore, the transform of $W^{-n^2/2}$ is also symmetric in both its real and imaginary parts.) It can be shown that by using this special value of N_o only $(L/2 + 1)$ points of $W^{-n^2/2}$ need be calculated and stored, and these $(L/2 + 1)$ complex points can be transformed using an $L/2$ point transform*. Hence the total storage required for the transform of $W^{-n^2/2}$ is $L + 2$ locations.

The only other modifications to the detailed procedures for evaluating the chirp z-transform presented in Section II are:

(*i*) following the L point inverse discrete Fourier transform of step *vii*, the data of array g_k must be rotated to the left by N_o locations,

(*ii*) the weighting factor of the g_k is $W^{k^2/2}W^{N_o k}$ rather than $W^{k^2/2}$.

* The technique for transforming two symmetric L point sequences using one $L/2$ point fast fourier transform was demonstrated by J. Cooley at the fast Fourier transform workshop, Arden House, Harriman, New York, October 1968. The appendix summarizes this technique.

The additional factor $W^{N_o k}$ represents a data shift of N_o samples to the right, thus compensating for the initial shift and keeping the effective positions of the data invariant to the value of N_o used.

Now we can estimate the storage required to perform the chirp z-transform. Assuming that the entire process is to take place in core, storage is required for V_r which takes $L + 2$ locations, for y_n, which takes $2L$ locations, and perhaps for some other quantities which we wish to save, such as the input or values of $W^{n^2/2}$ or $A^{-n}W^{n^2/2}$.

3.3 *Additional Considerations*

Since the chirp z-transform permits $M \neq N$, it is possible that occasions will arise where $M \gg N$ or $N \gg M$. In these cases, if the smaller number is small enough, the direct method of equation (10) is called for. However, if even the smaller number is large it may be appropriate to use the methods of sectioning described by Stockham.[3] Either the lap-save or lap-add methods may be used. Sectioning may also be used when problems, too big to be handled in core memory, arise. We have not actually encountered any of these problems and have not programmed the chirp z-transform with provision for sectioning.

Since the contour for the chirp z-transform is a straight line segment in the s-plane, it is apparent that repeated application of the chirp z-transform can compute the z-transform along a contour which is piecewise spiral in the z-plane or piecewise linear in the s-plane.

Let us briefly consider the chirp z-transform algorithm for the case of z_k all on the unit circle. This means that the z-transform is like a Fourier transform. Unlike the discrete Fourier transform, which by definition gives N points of transform for N points of data, the chirp z-transform does not require $M = N$. Furthermore the z_k need not stretch over the entire unit circle but can be equally spaced along an arc. Let us assume, however, that we are really interested in computing the N point discrete Fourier transform of N data points. Still the chirp z-transform permits us to choose any value of N, highly composite, somewhat composite, or even prime, without strongly affecting the computation time. An important application of the chirp z-transform may be computing discrete Fourier transforms when N is not a power of two and when the program or special purpose device available for computing discrete Fourier transforms by fast Fourier transform is limited to when N is a power of two.

There is also no reason why the chirp z-transform cannot be extended to the case of transforms in two or more dimensions with similar considerations. The two dimensional discrete Fourier transform becomes a two dimensional convolution which can be computed by fast Fourier transform techniques.

Caution: For the ordinary fast Fourier transform the starting point of the contour is still arbitrary; merely multiply the waveform x_n by A^{-n} before using the fast Fourier transform and the first point on the contour is effectively moved from $z = 1$ to $z = A$. However, the contour is still restricted to a circle concentric with the origin. The angular spacing of z_k for the fast Fourier transform can also be controlled to some extent by appending zeros to the end of x_n before computing the discrete Fourier transform (to decrease the angular spacing of the z_k) or by choosing only P of the N points x_n and adding together all the x_n for which the n are congruent modulo P; that is, wrapping the waveform around a cylinder and adding together the pieces which overlap (to increase the angular spacing).

IV. APPLICATIONS OF THE ALGORITHM

Because of its flexibility, the chirp z-transform algorithm discussed in the Section III has many potential applications.

4.1 *Enhancement of Poles*

One advantage of the chirp z-transform algorithm over the fast Fourier transform is its ability to evaluate the z-transform at points both inside and outside the unit circle. This is important in the investigation of systems whose transfer functions can be represented as ratios of polynomials in z; that is, in finding poles and zeros of a linear system. By evaluating the transform off the unit circle, one can make the contour pass closer to the poles and zeros of the system, thus effectively reducing the bandwidths and sharpening the transfer function.

For example, a five-pole system was simulated at a 10 kHz sampling frequency. The poles were located at center frequencies of 270, 2290, 3010, 3500 and 4500 Hz with bandwidths of 30, 50, 60, 87 and 140 Hz, respectively. The z-plane pole positions are shown in Fig. 5. (Those familiar with speech will recognize these numbers as resonance positions appropriate for the vowel *i* in the word *beet.*) The input to the system was a periodic impulse train of period 100 samples; that is, 100 pulses per second. Impulse invariant techniques were used to simulate the system.[6] The z-transform of one period of steady state data (100 samples) was evaluated on two spirals outside the unit circle, one on the unit circle, and two spirals inside the unit circle. Figure 6 shows the five contours as they would appear in the s-plane and the s-plane pole positions. The contours are seen to be equi-

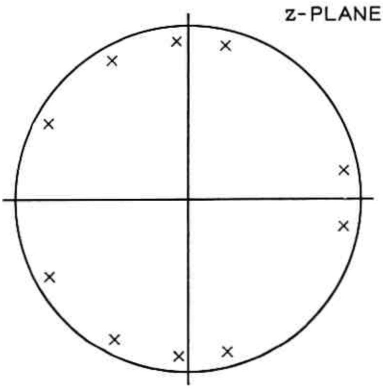


Fig. 5 — Representation of the z-plane locations of the poles of the linear system simulated in the text.

angularly spaced. The five sets of magnitude curves are shown in Fig. 7. The transform was evaluated at 50 equally spaced points from 0 to 4900 Hz, corresponding to $\varphi_o = -1/100$. The sharpening of the magnitude response in the region of the poles is quite pronounced. Figure 6 indicates that contour 5 should be near optimum since it intersects three of the poles.

This example is a somewhat idealized case in that spectral samples were taken every 100 Hz; that is, at the harmonics of the fundamental frequency. Figure 8 shows the case for spectral data taken every 25 Hz on contour 5 of Fig. 6, along with the case where the spectral resolution is the same as shown in Fig. 7. This figure places in evidence the true nature of the z-transform of a finite number of samples. It is clear from equation (2) that $X(z)$ has no poles anywhere in the z-plane except at $z = 0$. There are instead $N-1$ zeros which manifest themselves in the ripples seen in the upper curve of Fig. 8. In many cases the poles of the original system which generated the samples are still in evidence because the zeros tend to be arrayed at approximately equal angular increments except at the locations of the original poles. Hence a pole usually manifests itself by an absence of zeros in the vicinity of that pole in the z-plane. Zeros of transmission are often masked by these effects when only a finite number of samples are transformed. Examples of this effect are given after equation (23).

It is of interest to examine the ability of the chirp z-transform algorithm to determine the bandwidth of a pole as well as its center frequency. To investigate this point, synthetic samples were generated with the bandwidth of the lowest pole (B_1) variable from 10 to 320 Hz by factors of 2; all other bandwidths and center frequencies were held at values used in the previous example. Again a fixed 100 pulse per second source excited each of the systems. Figure 9 illustrates the six sets of poles and three contours used in the investigation. Contour 3 extends into the right half-plane (spiral outside the unit circle) and is only close to the lowest pole. Contour 2 corresponds to the unit circle in the z-plane (that is, the discrete Fourier transform). Contour 1 is an appropriate left-half plane contour (spiral inside the unit circle) used for investigating this system. The resulting set of 18 magnitude curves (six different sets of poles and three contours) are shown in Fig. 10. The rows of Fig. 10 show magnitude curves with a fixed bandwidth and variable contour, whereas the columns show curves on the same contour with variable bandwidths. There are 801 points plotted in each curve in the range 0 to 5,000 Hz.

Looking down any column it is seen that as B_1 increases, the level of the first resonance decreases steadily. The variation in fine spectral detail resulting from the distribution of zeros in the neighborhood of the poles of the original system is seen clearly in column 1. For ex-

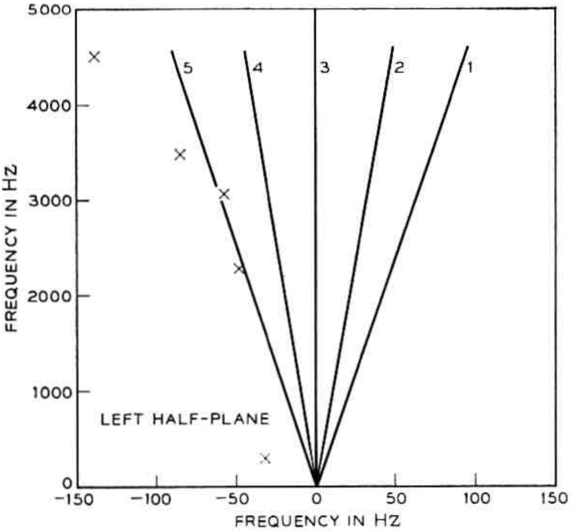


Fig. 6—The s-plane locations of the poles of Fig. 5 and five contours for evaluation of the z-transform.

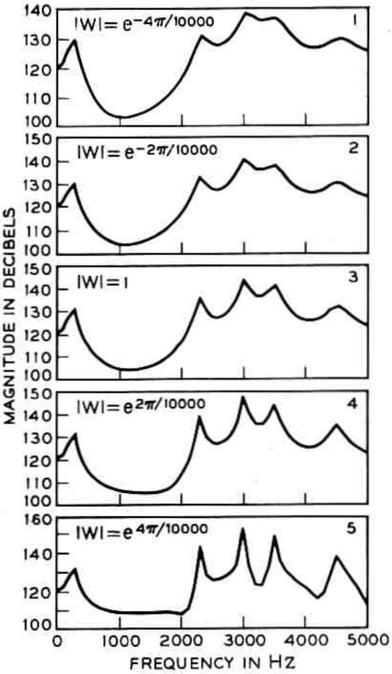


Fig. 7 — Magnitude curves corresponding to evaluation of the z-transform on the five contours of Fig. 6.

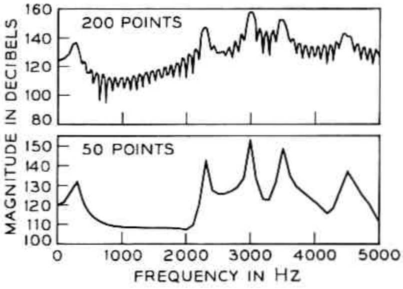


Fig. 8 — A comparison of high resolution and low resolution evaluations of the z-transform. The spacing of points is 25 Hz in the upper curve and 100 Hz in the lower curve.

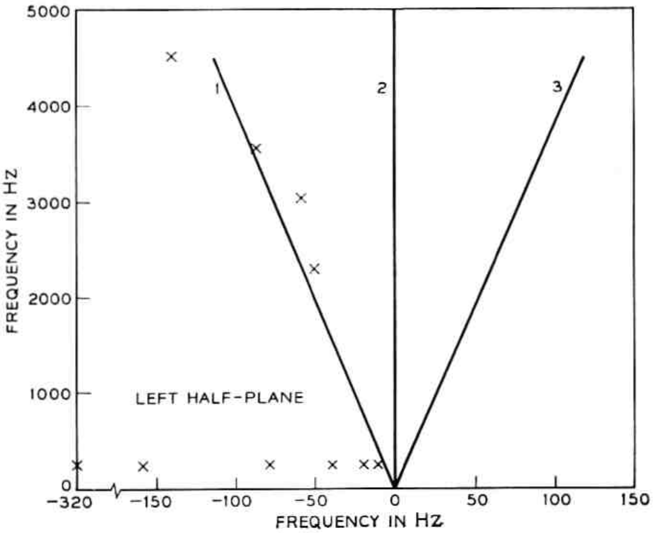


Fig. 9 — The pole locations and contours used to investigate the possibility of bandwidth determination using the chirp z-transform.

ample, the fifth resonance at 4,500 Hz is difficult to find in the upper plots and almost missing in the lower plots, because of the presence of a zero at the pole position. Furthermore, the frequency at which the magnitude is minimum, that is, the closest zero to contour 1 shifts from 2,500 to 2,700 to 800 to 1,100 Hz as bandwidth increases.

The plots in columns 2 and 3 show little or no variation from about 2,000 to 5,000 Hz where the appropriate contours are generally far away from the zeros of the distributions. The resonance at 4,500 Hz is always easy to locate on these plots, thus indicating the desirability of both detailed close-up examination of the transform (as on contour 1) and less detailed, further away looks at the magnitude curve (as on contours 2 and 3). The magnitude curves in the regions 0 to 2,000 Hz are still fairly sensitive to the exact zero distribution for contour 2, and slightly sensitive for contour 3. It would appear from Fig. 10 that there are cases when bandwidth can be determined either from the width or the magnitude of the resonance. Further investigation is necessary before quantitative techniques for determining bandwidths are available.

The choice of the optimum contours is highly dependent on the locations of the poles of the original system. In general there is no

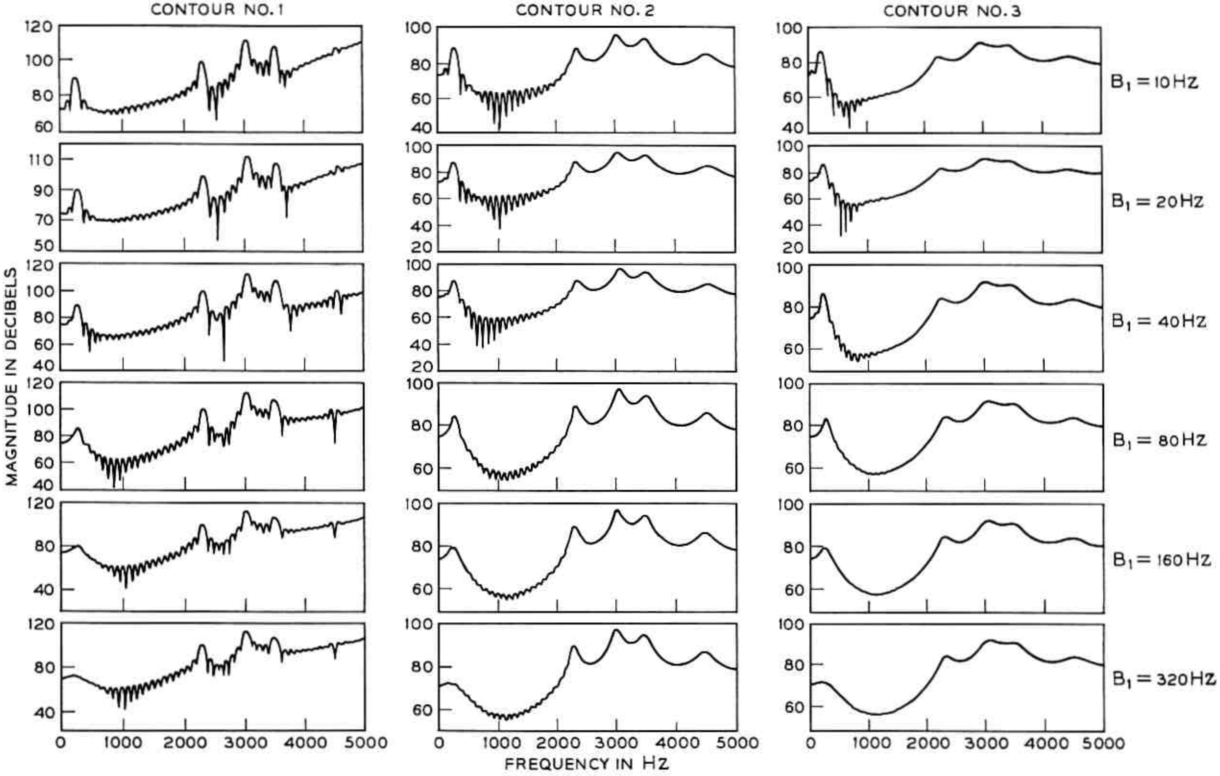


Fig. 10—Magnitude curves corresponding to the 18 different situations of Fig. 9. The rows show magnitude curves for fixed bandwidths and the columns show magnitude curves for fixed contours.

single contour on which all the poles are located since these contours are essentially lines of constant Q = (center frequency)/(bandwidth). Hence the choice of contour is highly dependent on which of the system poles is of most interest in the particular problem. However, some interesting observations can be made from studying magnitude curves for systems whose poles are constant Q poles. Such a system was simulated by keeping the pole center frequencies at the values used previously and setting the Q of each of the poles to 20. A 100 pulse per second impulse train was again used to excite the system and one period of steady state data was analyzed along contours 1 to 7 shown in Fig. 11. The pole positions of the system are shown in this figure and are seen to coincide with contour 6 exactly. The magnitude curves for these seven contours are shown in Fig. 12. These are high resolution spectra containing 801 points from 0 to 5,000 Hz. Notice that both magnitude curves 5 and 6 accentuate the poles equally except in the region of the fifth pole where curve 5 appears slightly better than curve 6. The fact that this occurs is not surprising in view of the fact that the pole is really manifested by an absence of a zero in an array of zeros of approximately the same magnitude and angular spacing.

Another anomaly which can be attributed to the way in which the

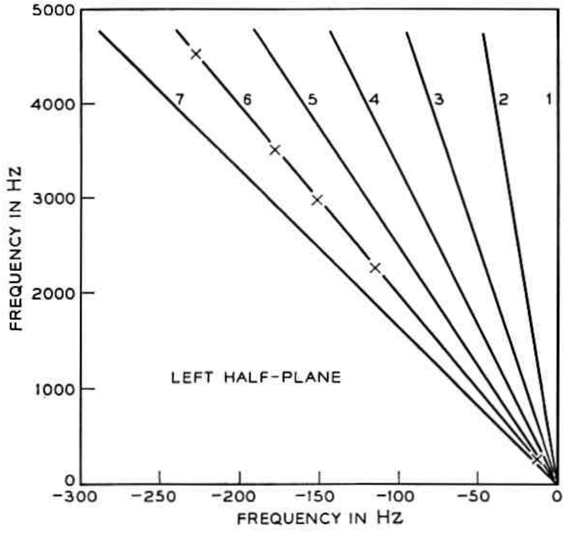


Fig. 11 — The s-plane locations of constant Q poles and contours on which the z-transform was evaluated.

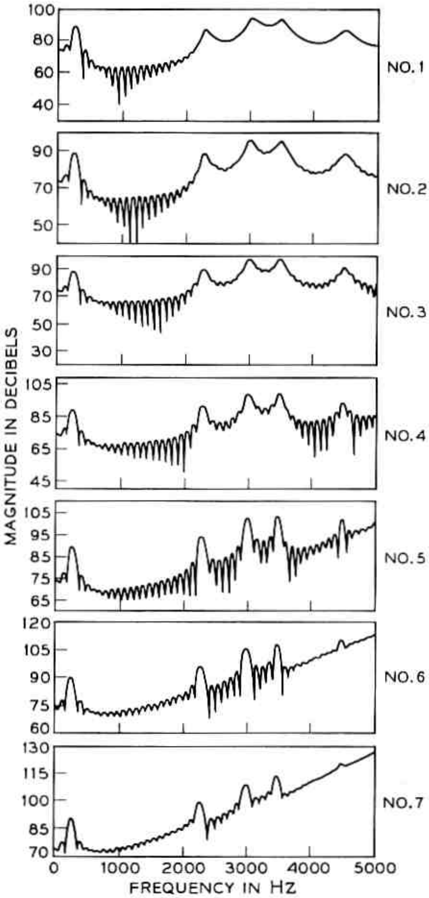


Fig. 12 — Magnitude curves for contours of Fig. 11 (constant Q poles).

zeros are distributed is evidenced by comparing curves 5 and 7. Based on the relative positions of the two contours with respect to the poles we would expect the magnitude curves for these contours to be identical, but the comparison shows that this is not the case in actual computation. This is the result of the fact that the zero distribution is not exactly symmetric so that contours which pass very close to the zeros may look considerably different from one another.

A final point of interest in Fig. 12 is the linear component in the last three curves which dominates at high frequencies. This effect is also shown in Fig. 14. Figure 13 shows the five contours used in obtaining the log magnitude plots in Fig. 14. It is clear that when the contour passes inside the original pole locations (and therefore inside the array of zeros in the z-plane), the log magnitude function exhibits a definite linear component. This effect is easily explained when $X(z)$ is written in the form

$$X(z) = Dz^{-(N-1)} \prod_{r=1}^{N-1} (1 - a_r^{-1}z). \tag{19}$$

where the a_r's are the zeros of $X(z)$. If we evaluate equation (19) at $z = W_o^{-k} \exp(-j2\pi\varphi_o k)$ we obtain for the magnitude

$$|X_k| = |D| W_o^{k(N-1)} \prod_{r=1}^{N-1} |[1 - a_r^{-1}W_o^{-k} \exp(-j2\pi\varphi_o k)]|. \tag{20}$$

For plotting in dB we define

$$20 \log_{10} |X_k| = 20 \log_{10} |D| + 20(N-1)k \log_{10} W_o + \sum_{r=1}^{N-1} 20 \log_{10} |[1 - a_r^{-1}W_o^{-k} \exp(-j2\pi\varphi_o k)]|. \tag{21}$$

In the examples we have shown, almost all of the zeros a_r have magnitudes slightly less than 1. Thus for contours inside these zeros (cor-

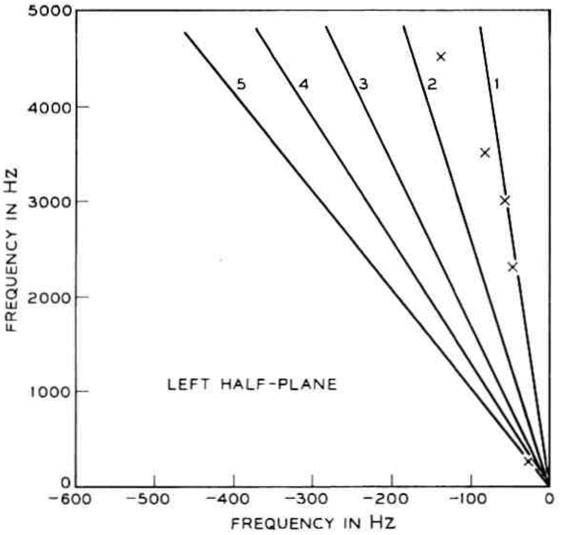


Fig. 13 — Contours and pole locations used to study the effect of passing inside the pole locations.

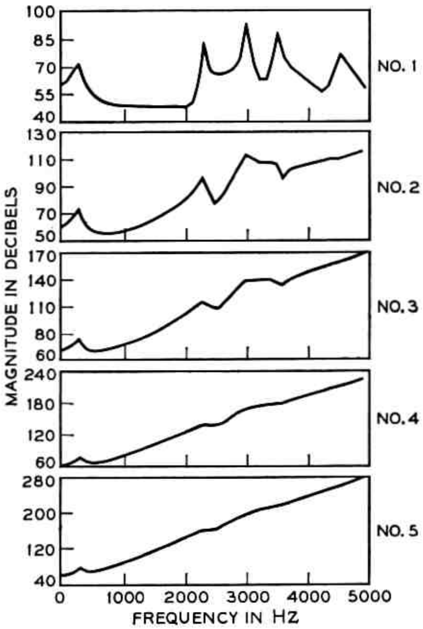


Fig. 14 — Magnitude curves for the contours of Fig. 13 showing a large linear component resulting from the $N-1$ poles of $X(z)$ at $z=0$.

responding to contours 2 through 5 in Fig. 13), W_o is greater than 1. Thus each term in the sum on the right side of equation (21) tends to decrease as k gets larger. In contrast, the second term on the right side of equation (21) represents a linear component with slope equal to 20 $(N-1)\log_{10} W_o$.

In Fig. 14, the mth curve corresponds to a value of $W_o = e^{4\pi m/10{,}000}$, $m = 1, 2, 3, 4, 5$. The value of φ_o is $-1/100$, N is 100, and the sampling rate is 10 kHz. Thus the frequency going from 0 to 5 kHz corresponds to k going from 0 to 50. For example, in the fifth curve $W_o = e^{20\pi/10{,}000}$ and the slope should be

$$20(N-1)\log_{10} W_o = \frac{20(99)20\pi \log_{10} e}{10{,}000} = 5.4. \qquad (22)$$

Thus the total dB change in going from 0 to 5 kHz should be on the order of $50(5.4) = 270$ dB. In Fig. 15a we show this case again. In Fig. 15b we show the result of evaluating

$$\sum_{n=-(N-1)}^{0} x_{n+N-1} z^{-n} = z^{(N-1)} X(z), \qquad (23)$$

using the same value of W_o and φ_o. Notice that this should remove the second term in equation (21) leaving the other terms unaffected. This observation is substantiated by Fig. 15b since the value at 5 kHz is very nearly 270 dB less than in Fig. 15a. Notice also that some of the resonances are still in evidence although not as clearly defined because the contour is passed relatively far inside the zeros of $X(z)$.

Another interesting question was investigated using this technique. As shown above, when the response of a linear system is truncated by repetitively pulsing the system and transforming a finite number of samples, the z-transform has only zeros except for poles at $z = 0$. However, the poles of the original system function can still be located in magnitude response curves by the absence of zeros in the appropriate regions. The question arises about what happens when the system function contains zeros. Suppose $h(nT)$ is the impulse response of a linear system with both poles and zeros in its z-transform $H(z)$. Since $H(z)$ has poles, $h(nT)$ will be an infinite sequence. There is clearly no reason to expect that the transform of only N of these samples will have zeros at the same location as the zeros of $H(z)$. However, the system zeros can be expected to have an effect on the distribution of zeros of the truncated z-transform.

To illustrate this point, the system response used in the above examples was modified by passing the output waveform through a system whose transfer function consisted of a complex conjugate zero pair. A periodic 100 pulse per second source was again used to excite the system and one period of steady state data was analyzed. The system pole-zero pattern and the contours of analysis are shown in

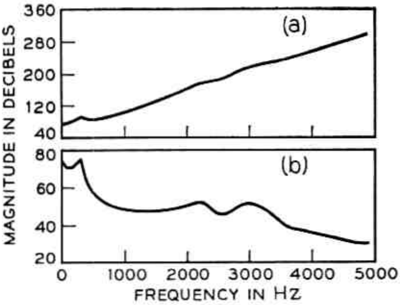


Fig. 15 — Magnitude curves obtained by evaluation of the z-transform on contour 5 of Fig. 13: (a) with the effect of the $N - 1$ poles at $z = 0$; (b) with the $N - 1$ poles removed by shifting the sequence x_n by N positions to the left.

Fig. 16. In one simulation a zero was placed at point *A* (500 Hz, 12.5 Hz), and in a second simulation the zero was at point *B*, (2,500 Hz, 60 Hz). The analysis was made at 801 points from 0 to 5,000 Hz along these contours. The resulting magnitude curves, along with a set of low resolution curves where the magnitude was computed every 100 Hz from 0 to 5,000 Hz, are shown in Figs. 17 and 18. The data of Fig. 17 are for the case where the transmission zero was at 500 Hz whereas for Fig. 18 the zero was at 2,500 Hz.

The high resolution data of Fig. 17 show no strong indication of the transmission zero; whereas the transmission poles are still very much in evidence. The low resolution data (evaluated at harmonics of the source) does indicate the presence of a zero along contour 1, but along the other contours the case is not so clear. The most unusual observation is that along contour 3, the contour closest to both the transmission zero and the poles, there is little or no indication of the zero; whereas the poles are still strongly in evidence. Along contour 4, at the high frequency end, there is noise in the magnitude spectrum. The source of this noise is discussed in Section V.

The indications from Fig. 17 are that a transmission zero can be more easily located on contours which are far from the zero than on

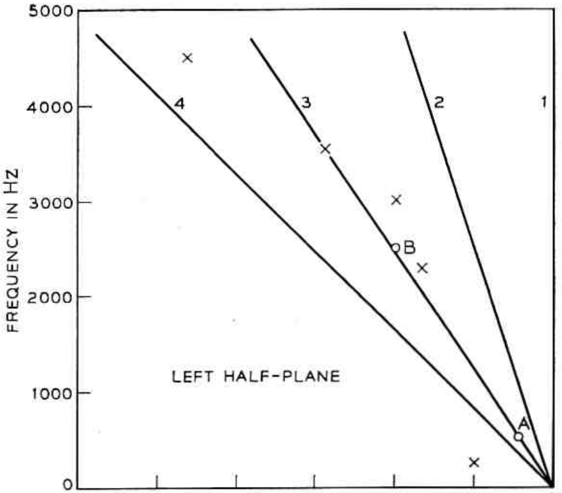


Fig. 16 — The *s*-plane locations of poles and zeros (at *A* and *B*) and contours used in studying the effect of zeroes on the magnitude curves.

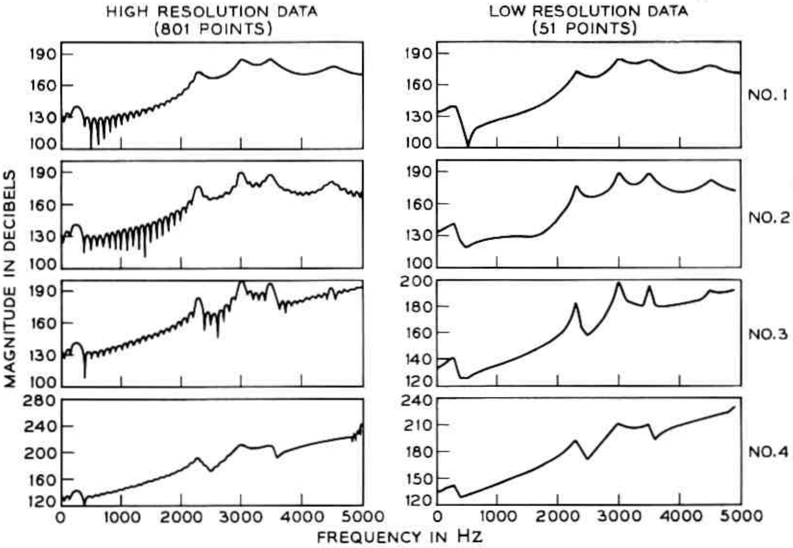


Fig. 17 — Magnitude curves for a zero at 500 Hz (position *A* in Fig. 16).

contours which traverse it. Furthermore it is much easier to locate on a low resolution spectrum than on a high resolution spectrum. Hence zeros, unlike poles, are not generally easy to locate from spectra.

The zero of Fig. 17 was at 500 Hz and in a region where the high resolution spectra displayed a large amount of ripple from the truncation zeros of the data. Figure 18 shows similar magnitude curves for the zero at 2,500 Hz, a region with much less ripple in the spectrum. The magnitude response curves show effects entirely similar to those of Fig. 17. The zero is most easily locatable for contour 1, the standard fast Fourier transform. In contour 3, which again passes through the zero, there is no indication of the zero. Also the low resolution data tends to show the zero better than the high resolution data. One important implication of these results is that one could not use these techniques to accurately find the position of complex transmission zeros. In many cases it would be difficult to differentiate between dips in the spectrum between poles and dips caused by complex zeros, thus indicating the difficulty of locating even the center frequency of a zero.

The chirp *z*-transform algorithm has been applied to the spectral analysis of speech in order to aid in automatic detection of the time varying resonances (poles or formants) of speech. Voiced speech can

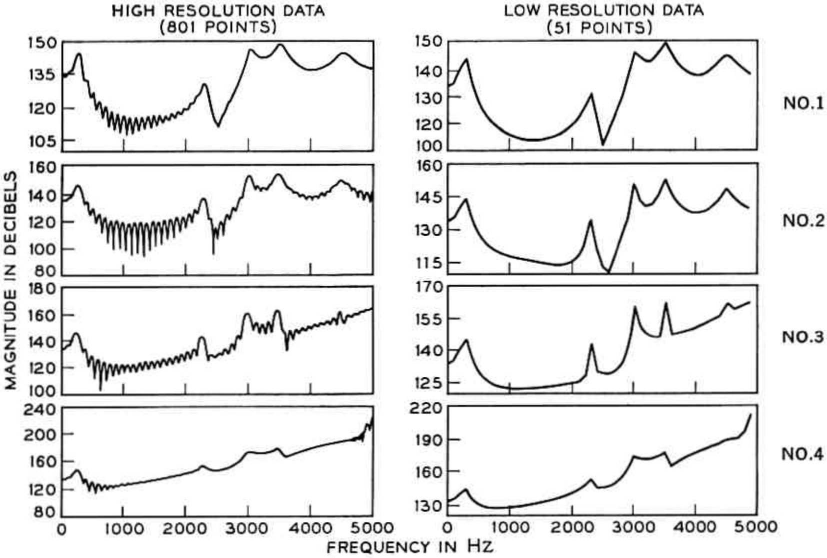


Fig. 18 — Magnitude curves for a zero at 2,500 Hz (position *B* in Fig. 16).

be modelled as the convolution of a source waveform with a vocal tract impulse response. The vocal tract impulse response is essentially a sum of damped exponentials, each exponential corresponding to a mode or pole of the vocal tract transfer function. It is of interest to speech researchers to detect these time varying resonances. The chirp z-transform algorithm has been applied to individual periods of voiced speech with a high degree of success. Figure 19 shows the result of applying the chirp z-transform algorithm along the two contours shown at the upper left of the figure, to a period of voiced speech. The upper contour corresponds to the standard fast Fourier transform contour; the lower to a suitably chosen spiral contour. The magnitude function along the upper contour indicates a single wide peak in the region 2,000 to 2,500 Hz, whereas the magnitude along the lower contour shows two isolated peaks in this region corresponding to the physical knowledge that there actually are supposed to be two peaks in this region. Variations on the chirp z-transform algorithm for spectral analyses of speech have been studied and will be reported on in a subsequent paper.[7]

4.2 *High Resolution, Narrow Band Frequency Analysis*

One very useful application of the chirp z-transform algorithm is the ability to efficiently evaluate high resolution, narrow frequency band spectra. Using standard fast Fourier transform techniques, in order to achieve a frequency resolution of $\leqq \Delta F$, with a sampling frequency of the data of $1/T$, requires $N \geqq 1/(T \cdot \Delta F)$ points. For very small ΔF, this implies very large values of N. The crucial issue is that what is often required is high resolution for a limited range of frequencies and low resolution for the remainder of the spectrum. An example of such a circumstance is the design of band-pass or low-pass filters. Usually what is desired is a microscopic look at details of the frequency response in the pass-band and only a gross look outside the pass-band.

The chirp z-transform algorithm is extremely well suited for such cases since it allows selection of initial frequency and frequency spacing, independent of the number of time samples. Hence high resolution data over a narrow frequency range can be attained at low cost.

To illustrate these points, simple rectangular band-pass filters were simulated by symmetrically truncating a delayed impulse response. The impulse response used was

$$h(nT) = \alpha \sin [\pi(F_2 - F_1)(n - \tfrac{1}{2} - m)T] \cdot \cos [\pi(F_2 + F_1)(n - \tfrac{1}{2} - m)T], \qquad 0 \leqq n \leqq 2m \qquad (24)$$

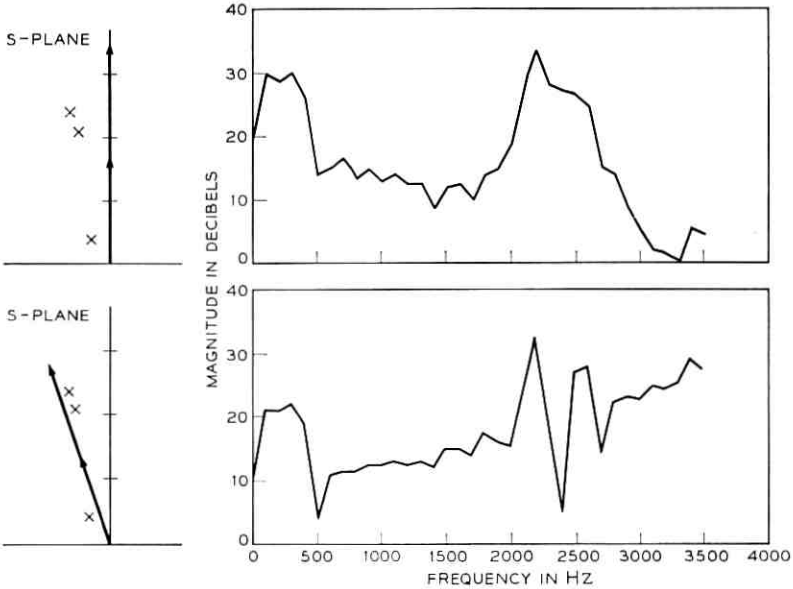


Fig. 19 — Magnitude curves from evaluation of the z-transform of one period of natural speech. The contour for the upper plot is the unit circle in the z-plane while the contour for the lower curve is a spiral inside the unit circle.

where

$2m$ = number of terms in the truncated impulse response
$1/T$ = sampling frequency = 10,000 Hz
F_1 = lower cutoff frequency in Hz
F_2 = upper cutoff frequency in Hz.

Values for m of 100 and 500 were used with $F_1 = 900$ Hz and $F_2 = 1{,}100$ Hz. Figure 20 shows plots of equation (24) for these two cases. A standard 1,600 point fast Fourier transform was calculated and the magnitude response for $m = 100$ is shown in the upper half of Fig. 21. In order to investigate the pass-band and transition region more carefully the chirp z-transform algorithm was used to give a 1.25 Hz resolution over the band from 500 to 1,500 Hz. The contour used was identical to the contour for the fast Fourier transform. The resulting magnitude response curve is shown in the lower half of Fig. 21. To achieve this high a resolution would have required an 8,000 point fast Fourier transform, instead of the 1,000 point transforms actually used. (Similar expansions of regions of the phase curve were made for this filter but are not shown.)

Figure 22 shows similar effects for the case $m = 500$. The applicability of the chirp z-transform algorithm for such frequency expansions is a powerful tool for close examination of small frequency bands, as well as for debugging implementations of digital filters. For example, one could easily check if a desired filter met its design specification of in-band ripple, transition ratio, and so on.[8]

One situation where the chirp z-transform algorithm may be quite useful is when we are confronted with an extremely long sequence for which we desire a fine grained spectrum over a narrow band of frequencies. Suppose we have a sequence of P samples and desire M spectral samples where $M \ll P$. That is, we wish to evaluate

$$X_k = \sum_{n=0}^{P-1} x_n A^{-n} W^{nk}, \qquad k = 0, 1, \cdots, M-1. \tag{25}$$

The sum in equation (25) can be broken up into r sums over N points as follows

$$X_k = \sum_{q=0}^{r-1} A^{-qN} W^{kqN} \left[\sum_{n=0}^{N-1} x_{n+qN} A^{-n} W^{nk} \right], \qquad k = 0, 1, \cdots, M-1 \tag{26}$$

where $rN \geqq P$. Each of the r sums in the brackets can be evaluated using the chirp z-transform algorithm, requiring storage on the order of

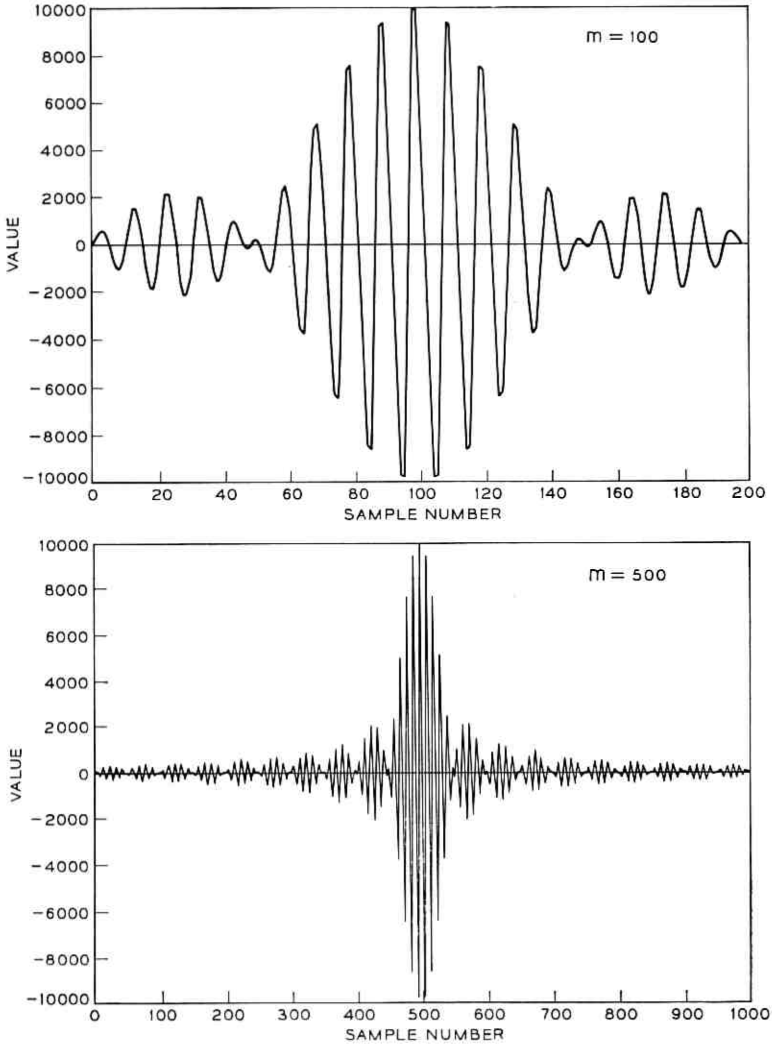


Fig. 20 — The impulse responses of simple band-pass filters.

$3(N + M - 1)$ locations. In addition we require $2M$ locations in which to accumulate the M complex values of the transform. Although 2 fast Fourier transforms and $2M$ complex multiplications are required for each of the r transforms, it is quite possible that a saving in total time may result from this method as opposed to evaluation of a P point transform using auxiliary storage such as drum, disk or tape.

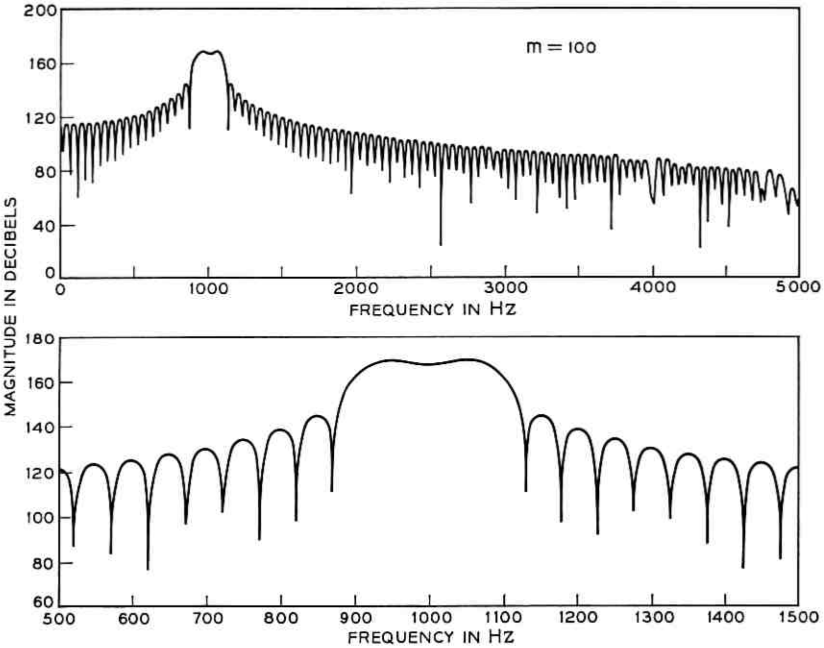


Fig. 21 — Frequency response curves for upper impulse response (200 samples) in Fig. 20. Upper curve obtained with 1,600 point fast Fourier transform (resolution 6.25 Hz). Lower curve obtained with chirp *z*-transform algorithm (1.25 Hz resolution).

4.3 *Time Interpolation or Sampling Rate Changing*

The flexibility of the chirp *z*-transform algorithm for obtaining high resolution in frequency has been explained and illustrated in Section 4.2. A similar procedure applies to interpolation between samples of a bandlimited time function using samples of the frequency spectrum.[9] In this section, we discuss how the discrete Fourier transform can be used to perform interpolation on a set of samples and the advantages and disadvantages of using the chirp *z*-transform algorithm for this.

4.3.1 *Bandlimited Interpolation Using the discrete Fourier transform*

Assume that we have available N samples $x(nT)$, $n = 0, 1, 2, \ldots,$ $N - 1$, of a bandlimited waveform $x(t)$. The sampling interval T is assumed less than or equal to the Nyquist interval. The total time interval spanned by these samples is therefore NT seconds. We wish to obtain equally spaced samples of $x(t)$ at a sampling interval T', where

T' is less than or equal to the Nyquist interval. These samples are denoted by $x(mT')$, $m = 0, 1, \ldots, N' - 1$, where $N'T' = NT$. (Notice that we are assuming N' is an integer. This assumption will be dropped later.)

If all the samples $x(nT)$ are available, the samples $x(mT')$ can be obtained from

$$x(mT') = \sum_{n=-\infty}^{\infty} x(nT) \frac{\sin \frac{\pi}{T}(mT' - nT)}{\frac{\pi}{T}(mT' - nT)}. \tag{27}$$

Thus the interpolation can be viewed as the result of convolving the interpolation function $[\sin(\pi t/T)]/[(\pi/T)\,t]$ with the samples $x(nT)$ and then resampling with period T'. It is well known that convolution may be done using the discrete Fourier transform, and we will show how the resampling can also be affected by properly augmenting the

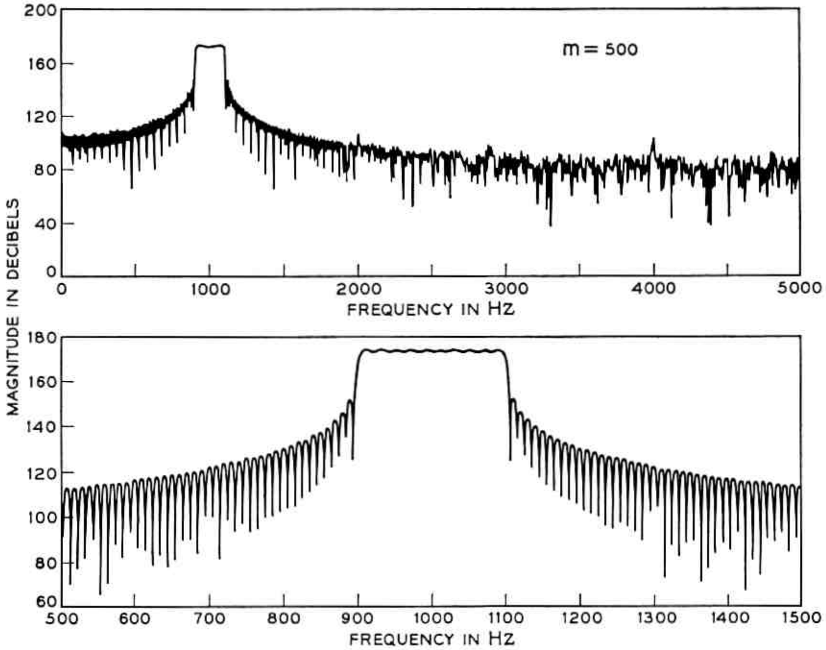


Fig. 22 — Frequency response curves for lower impulse response (1,000 samples) of Fig. 20. Upper curve obtained with 1,600 point fast Fourier transform (resolution 6.25 Hz). Lower curve obtained with chirp z-transform algorithm (1.25 Hz resolution).

transform with zeros. Because the discrete Fourier transform uses only a finite number of samples, we shall encounter errors similar (but not identical) to using only N terms in equation (27).

The discrete Fourier transform of the given samples is

$$X_N(k) = \sum_{n=0}^{N-1} x(nT) \exp\left(-j \frac{2\pi}{NT} knT\right), \qquad k = 0, 1, \cdots, N-1. \tag{28}$$

(Notice that we have changed notation in this section in order to make explicit the number of samples and the sampling period.) We define

$$X_N'(k) = X_N(k) H_N(k), \tag{29}$$

where $H_N(k)$ is the N point discrete Fourier transform of the interpolation function to be convolved with the samples $x(nT)$. [Notice that this convolution is equivalent to cyclic convolution of a periodic impulse response $h(nT)$ with the samples $x(nT)$.]

In order to change the sampling to period T', we split $X_N'(k)$ about $k = N/2$ and expand (by inserting zeros) or contract (by discarding zeros) the transform according to the following equations

$$\begin{aligned}
X_{N'}'(k) &= X_N'(k) && \begin{cases} 0 \leqq k < N'/2 & N' < N \\ 0 \leqq k < N/2 & N' > N \end{cases} && \text{(30a)}\\
&= 0 && \qquad k = N'/2 \qquad\qquad N' < N && \text{(30b)}\\
&= \tfrac{1}{2} X_N'(k) && \qquad k = N/2 \qquad\qquad N' > N && \text{(30c)}\\
&= 0 && N/2 < k < N' - N/2 \qquad N' > N && \text{(30d)}\\
&= \tfrac{1}{2} X_N'(k - N' + N) && \qquad k = N' - N/2 \qquad N' < N && \text{(30e)}\\
&= X_N'(k - N' + N) && \begin{cases} N'/2 < k < N' & N' < N \\ N' - N/2 < k < N' & N' > N \end{cases}. && \text{(30f)}
\end{aligned}$$

Equations (30b), (30c), and (30e) are required only when N' and N are even integers and equation (30d) is required only when $N' > N$.

The N' point inverse discrete Fourier transform of $X_{N'}'(k)$ is defined to be

$$x'(mT') = \frac{1}{N'} \sum_{k=0}^{N'-1} X_{N'}'(k) \exp\left(j \frac{2\pi}{N'T'} kmT'\right), \qquad m = 0, 1, \cdots, N'-1. \tag{31}$$

For example, if N is even and $N' > N$, we can show using equations (28), (29), and (30) that

$$x'(mT') = \frac{1}{N'} \sum_{k=-N/2}^{N/2} H_N(k)\left[\sum_{n=0}^{N-1} x(n) \exp\left(-j\frac{2\pi}{NT}knT\right)\right] \cdot \exp\left(j\frac{2\pi}{N'T'}kmT'\right), \tag{32}$$

where $m = 0, 1, \cdots, N' - 1$, and the terms corresponding to $k = \pm N/2$ are understood to be multiplied by $1/2$ since N is even. By interchanging the order of summation and using the fact that $N'T' = NT$, we obtain

$$x'(mT') = \frac{N}{N'} \sum_{n=0}^{N-1} x(nT)h(mT' - nT) \tag{33}$$

where

$$h(mT' - nT) = \frac{1}{N} \sum_{k=-N/2}^{N/2} H_N(k) \exp\left[j\frac{2\pi}{NT}k(mT' - nT)\right]. \tag{34}$$

Notice that equation (33) has the desired form for interpolation [see equation (27)], however the values of $x'(mT')$ are clearly not exactly equal to the desired interpolated values $x(mT')$. This is so because only N samples are used and because of the form of $h(mT - nT)$. As an example, suppose

$$H_N(k) = \begin{cases} 1 & -N/2 < k < N/2 \\ \frac{1}{2} & k = \pm N/2 \end{cases}. \tag{35}$$

(This is equivalent to splitting $X_N(k)$ at $k = N/2$ and inserting $N' - N$ zeros between the two halves of the transform). If we evaluate equation (34) for this case, we obtain

$$h(mT' - nT) = \frac{\sin\frac{\pi}{T}(mT' - nT)}{N\tan\frac{\pi}{NT}(mT' - nT)}. \tag{36}$$

This function is plotted in Fig. 23a where $\theta = (mT' - nT)/T$, and $N = 8$. Clearly

$$h(mT' - nT) \approx \frac{\sin\frac{\pi}{T}(mT' - nT)}{\frac{\pi}{T}(mT' - nT)} \tag{37}$$

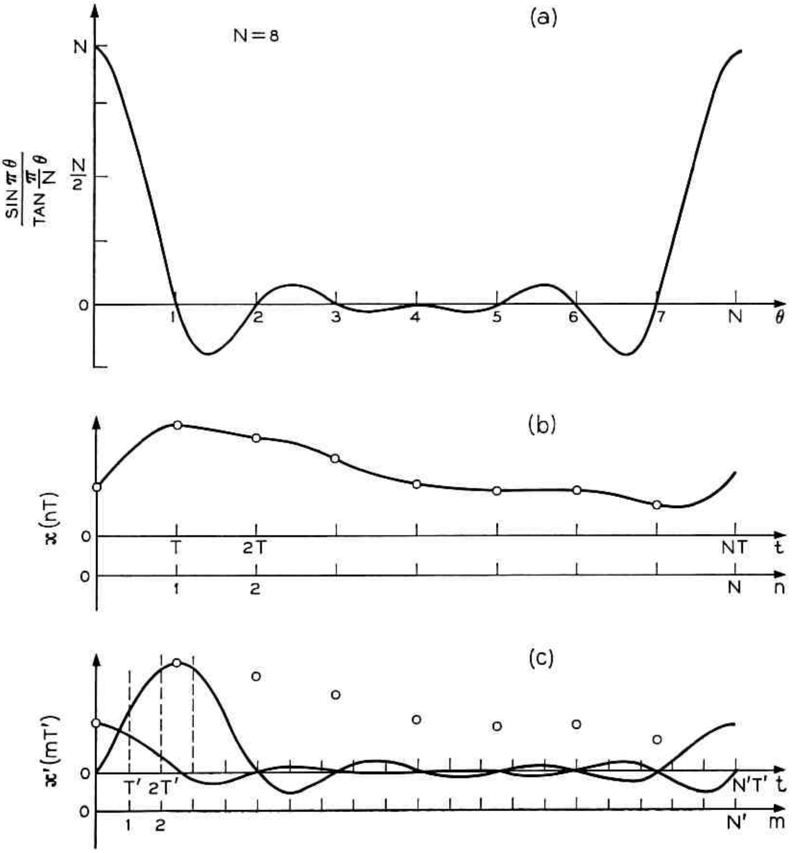


Fig. 23 — An illustration of bandlimited interpolation using the discrete Fourier transform: (a) the periodic function which is convolved with the original samples; (b) a bandlimited time function showing samples with spacing T; (c) How the interpolated values $x'(mT')$ are formed.

when $\pi(mT' - nT)/NT$ is small, that is, in a region where $h(mT' - nT)$ is significantly different from zero. Figure 23b shows a segment of a waveform $x(t)$ and samples $x(nT)$. In Fig. 23c, we have shown just two of the terms in equation (33). This figure places in evidence the nature of the interpolation which is performed. The errors are likely to be greatest at either end of the segment, since the interpolated values at one end depend on the samples at the other end in a way which is not at all consistent with equation (27). The error caused by this effect will be most significant in the regions $0 \leqq mT' < 2T$ and

$T(N - 2) \leqq mT' \leqq TN$. The remainder of the values will have essentially the same error associated with using only N terms in equation (27).

Notice that equation (36) is not the only interpolation function which can be used. Other choices of $H_N(k)$ may lead to interpolation functions which are in some sense more desirable. For example, Fig. 24 shows four different choices for $H_N(k)$ and their associated interpolation functions or impulse responses. (The impulse responses were shifted modulo N' as an aid in plotting.) It can be seen from Fig. 24, that removing the sharp cutoff in $H_N(k)$ greatly shortens the effective duration of the impulse response, thus tending to minimize the end effects discussed previously. Clearly the approximation to $(\sin \pi t)/\pi t$ interpolation is not as good as equation (36), but in many cases such smoothing of the interpolated values may not be objectionable.

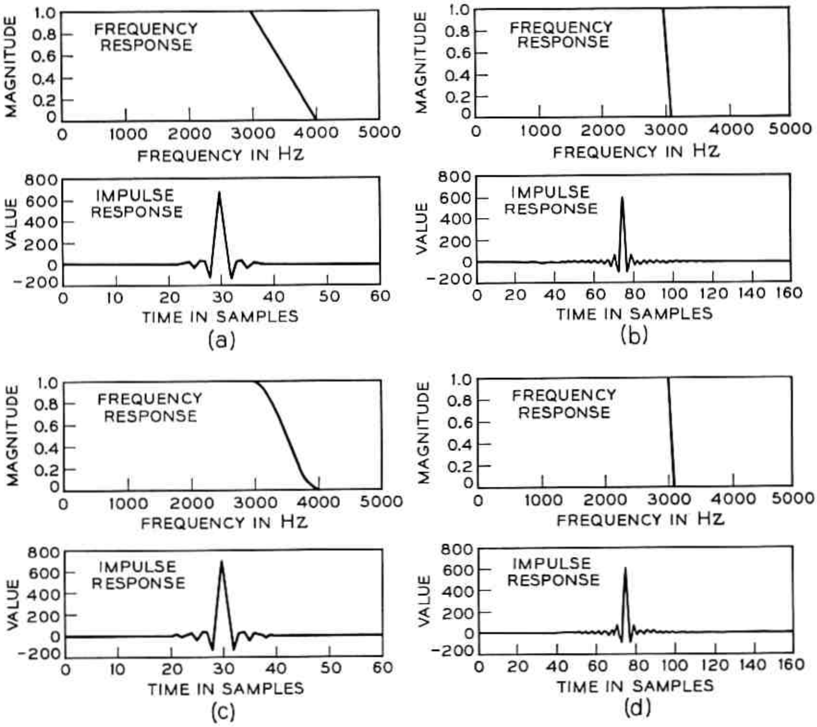


Fig. 24 — A set of four simple frequency responses and corresponding impulse responses which could be used for interpolation.

4.3.2 *Computational Considerations in Bandlimited Interpolation*

In Section 4.3.1 we discussed a method of bandlimited interpolation based on the discrete Fourier transform. The operations involved are summarized in Fig. 25. The sequence $\{X_N(k)\}$ may be evaluated using the fast Fourier transform. In this case, N will be restricted to a value compatible with the available fast Fourier transform routine, for example, N would be highly composite. The transform is then multipled by $H_N(k)$ and expanded or contracted according to equation (30). Then we must compute the inverse discrete Fourier transform with N' points. This can be done using the fast Fourier transform provided that

(*i*) $N' = NT/T'$ is an integer and compatible with the available fast Fourier transform routine.

(*ii*) Enough high speed storage (a minimum of N' locations) is available.

[Notice that *i* applies for either $N' > N$ or $N' < N$ while *ii* will probably not be a problem except when $N' \gg N$.]

In many cases it may not be possible to meet one or both of the above conditions; then the chirp z-transform algorithm can be very useful. The N' point inverse discrete Fourier transform may be computed using $W_o = 1$ and $\varphi_o = +\ 1/N'$, where N' need not even be an integer. Thus we can compute M interpolated values using

$$x'(mT) = \frac{1}{N'} W^{m^2/2} \sum_{k=-N/2}^{N/2} [X'_{N'}(k) W^{k^2/2}] W^{-(m-k)^2/2} \tag{38}$$

where $X'_{N'}(k)$ is determined by equation (30) and

$$X'_{N'}(-k) = X'_{N'}(N' - k), \qquad k = 1, 2, \cdots, N/2. \tag{39}$$

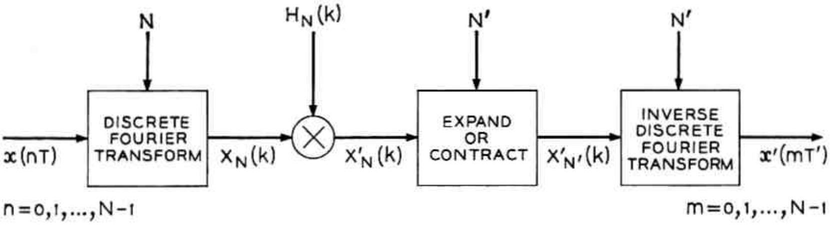


Fig. 25 — Illustration of the steps involved in bandlimited interpolation using the discrete Fourier transform.

Assuming that the transform of $W^{-k^2/2}$ is available, equation (38) can be evaluated using two L point fast Fourier transforms where L is the smallest integer which is greater than $N + M - 1$ and which is compatible with an available fast Fourier transform routine.

Alternatively we can evaluate

$$y'(mT') = \frac{1}{N'} W^{m^2/2} \sum_{n=0}^{N/2} [Y'_{N'}(k)W^{k^2/2}]W^{-(m-k)^2/2} \tag{40}$$

where

$$Y'_{N'}(k) = \begin{cases} X'_{N'}(k) & k = 0 \\ 2X'_{N'}(k) & 0 < k \leqq N/2 \end{cases}. \tag{41}$$

It can be shown that

$$y'(mT') = x'(mT') - j\hat{x}'(mT') \tag{42}$$

where $\hat{x}'(mT')$ is the inverse discrete Fourier transform of

$$\hat{X}'_{N'}(k) = j\,\mathrm{sgn}_{N'}\,(k)\cdot X'_{N'}(k) \tag{43}$$

and

$$\mathrm{sgn}_{N'}\,(k) = \begin{cases} 0 & k = 0, N'/2 \\ 1 & 0 < k < N'/2. \\ -1 & N'/2 < k < N' - 1 \end{cases} \tag{44}$$

From equation (41) and (44) it can be shown that $\hat{x}'(mT')$ is an approximation to the Hilbert transform of $x'(mT)$. In this case we require at least $(N/2 + M - 1)$ point transforms to compute M interpolated values. This is at the expense of not being able to do two interpolations at once as is possible with equation (38) (obtaining one interpolation as a real output and one as an imaginary output); however we do obtain an approximation to the Hilbert transform of $x'(mT')$ which may be of value in some applications.

If sufficient core storage is not available to compute an N' point fast Fourier transform, we can compute the interpolated values in sections and piece these sections together as is commonly done in high speed convolution. The chirp z-transform algorithm allows us to compute as many as $2M$ interpolated points at a time, where M can be chosen so that the fast Fourier transforms can be done using only core storage. Probably the most significant advantage, though, is the ability to efficiently interpolate to arbitrary sampling intervals.

As an example of the ideas discussed in this section, consider the waveforms shown in Fig. 26. Figure 26a shows 500 samples of a speech waveform where the sampling rate was 20 kHz. ($T = 5 \times 10^{-5}$ second). The samples are connected by straight lines in the figure. Figure 26b shows the 500 samples of the waveform in (a) after filtering with a nonrecursive filter of the type shown in Fig. 24, whose gain was zero after 3.2 kHz. Figure 26c shows 160 samples of the result of a change of sampling rate from 20 kHz to 6.4 kHz. The value of N was 700 and $N' = (6{,}400)(700)/20{,}000 = 224$. It is difficult to judge quantitatively from such a figure the accuracy of the interpolation. It does seem safe to conclude that the error is not extreme. Our experience has been that there is significant error only in the first and last few samples of the N' output samples. Using the chirp z-transform algorithm, these "bad" samples need not ever be computed, for example, only M "good" values need be computed.

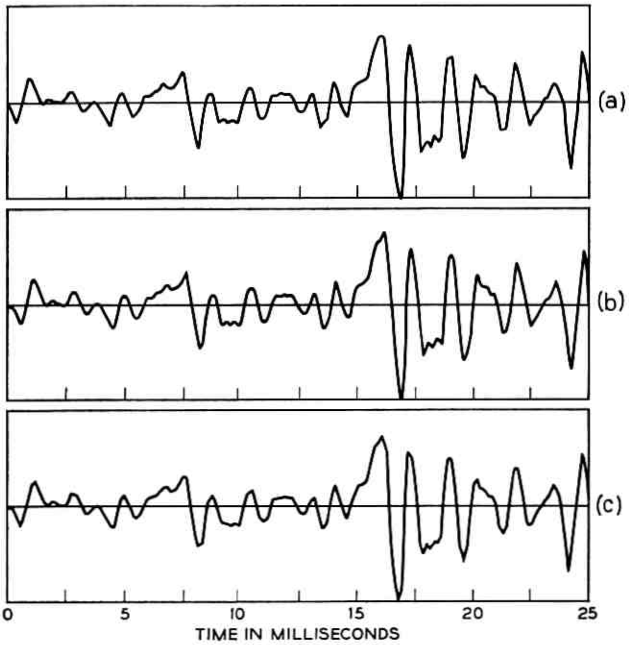


Fig. 26 — An example of interpolation for the purpose of changing the sampling rate: (a) 500 samples of speech at 20 kHz sampling rate; (b) 500 samples of (a) after low pass filtering to 3.2 kHz; (c) 160 samples of (a) after changing the sampling rate to 6.4 kHz using the chirp z-transform. (In all cases the samples are connected by straight lines).

If one wishes to low-pass filter a waveform and then go to a lower sampling rate, the filtering and interpolation can be combined if we use a nonrecursive filter. That is, the discrete Fourier transform of the filter impulse response can be simply combined with $H_N(k)$.

V. LIMITATIONS

Several times we have pointed out shortcomings of the chirp z-transform algorithm. One limitation in using it to evaluate the z-transform off the unit circle stems from the fact that we may be required to compute $W_o^{\pm n^2/2}$ for large n. If W_o differs very much from 1.0, $W_o^{\pm n^2/2}$ can become very large or very small when n becomes large. (We require a large n when either M or N become large, since we need to evaluate $W^{n^2/2}$ for n in the range $-N < n < M$.) For example, if $W_o = e^{-2\cdot 5/10,000} \approx 0.999749$, and $n = 1,000$, $W_o^{\pm n^2/2} = e^{\pm 125}$ which exceeds the single precision floating point capability of most computers by a large amount. Hence the tails of the functions $W^{\pm n^2/2}$ can be greatly in error, thus causing the tails of the convolution (the high frequency terms) to be grossly inaccurate. The low frequency terms of the convolution will also be slightly in error, but these errors generally are negligible.

An example of this effect is shown in Fig. 27. The contour for the five curves in this figure was held fixed (contour 5 in Fig. 6) and the number of frequency points in the range 0 to 5,000 Hz was increased in steps of 2 from 50 to 800. Spectral samples are plotted every 100 Hz for comparison. (This example was programmed using single-precision floating-point arithmetic on a GE 635 computer with a 36 bit word length.) It is seen that as the number of output points increases, errors in the high frequency region become large and completely mask the fifth resonance for the 800 point case. The effects of the inaccuracy in $W^{\pm n^2/2}$ can also be seen at low frequencies. For example, the spectral magnitude at 0 Hz goes from about 120 dB to 134 dB as the number of points goes from 50 to 800. These small errors generally do not affect the gross spectral characteristics as seen in Fig. 27. The resonances are easy to locate in all cases until the errors get exceedingly large. One can push the maximum point limit higher than 800 (in this case) by using double precision arithmetic.

The limitation of contour distance in or out from the unit circle is again the result of computation of $W^{\pm n^2/2}$. As W_o deviates significantly from 1.0, the number of points for which $W^{\pm n^2/2}$ can be accurately computed decreases. It is of importance to stress, however, that for

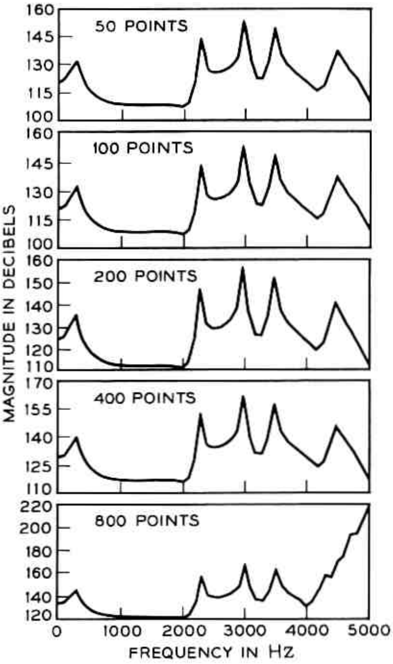


Fig. 27 — A comparison of magnitude plots for varying number of points on the same spiral contour. The fifth plot shows the effect of errors in evaluating $W^{n^2/2}$ for large n. (Points are plotted every 100 Hz in each curve to aid in comparison.)

$W_o = 1$ there is no limitation of this type since $W^{\pm n^2/2}$ is always of magnitude 1.

The other main limitation of the chirp z-transform algorithm stems from the fact that two L point fast Fourier transforms and one $L/2$ point fast Fourier transform must be evaluated where L is the smallest convenient integer greater than $N + M - 1$ as previously mentioned. We need one fast Fourier transform and $2L$ storage locations for the transform of $x_n A^{-n} W^{n^2/2}$; one fast Fourier transform and $L+2$ storage locations for the transform of $W^{-n^2/2}$; and one fast Fourier transform for the inverse transform of the product of these two transforms. We do not know a way of computing the transform of $W^{-n^2/2}$ either recursively or by a specific formula (except in some trivial cases.) Thus we must compute this transform and store it in an extra $L + 2$ storage locations. Of course, if many transforms are to be done with the same value of L we need not compute the transform of $W^{-n^2/2}$ each time.

We can compute the quantities $A^{-n} W^{n^2/2}$ recursively, as they are needed, to save computation and storage. This is easily seen from the fact that

$$A^{-(n+1)}W^{(n+1)^2/2} = (A^{-n}W^{n^2/2})\cdot W^nW^{\frac{1}{2}}A^{-1}. \tag{45}$$

If we define

$$C_n = A^{-n}W^{n^2/2} \tag{46}$$

and

$$D_n = W^nW^{\frac{1}{2}}A^{-1} \tag{47}$$

then

$$D_{n+1} = W\cdot D_n \tag{48}$$

and

$$C_{n+1} = C_n\cdot D_n\ . \tag{49}$$

Setting $A = 1$ in equations (45) through (49) provides an algorithm for the coefficients required for the output sequence. A similar recursion formula can be obtained for generating the sequence $A^{-n}W^{(n-N_o)^2/2}$. The user is cautioned that recursive computation of these coefficients may be a major source of numerical error, especially when $W_o \approx 1$ or $\varphi_o \approx 0$.

VI. SUMMARY

We give a computational algorithm for numerically evaluating the z-transform of a sequence of N time samples. This algorithm, the chirp z-transform algorithm, enables the evaluation of the z-transform at M equiangularly spaced points on contours which spiral in or out (circles being a special case) from an arbitrary starting point in the z-plane. In the s-plane the equivalent contour is an arbitrary straight line.

The chirp z-transform algorithm has great flexibility in that neither N or M need be composite numbers; the output point spacing is arbitrary; the contour is fairly general and N need not be the same as M. The flexibility of the chirp z-transform algorithm comes from being able to express the z-transform on the above contours as a convolution, permitting the use of well-known high speed convolution techniques to evaluate the convolution.

Applications of the chirp z-transform algorithm include enhancement of poles for use in spectral analysis, high resolution narrowband frequency analysis, and time interpolation of data from one sampling rate to any other sampling rate. These applications are explained in detail. The chirp z-transform algorithm also permits use of a radix 2 fast Fourier transform program or device to compute the discrete Fourier transform of an arbitrary number of samples. Examples were presented illustrating how the chirp z-transform algorithm was used in specific cases. It is anticipated that other applications will be found.

VII. ACKNOWLEDGMENT

The authors would like to acknowledge the help and assistance of Miss Carol McGonegal in programming many of the examples shown in this paper and Miss Debbie Hougak for valuable clerical help in preparing the manuscript.

APPENDIX

Fast Fourier Transforms for Two Real L *Point Sequences*

The purpose of this appendix is to show how the fast Fourier transforms of two real, symmetric L point sequences can be obtained using one $L/2$ point fast Fourier transform.

Let x_n and y_n be two real, symmetric L point sequences with corresponding discrete Fourier transforms X_k and Y_k. By definition,

$$\begin{aligned} x_n &= x_{L-n} \qquad n = 0, 1, 2, \cdots, L-1; \\ y_n &= y_{L-n} \end{aligned}$$

it is easily shown that X_k and Y_k are real, symmetric L point sequences, so that

$$\begin{aligned} X_k &= X_{L-k} \qquad k = 0, 1, 2, \cdots, L-1. \\ Y_k &= Y_{L-k} \end{aligned}$$

Define a complex, $L/2$ point sequence u_n whose real and imaginary parts are

$$\left.\begin{aligned} \text{Re}\,[u_n] &= x_{2n} - y_{2n+1} + y_{2n-1} \\ \text{Im}\,[u_n] &= y_{2n} + x_{2n+1} - x_{2n-1} \end{aligned}\right\} \qquad n = 0, 1, \cdots, L/2 - 1.$$

The $L/2$ point discrete Fourier transform of u_n is denoted U_k and is calculated by the fast Fourier transform. The values of X_k and Y_k may be computed from U_k using the relations

$$X_k = \tfrac{1}{2}\{\text{Re}\,[U_k] + \text{Re}\,[U_{L/2-k}]\} - \frac{1}{4\sin\frac{2\pi}{L}k}\{\text{Re}\,[U_k] - \text{Re}\,[U_{L/2-k}]\}$$

$$Y_k = \tfrac{1}{2}\{\text{Im}\,[U_k] + \text{Im}\,[U_{L/2-k}]\} - \frac{1}{4\sin\frac{2\pi}{L}k}\{\text{Im}\,[U_k] - \text{Im}\,[U_{L/2-k}]\}$$

$$\text{for} \quad k = 1, 2, \cdots, L/2 - 1.$$

The remaining values of X_k and Y_k are obtained from the relations

$$X_o = \sum_{n=0}^{L-1} x_n$$

$$Y_o = \sum_{n=0}^{L-1} y_n$$

$$X_{L/2} = \sum_{n=0}^{L-1} (-1)^n x_n$$

$$Y_{L/2} = \sum_{n=1}^{L-1} (-1)^n y_n \,.$$

REFERENCES

1. Cooley, J. W. and Tukey, J. W., "An Algorithm for the Machine Calculation of Complex Fourier Series," Mathematics of Computation, *19,* No. 90 (April 1965), pp. 297–301.
2. "What is the Fast Fourier Transform?," G-AE Subcommittee on Measurement Concepts, IEEE Tran. Audio and Electroacoustics, *AU-15,* No. 2, (June 1967), pp. 45–55.
3. Stockham, T. G., "High Speed Convolution and Correlation," 1966 Spring Joint Computer Conference, Amer. Federation of Inform. Processing Soc. Proc., *28,* Washington, D. C., April 1966, pp. 229–233.
4. Helms, H. D. "Fast Fourier Transform Method of Computing Difference Equations and Simulating Filters," IEEE Trans. Audio and Electroacoustics, *AU-15,* No. 2 (June 1967), pp. 85–90.
5. Bluestein, L. I., "A Linear Filtering Approach to the Computation of the Discrete Fourier Transform," 1968 Northeast Electronics Research and Engineering Meeting Record, *10,* November 1968, pp. 218–219.
6. Rader, C. M. and Gold, B., "Digital Filter Design Techniques in the Frequency Domain," Proc. IEEE, *55,* No. 2, (February 1967), pp. 149–171.

7. Schafer, R. W. and Rabiner, L. R., Automatic Formant Analysis of Speech Using the Chirp z-Transform Algorithm" to be presented at the 1969 IEEE International Conference on Communications, Boulder, Colorado, June, 1969.
8. Gold, B. and Jordan, K. L., "A Direct Search Procedure for Designing Finite Duration Impulse Response Filters," IEEE Trans. on Audio Electroacoustics, *AU-17,* No. 1 (March 1969) pp. 33–36.
9. Gentleman, W. M. and Sande, G., "Fast Fourier Transforms for Fun and Profit," 1966 Fall Joint Computer Conference, American Federation of Information Processing Societies Proc., *29,* Washington, D. C., November 1966, pp. 563–578.

Some Network-Theoretic Properties of Nonlinear DC Transistor Networks

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

(Manuscript received September 9, 1968)

This paper extends, in several directions, some of the results of earlier work concerned with the existence and uniqueness of solutions of the dc equations of nonlinear transistor networks. In particular, here we develop techniques which enable us to deal directly with a more complicated transistor model.

I. INTRODUCTION

Several results are presented in Ref. 1 concerning the equation

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

(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, a necessary and sufficient condition on A is given 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. Several circuit-theoretic implications of the results are also described in Ref. 1; for example, it is shown that the short-circuit admittance matrix of the linear portion of the dc model of an interesting class of switching circuits must violate a certain dominance condition.

In Ref. 1 the word *transistor* was used to refer to the three-terminal device whose dc equivalent circuit is shown in Fig. 1(a). Although this equivalent circuit is frequently used in the design and computer analysis of transistor networks it is, from a physical standpoint, somewhat incomplete. A more exact dc model of a physical transistor is that of Fig. 1(b) in which the presence of series resistance in each of the transistor's leads has been accounted for.

In this paper we report on several extensions of the previous results. The motivation for much of this work was to enable the model of Fig. 1(b) to be taken into account. In addition, we present here

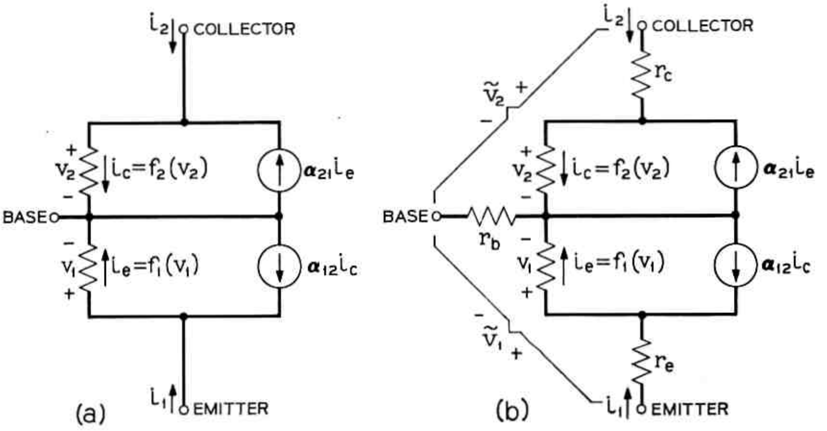


Fig. 1 — DC transistor models.

further material concerning cases in which (in accordance with standard assumptions) the nonlinear functions of Fig. 1(b) do not map E^1 *onto* itself. Finally, we prove a considerably stronger result than that of Ref. 1, to the effect that a certain class of networks cannot be bistable.

We now summarize some of the material of Ref. 1 that will be needed in the sequel:

For each positive integer n, we 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 $x \equiv (x_1, \cdots, x_n)^t \in E^n$, $F(x) \equiv (f_1(x_1), \cdots, f_n(x_n))^t$.

The origin in E^n will be denoted by θ. Throughout this article we consider only matrices whose elements are real. If D is a diagonal matrix then $D > 0$ ($D \geqq 0$) means that each element on the main diagonal of D is positive (nonnegative).

The classes of matrices P and P_0 have been defined by M. Fiedler and V. Pták in Refs. 2 and 3. They prove that these classes can be defined by any one of several equivalent properties. We shall need only the following characterization of the classes P and P_0 : A square matrix A is a member of the class P (P_0) if and only if all principal minors of A are positive (nonnegative). In the appendix it is proved that $A \in P_0$ if and only if det $[A + D] \neq 0$ for every diagonal matrix $D > 0$.

† If M is an arbitrary matrix, then the transpose of M is denoted in this article by M^t.

The following theorem is proved in Ref. 1:

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

We say that an $n \times n$ matrix A is *strongly* (*weakly*) *row-sum dominant* if and only if the elements a_{ij} of A satisfy

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

Similarly, a *strongly* (*weakly*) *column-sum dominant* matrix is one that satisfies

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

The square matrix A is said to be *dominant* (*strongly dominant*) if and only if A is weakly (strongly) row-sum dominant and symmetric.

If a square matrix A is strongly column-sum or row-sum dominant then A is nonsingular, in fact $A \varepsilon P$.

The following theorem is also proved in Ref. 1:

Theorem 2: 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*).

An analogous theorem involving row-sum dominant matrices is also true, and can be proved with trivial modifications of the proof of Theorem 2 given in Ref. 1.

II. FURTHER RESULTS CONCERNING THE EXISTENCE AND UNIQUENESS OF SOLUTIONS

The proof of Theorem 1 given in Ref. 1 exploits the fact that the straight line described by the equation $y = -ax + b$ has exactly one intersection with the graph of each strictly monotone increasing function $f(x)$ which maps E^1 onto E^1 if and only if $a \geqq 0$.

It happens that a useful result that is slightly more general than that of Theorem 1 can be proved easily if use is made of a proposition that is similar to, but stronger than, the elementary fact mentioned in the preceding paragraph. That proposition is stated below.

Definition: For all α, β with $-\infty \leqq \alpha < \beta \leqq \infty$, let $I(\alpha, \beta)$ denote the interval $I(\alpha, \beta) = \{x : \alpha < x < \beta\}$.

The following proposition is quite easily verified:

Proposition: For $-\infty \leqq \alpha < \beta \leqq \infty$, the straight line described by the equation $y = -ax + b$ has exactly one intersection with the graph of each strictly monotone increasing function $f(x)$ which maps $I(\alpha, \beta)$ onto E^1 if and only if $a \geqq 0$.

Definition: For each positive integer n and each pair of n-vectors α, β whose components α_i, β_i lie in the extended real number system, with $\alpha < \beta$ (that is, with $-\infty \leqq \alpha_i < \beta_i \leqq \infty$ for $i = 1, \cdots, n$) let $\mathfrak{F}^n(\alpha, \beta; E^n)$ denote that collection of mappings of $I(\alpha_1, \beta_1) \times \cdots \times I(\alpha_n, \beta_n)$ onto E^n defined by: $F \in \mathfrak{F}^n(\alpha, \beta; E^n)$ if and only if there exist, for $i = 1, \cdots, n$, strictly monotone increasing functions f_i mapping (α_i, β_i) onto E^1 such that for each $x \equiv (x_1, \cdots, x_n)^t \in I(\alpha_1, \beta_1) \times \cdots \times I(\alpha_n, \beta_n)$,

$$F(x) \equiv (f_1(x_1), \cdots, f_n(x_n))^t.$$

Let the collection of strictly monotone increasing mappings of E^n onto $I(\alpha_1,\beta_1) \times \cdots \times I(\alpha_n,\beta_n)$ be similarly defined, and denoted by $\mathfrak{F}^n(E^n; \alpha, \beta)$. Note that $F \in \mathfrak{F}^n(\alpha, \beta; E^n)$ if and only if F^{-1} exists and $F^{-1} \in \mathfrak{F}^n(E^n; \alpha, \beta)$. Also, in case $I(\alpha_1,\beta_1) \times \cdots \times I(\alpha_n,\beta_n) = E^n$, then $\mathfrak{F}^n(\alpha, \beta; E^n) = \mathfrak{F}^n(E^n; \alpha, \beta) = \mathfrak{F}^n$.

Using the above proposition it is now easy to prove:

Theorem 3: For the n-vectors $\alpha < \beta$ whose components lie in the extended real number system, if A is an $n \times n$ matrix then there exists a unique solution of (1) *for each $F \in \mathfrak{F}^n(\alpha, \beta; E^n)$ and each $B \in E^n$ if and only if $A \in P_0$.*

Proof: (*if*) The proof of this part of the theorem is identical to the proof (given in Ref. 1) of the corresponding part of Theorem 1 with the exception that appropriate use is made of the above proposition. Since the necessary modifications are quite obvious we omit the details.

(*only if*) Suppose $A \notin P_0$. Then there exists a diagonal matrix $D \equiv \operatorname{diag}[d_1, \cdots, d_n] > 0$ such that $\det[A + D] = 0$. Let x^0 be an arbitrary point in $I(\alpha_1, \beta_1) \times \cdots \times I(\alpha_n, \beta_n)$ and let y^0 be an arbitrary point in E^n. Let

$$B = y^0 + Ax^0.$$

Let $\delta > 0$ be chosen such that

$$\alpha_i < x_i^0 - \delta < x_i^0 + \delta < \beta_i, \qquad \text{for } i = 1, \cdots, n,$$

and choose $F \equiv (f_1(\cdot), \cdots, f_n(\cdot))^t$ in $\mathfrak{F}^n(\alpha, \beta; E^n)$ such that for $i = 1, \cdots, n$, and for $x_i^0 - \delta < x_i < x_i^0 + \delta$,

$$f_i(x_i) = y_i^0 + d_i(x_i - x_i^0).$$

Thus, $F(x^0) = y^0$ and hence, x^0 is a solution of (1) for this choice of F.

Since $\det [A + D] = 0$, there exists some n-vector $x^* \neq \theta$ having the property that

$$Ax^* + Dx^* = \theta.$$

Thus, for each real number ϵ,

$$y^0 + D\epsilon x^* + A(x^0 + \epsilon x^*) = B.$$

In particular, if $\epsilon \neq 0$ is chosen such that $|\epsilon|$ is sufficiently small, then $| \epsilon x_i^* | < \delta$ for $i = 1, \cdots, n$. Hence, for such ϵ, if $x = x^0 + \epsilon x^*$, $F(x) = y^0 + D\epsilon x^*$ and therefore $x \neq x^0$ is also a solution of (1). □

An important special case of Corollary 3 of Ref. 1 is:

Corollary 1: For the n-vectors $\alpha < \beta$ whose components lie in the extended real number system, if A is an $n \times n$ matrix then there exists a unique solution of (1) *for each $F \in \mathfrak{F}^n(E^n; \alpha, \beta)$ and each $B \in E^n$ if $A \in P$.*

Theorem 3 may be used to prove a sharper (and, from the viewpoint of transistor networks, a more useful) result than Corollary 1. We have:

Theorem 4: For the n-vectors $\alpha < \beta$ whose components lie in the extended real number system (in the real number system), if A is an $n \times n$ matrix then there exists a unique solution of (1) *for each $F \in \mathfrak{F}^n(E^n; \alpha, \beta)$ and each $B \in E^n$ if (and only if) $A \in P_0$ and det $A \neq 0$.*

Proof: (*if*) As pointed out in Ref. 1, $A \in P_0$ and det $A \neq 0$ imply that $A^{-1} \in P_0$. Also, F^{-1} exists and $F^{-1} \in \mathfrak{F}^n(\alpha, \beta; E^n)$. Now x satisfies (1) if and only if y satisfies

$$F^{-1}(y) + A^{-1}y = A^{-1}B, \tag{2}$$

where $y = F(x)$. But, according to Theorem 3, there exists a unique y which satisfies (2).

(*only if*) We assume here that the components of α and β are real. Suppose $A \notin P_0$. Then, in a manner similar to that used in the proof of the "only if" part of Theorem 3, we can choose a mapping $F \in \mathfrak{F}^n(E^n; \alpha, \beta)$ and a point $B \in E^n$, such that the solution of (1) is not unique.

If, on the other hand, det $A = 0$, then there exists $x^* \neq \theta$ such that $A^t x^* = \theta$. Assume that (1) has a solution x for each $B \in E^n$. Then, since $\langle x^*, Ax \rangle = 0$ for all x, we have

$$\langle x^*, F(x) \rangle = \langle x^*, B \rangle,$$

for each $B \in E^n$ (and the corresponding x). It is clear, since the components of α and β are finite, that there exists some constant M such that

$$|\langle x^*, F(x)\rangle| \leqq M$$

for all $x \in E^n$. But B can certainly be chosen such that $\langle x^*, B\rangle > M$. This contradiction completes the proof of the theorem. □

The following theorem provides an alternative method of characterizing the class of matrices that are in P_0 and are nonsingular (compare with the theorem of the appendix).

Theorem 5: If A is a real square matrix then $A \in P_0$ and det $A \neq 0$ if and only if det $[A + D] \neq 0$ for every diagonal matrix $D \geqq 0$.

Proof: (*if*) It is clear, by the theorem of the appendix, that $A \in P_0$, since det $[A + D] \neq 0$ for all diagonal $D > 0$. Moreover, det $A \neq 0$, by hypothesis.

(*only if*) It is shown in Ref. 1 that, for each $A \in P_0$ and each diagonal $D \geqq 0$, $A + D \in P_0$. It suffices, therefore, to show that if $D_i = \text{diag}\,[0, \cdots, 0, d_i, 0, \cdots, 0]$ with $d_i \geqq 0$, and $A \in P_0$ with det $A > 0$, then det $[A + D_i] > 0$. Letting A_i denote the principal submatrix obtained from A by deleting the ith row and the ith column, we have

$$\det\,[A + D_i] = \det A + d_i \det A_i\,.$$

But det $A > 0$ and d_i det $A_i \geqq 0$. □

III. APPLICATION TO EQUATIONS FOR TRANSISTOR NETWORKS

In the analysis of a transistor network one could account for the presence of series lead resistance, while using the model of Fig. 1(a) to represent the transistor, by including appropriate additional resistors in the rest of the network. Indeed, there is at least one good reason for doing this. When treated in this manner, the presence of nonzero series resistance in the base, collector, and emitter leads of each transistor ensures that the y-parameter matrix exists for the circuit to which the transistors are connected—and hence ensures that the transistor network can be described by an equation having the form of (1). On the other hand, there are also good reasons for representing the transistor, for analysis purposes, by the model of Fig. 1(b). Using this model it will be shown, for example, that it is often possible to determine that there is a unique solution of the equation describing a given transistor network *regardless* of the (nonnegative) values of the transistors' series lead resistances. Since these resistances are usually parasitic and unavoidable in nature it is significant that one might be able to show that their introduction in, say, a certain monostable circuit will not cause the circuit to become bistable.

Using the model of Fig. 1(b) it is quite easy to see that the port variables for the transistor, when considered as a nonlinear two-port network, obey the following relationship

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

where

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} \tilde{v}_1 \\ \tilde{v}_2 \end{pmatrix} - \begin{bmatrix} r_e + r_b & r_b \\ r_b & r_e + r_b \end{bmatrix} \begin{pmatrix} i_1 \\ i_2 \end{pmatrix}.$$

As in Ref. 1 we assume that $0 < \alpha_{12} < 1$, $0 < \alpha_{21} < 1$, and that both of the functions f_1 and f_2 are strictly monotone increasing mappings of E^1 into E^1.

Suppose an electrical network is synthesized containing transistors, resistors (that is, linear resistors having nonnegative resistance), independent voltage and current sources, and nonlinear resistors which are described by strictly monotone increasing conductance functions (and which shall henceforth be called "diodes"). Suppose the network contains n transistors and d diodes ($n + d > 0$). For $k = 1, \cdots, n$ let x_{2k-1}, x_{2k}, $\tilde{x}_{2k-1}$, $\tilde{x}_{2k}$, y_{2k-1}, and y_{2k} denote the voltage and current variables v_1, v_2, $\tilde{v}_1$, $\tilde{v}_2$, i_1, and i_2, respectively, for the kth 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; also (for $k = 1, \cdots, d$) let $\tilde{x}_{2n+k} = x_{2n+k}$. Let these variables be related by $y_{2n+k} = f_{2n+k}(x_{2n+k})$. Then, if $x = (x_1, \cdots, x_{2n+d})^t$, $\tilde{x} = (\tilde{x}_1, \cdots, \tilde{x}_{2n+d})^t$, and $y = (y_1, \cdots, y_{2n+d})^t$, we have

$$y = TF(x), \qquad x = \tilde{x} - Ry, \tag{3}$$

where $T = \text{diag}[T_1, T_2]$, with T_1 a block diagonal matrix with n 2×2 diagonal blocks of the form

$$\begin{bmatrix} 1 & -\alpha_{12}^{(k)} \\ -\alpha_{21}^{(k)} & 1 \end{bmatrix}, \tag{4}$$

and T_2 the $d \times d$ identity matrix. Also, $R = \text{diag}[R_1, R_2]$, with R_1 a block diagonal matrix with n 2×2 diagonal blocks of the form

$$\begin{bmatrix} r_e^{(k)} + r_b^{(k)} & r_b^{(k)} \\ r_b^{(k)} & r_c^{(k)} + r_b^{(k)} \end{bmatrix}, \tag{5}$$

and R_2 the $d \times d$ null matrix.

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 $\tilde{x}$ and y:

$$y = -G\tilde{x} + \tilde{u} \tag{6}$$

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

The vectors $\tilde{x}$ and y can easily be eliminated from (3) and (6), resulting in the equation

$$TF(x) + [I + GR]^{-1}Gx = u, \tag{7}$$

where we have defined the vector u by

$$u = [I + GR]^{-1}\tilde{u}.$$

According to Theorem 6, below, the matrix $[I + GR]$ must be nonsingular.

In case the matrix R contains all zeros (that is, in case all series lead resistors are omitted from the transistors) (7) reduces immediately to the equation which was studied in Ref. 1. Even when R does not contain all zeros, however, the results of Ref. 1 can be applied directly to (7). By applying Theorem 2 we have: *If the matrix* $[I + GR]^{-1}$ G *is dominant*† *then there is at most one solution of* (7). *If, furthermore,* F *maps* E^n *onto* E^n, *or if* $[I + GR]^{-1}G$ *is strongly dominant, then there exists a unique solution of* (7).

Making use of Theorem 4, we also have the stronger result: *There exists a unique solution of* (7) *if* $[I + GR]^{-1}G$ *is dominant and* G *is nonsingular.*

Although it is not, in general, true that the inverse of a strongly column-sum (row-sum) dominant matrix is strongly row-sum (column-sum) dominant, the statement is true when the order of the matrix is less than three. This elementary observation turns out to be quite useful in the proof of Theorem 6, which yields results that focus attention on the properties of G, concerning the existence and uniqueness of a solution of (7).

† For symmetric matrices the properties (*i*) weak column-sum dominance, and (*ii*) dominance, are identical. Since it is easily verified that for symmetric G and R, $[I + GR]^{-1}G$ is also symmetric, we simply specifiy that $[I + GR]^{-1}G$ be dominant.

Theorem 6: Let A (B) be the direct sum of n 2×2 and d 1×1 strongly column-sum (weakly row-sum) dominant matrices. Let B be symmetric and let C be a square matrix of order $2n + d$. Then:

(*i*) $\det [I + CB] \neq 0$, *provided that C is positive semidefinite,*
(*ii*) $A^{-1}[I + CB]^{-1}C \;\varepsilon\; P_0$, *provided that C is dominant,*
(*iii*) $A^{-1}[I + CB]^{-1}C \;\varepsilon\; P$, *provided that C is strongly dominant.*

Proof: (*i*) Here C is positive semidefinite. Let $B^{\frac{1}{2}}$ be the symmetric nonnegative square root of B, so that $I + CB = I + CB^{\frac{1}{2}}B^{\frac{1}{2}}$. Since (see Appendix A of Ref. 4) $\det [I + CB^{\frac{1}{2}}B^{\frac{1}{2}}] = \det [I + B^{\frac{1}{2}}CB^{\frac{1}{2}}]$, and since $I + B^{\frac{1}{2}}CB^{\frac{1}{2}}$ is positive definite, we have $\det [I + CB] > 0$.

(*ii*) Here C is dominant (which, as is well known, implies that C is positive semidefinite and hence, by (*i*), implies that $[I + CB]^{-1}$ exists). Suppose $A^{-1}[I + CB]^{-1}C \not\varepsilon P_0$. Then, by the theorem of the appendix, there exists a diagonal matrix $D > 0$ such that $A^{-1}[I + CB]^{-1}C + D$ is singular. But

$$A^{-1}[I + CB]^{-1}C + D = A^{-1}[I + CB]^{-1}[C(D^{-1}A^{-1} + B) + I]AD,$$

which means that $C(D^{-1}A^{-1} + B) + I$ must be singular. Since A is a direct sum of 1×1 and 2×2 strongly column-sum dominant matrices, it follows that A^{-1} is a direct sum of 1×1 and 2×2 strongly row-sum dominant matrices. Thus, $D^{-1}A^{-1}$ and hence $D^{-1}A^{-1} + B$ is strongly row-sum dominant. Therefore, $(D^{-1}A^{-1} + B)$ is nonsingular, and $(D^{-1}A^{-1} + B)^{-1}$ is strongly column-sum dominant. But,

$$C(D^{-1}A^{-1} + B) + I = [C + (D^{-1}A^{-1} + B)^{-1}](D^{-1}A^{-1} + B)$$

in which the right-hand side is nonsingular since $C + (D^{-1}A^{-1} + B)^{-1}$ is strongly column-sum dominant, which is a contradiction.

(*iii*) Here C is strongly dominant. Since $C(I + BC) = (I + CB)C$, we have $\det(I + BC) > 0$ and

$$(I + CB)^{-1}C = C(I + BC)^{-1}.$$

Suppose that there is no constant $\delta > 0$ such that $A^{-1}\, C(I + BC)^{-1} - \delta I \;\varepsilon\; P_0$. Then, for each $\delta > 0$ there is a diagonal matrix $D > 0$ such that $A^{-1}\, C(I + BC)^{-1} - \delta I + D$ is singular. But,

$$\begin{aligned} A^{-1}C(I + BC)^{-1} - \delta I + D & \\ &= A^{-1}[C - \delta A(I + BC) + AD(I + BC)](I + BC)^{-1} \\ &= D\{I + BC + D^{-1}A^{-1}[C - \delta A(I + BC)]\}(I + BC)^{-1} \\ &= \{D + [DB + A^{-1} - \delta(C^{-1} + B)]C\}(I + BC)^{-1}, \end{aligned}$$

which leads to the conclusion that for each $\delta > 0$ there is a $D > 0$ such that $D + [DB + A^{-1} - \delta(C^{-1} + B)]C$ is singular. We now establish a contradiction:

For all $x \varepsilon E^n$, let $\| x \| = \max_i | x_i |$. If $x, y \varepsilon E^n$ such that $\| x \| = 1$ and

$$[DB + A^{-1}]y = x$$

then it is easy to show that

$$\| y \| \leqq \max_k \frac{1}{\alpha_{kk} - \sum_{j \neq k} | \alpha_{kj} |}$$

in which the α_{kj} are the elements of A^{-1}. Thus, the norm of $[DB + A^{-1}]^{-1}$ can be bounded from above uniformly in $D > 0$. Therefore,

$$D + [DB + A^{-1} - \delta(C^{-1} + B)]C = (DB + A^{-1})\{(DB + A^{-1})^{-1}D + [I - \delta(DB + A^{-1})^{-1}(C^{-1} + B)]C\}$$

in which $\delta > 0$ can be chosen so small that $[I - \delta(DB + A^{-1})^{-1}(C^{-1} + B)]C$ is strongly column-sum dominant for all $D > 0$. Since $(DB + A^{-1})^{-1}D$ is also column-sum dominant, we have a contradiction. It follows that for some $\delta > 0$, $A^{-1}(I + CB)^{-1}C - \delta I \varepsilon P_0$ and hence, by Theorem 1 of Ref. 1, $A^{-1}(I + CB)^{-1}C \varepsilon P$. □

The matrices T, R, and G of (7) satisfy the hypotheses on A, B, and C, respectively, of Theorem 6 if it happens that G is dominant (strongly dominant for (*iii*)). Thus, we have the result: *If the y-parameter matrix G is dominant then there is at most one solution of* (7). *If, furthermore, F maps E^n onto E^n, or if G is strongly dominant, then there exists a unique solution of* (7).

Making use of Theorem 4 and since det $C \neq 0$ implies det $[A^{-1}(I + CB)^{-1}C] \neq 0$, we also have: *There exists a unique solution of* (7) *if G is dominant and nonsingular.*

These results show that if the solution of the equation

$$TF(x) + Gx = \tilde{u}, \tag{8}$$

describing a given transistor network (with the transistors represented by the model of Fig. 1(a) is shown to (exist and) be unique by showing that the y-parameter matrix G is dominant (and det $G \neq 0$, or that F maps E^n onto E^n), then any other network obtained from the original by adding arbitrary (nonnegative) resistances in series with any of the transistor leads will be described by (7) and, furthermore, the solution of (7) will also (exist and) be unique. Thus, the addition of series lead resistance does not affect the existence and uniqueness of the solution, provided G is dominant.

We now prove another result concerning the relationship between the existence and uniqueness of solutions of the two equations (7) and (8). We prove that, roughly speaking, whenever (8) has a unique solution for all transistors and diodes then so does (7). More precisely, let us define, for a given transistor network, the class of matrices $\mathfrak{I}$:

Definition: Let (8) describe the given network for some choice of transistor parameters α_{12} , α_{21} , for each transistor. Let $\mathfrak{I}$ then denote that class of matrices T obtained by considering all possible combinations of values of α_{12} , α_{21} $(0 < \alpha_{12} < 1, 0 < \alpha_{21} < 1)$ for each transistor.

We then have:

Theorem 7: If (8) *has a unique solution for each* $T \varepsilon \mathfrak{I}$, *and each* $F \varepsilon \mathfrak{F}^n(E^n; \alpha, \beta)$ *for all* $\alpha < \beta$ *whose components lie in the extended real number system then, for each* R, *so does* (7).

Proof: The hypotheses imply (using Theorem 4) that $T^{-1}G \varepsilon P_0$ and $\det [T^{-1}G] \neq 0$ for each $T \varepsilon \mathfrak{I}$. Thus, G^{-1} exists. Letting

$$H \equiv [I + GR]^{-1}G,$$

H^{-1} exists and,

$$H^{-1} = G^{-1} + R.$$

As pointed out in Ref. 1, since $\det [T^{-1}G] \neq 0$, $T^{-1}G \varepsilon P_0$ for every $T \varepsilon \mathfrak{I}$ implies that $G^{-1}T \varepsilon P_0$ for every $T \varepsilon \mathfrak{I}$. Hence

$$\det [G^{-1}T + D] > 0, \qquad \text{for all } T \varepsilon \mathfrak{I} \text{ and all } D > 0.$$

But then,

$$\det [G^{-1} + DT^{-1}] > 0, \qquad \text{for all } T \varepsilon \mathfrak{I} \text{ and all } D > 0.$$

Now, due to the special structure of the matrix R (that is, block diagonal with dominant blocks that are "compatible" with T^{-1}) it is clear that, for any such R, any diagonal $D > 0$, and any $T \varepsilon \mathfrak{I}$, there exists a diagonal $\Delta > 0$ and some $M \varepsilon \mathfrak{I}$, such that $R + DT^{-1} = \Delta M^{-1}$. Hence, it is clear that

$$\det [G^{-1} + R + DT^{-1}] > 0, \qquad \text{for all } T \varepsilon \mathfrak{I} \text{ and all } D > 0.$$

It easily follows that $H^{-1}T \varepsilon P_0$ and hence $T^{-1}H \varepsilon P_0$ for all $T \varepsilon \mathfrak{I}$. Applying Theorem 4, we thus have that there exists a unique solution of (7) for each $T \varepsilon \mathfrak{J}$, and each $F \varepsilon \mathfrak{F}^n(E^n; \alpha, \beta)$ for all $\alpha < \beta$ whose components lie in the extended real number system. □

It is not difficult to show that there exist transistor networks for which $[I + GR]^{-1}G$ is dominant while G is not, and also networks for which G is dominant while $[I + GR]^{-1}G$ is not. For the first case, pick any network for which G is not dominant and det $G \neq 0$. If the values of the series lead resistors in each transistor lead are then allowed to become large, since

$$[I + GR]^{-1}G = [I + R^{-1}G^{-1}]^{-1}R^{-1},$$

and since each element of R^{-1} approaches zero as the lead resistor values approach infinity, we see that $[I + GR]^{-1}G \to R^{-1}$. But R^{-1} is strongly dominant and hence there certainly exist sufficiently large values for the lead resistors such that $[I + GR]^{-1}G$ is dominant. The network of Fig. 2 is an example of the other case. For this network,

$$G = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}, \qquad R = \begin{bmatrix} 9 & 9 & 0 & 0 \\ 9 & 9 & 0 & 0 \\ 0 & 0 & 9 & 9 \\ 0 & 0 & 9 & 9 \end{bmatrix},$$

while

$$[I + GR]^{-1}G = \frac{1}{37}\begin{bmatrix} 19 & -18 & -19 & 18 \\ -18 & 19 & 18 & -19 \\ -19 & 18 & 19 & -18 \\ 18 & -19 & -18 & 19 \end{bmatrix}.$$

IV. A SPECIAL CLASS OF TRANSISTOR NETWORKS

Transistor networks in which the base terminal of each transistor is connected to a common node are considered in Ref. 1 using the model of Fig. 1(a) to represent the transistor. It is shown there that there is at most one pair of base-collector and base-emitter voltages for each transistor in such a network—even in the cases in which the network is not described by an equation having the form of (1).

In this section we show that the class of common-base transistor networks is but a subset of a considerably more extensive special class of transistor networks for which the same statement is true. We show that there is at most one pair of base-collector and base-emitter volt-

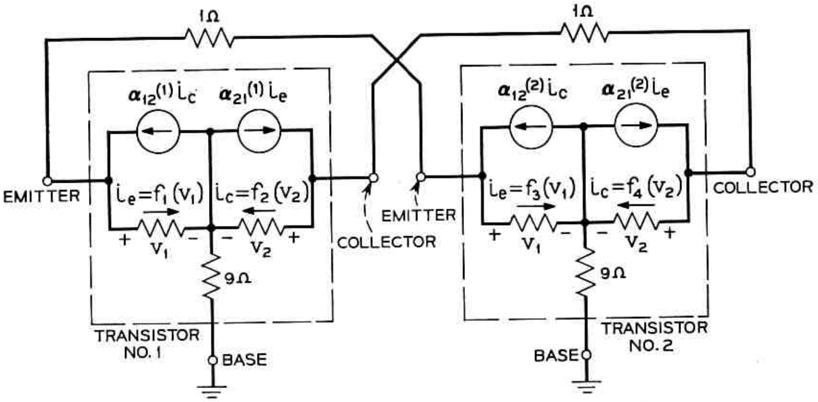


Fig. 2 — A two-transistor network.

ages for each transistor in any dc network which has the structure shown in Fig. 3. The box at the top of Fig. 3 represents, assuming that there are n transistors, any $(2n + 1)$-terminal network consisting of independent voltage and current sources, resistors (that is, linear resistors having nonnegative resistance), and diodes (that is, nonlinear resistors which are described by strictly monotone increasing conductance functions). Each of the n boxes at the bottom of Fig. 3 represents an

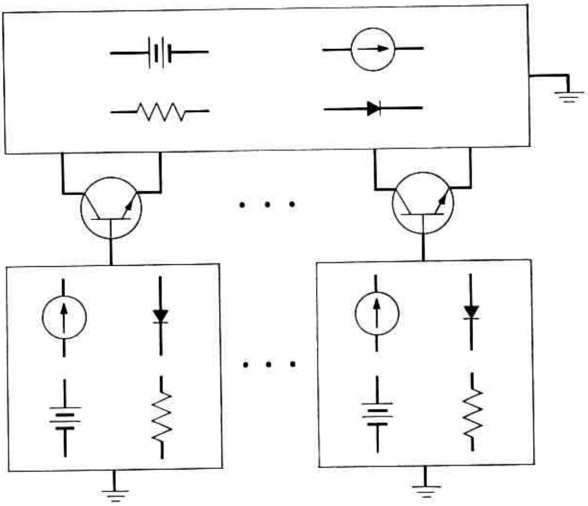


Fig. 3 — A special class of transistor networks.

arbitrary 2-terminal network consisting of independent sources, resistors, and diodes. Each of the transistors in Fig. 3 is represented by the model of Fig. 1(b), in which the value of each of the resistors r_b , r_c , r_e may be any nonnegative number. In this regard, we note here that it suffices in what follows to show, for each transistor, the uniqueness of the voltages v_1 and v_2 (in Fig. 1(b)) since, clearly, the voltages $\tilde{v}_1$ and $\tilde{v}_2$ are then uniquely determined.

As in Ref. 1 we assume, temporarily, that no diodes are present in the network. This assumption allows each of the n boxes at the bottom of Fig. 3 to be replaced by either a current source or else a Thévenin's equivalent circuit in which the value of the Thévenin's resistor is not infinite. Let us temporarily ignore the possibility that any of these boxes is equivalent to a current source. Following the technique presented in Section IX of Ref. 1, we may then consider the network of Fig. 4 instead of that of Fig. 3. In Fig. 4 we have explicitly shown the base, emitter, and collector resistors of each transistor, and we consider the Thévenin's resistor of each base circuit to be lumped in with the corresponding base resistor. The m-vectors v^* and i^* ($m \leqq 2n$) and the $2n$-vectors v' and i' are related by the four equations:

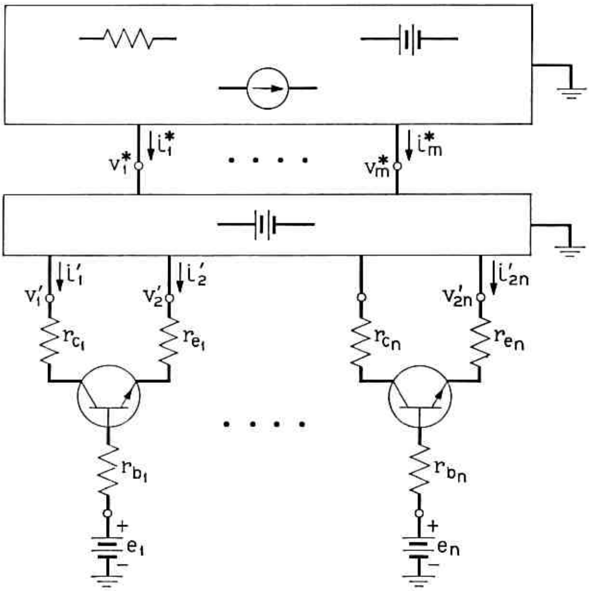


Fig. 4 — Network derived from that of Fig. 3.

$$i^* = -Gv^* + b, \tag{9}$$

$$i^* = Qi', \tag{10}$$

$$v' = Q^t v^* + c, \tag{11}$$

$$i' = TF(v' - e - Ri'), \tag{12}$$

in which b, c, and e are vectors whose elements are constants, G is a dominant matrix, Q is an $m \times 2n$ matrix having the property that whenever the $2n \times 2n$ matrix M is strongly column-sum dominant then so is the $m \times m$ matrix QMQ^t, T and R are $2n \times 2n$ block diagonal matrices having 2×2 diagonal blocks of the form (4) and (5), respectively.

We now show that the vectors v^*, i^*, v', and i' which satisfy (9) through (12) are unique (if they exist). Let $\{v^*_{(1)}, i^*_{(1)}, v'_{(1)}, i'_{(1)}\}$ and $\{v^*_{(2)}, i^*_{(2)}, v'_{(2)}, i'_{(2)}\}$ denote two sets of vectors, each of which satisfies (9) through (12). Subtracting corresponding equations, and observing the strictly monotone character of F, we see that there exists a diagonal matrix $D > 0$ such that:

$$i^*_{(1)} - i^*_{(2)} = -G(v^*_{(1)} - v^*_{(2)}), \tag{13}$$

$$i^*_{(1)} - i^*_{(2)} = Q(i'_{(1)} - i'_{(2)}), \tag{14}$$

$$v'_{(1)} - v'_{(2)} = Q^t(v^*_{(1)} - v^*_{(2)}), \tag{15}$$

$$i'_{(1)} - i'_{(2)} = TD(v'_{(1)} - v'_{(2)} - R(i'_{(1)} - i'_{(2)})). \tag{16}$$

But (15) and (16) imply

$$[I + TDR](i'_{(1)} - i'_{(2)}) = TDQ^t(v^*_{(1)} - v^*_{(2)}).$$

However, since

$$[I + TDR] = T[T^{-1} + DR],$$

in which T is strongly column-sum dominant (T^{-1} is strongly row-sum dominant), and DR is weakly row-sum dominant, we have $\det [I + TDR] \neq 0$, and hence,

$$i'_{(1)} - i'_{(2)} = [I + TDR]^{-1}TDQ^t(v^*_{(1)} - v^*_{(2)}). \tag{17}$$

Substituting this into (14) and then (13), however, yields:

$$\{Q[I + TDR]^{-1}TDQ^t + G\}(v^*_{(1)} - v^*_{(2)}) = \theta.$$

Now if $Q[I + TDR]^{-1}TDQ^t + G$ can be shown to be nonsingular then $v^*_{(1)} - v^*_{(2)} = \theta$ and hence, by (13), (15), and (17): $i^*_{(1)} - i^*_{(2)} = \theta$, $v'_{(1)} - v'_{(2)} = \theta$, and $i'_{(1)} - i'_{(2)} = \theta$, which, together, show that the vectors which satisfy (9) through (12) are unique. Since G is dominant it suffices to show that $[I+TDR]^{-1}TD$ (and hence $Q[I+TDR]^{-1}TDQ^t$) is strongly column-sum dominant. But

$$[I + TDR]^{-1}TD = [D^{-1}T^{-1} + R]^{-1},$$

which is the inverse of the direct sum of 2 × 2 strongly row-sum dominant matrices and is, therefore, strongly column-sum dominant.

Let us now consider the case in which diodes are present in the box at the top of Fig. 3. In this case, arguing as in Section IX of Ref. 1, if the set of base-emitter and base-collector voltages for Fig. 3 was not unique, we could replace all of the diodes by an appropriate series combination of a voltage source and a (nonnegative) resistor and thus synthesize a network of the type just considered, for which the set of base-emitter and base-collector voltages is not unique. This is a contradiction, and hence establishes that the set of base-emitter and base-collector voltages for the network of Fig. 3 is unique even when diodes are present in the top box.

A somewhat similar argument may now be used to show the uniqueness of the voltage across each of the diodes in the box at the top of Fig. 3. Assume that there exist two sets of branch voltages and currents, S_1 and S_2, which satisfy Kirchoff's and Ohm's laws for the network of Fig. 3. Since we have just proved the uniqueness of the base-emitter and base-collector voltages of each transistor, the elements of S_1 and S_2 which correspond to any such voltage must be identical. Thus, if each transistor is replaced by, say, an appropriate pair of voltage sources, the sets S_1 and S_2 still satisfy Kirchoff's and Ohm's laws for the modified network. Let us now choose (arbitrarily) any diode in the network and, as in the previous argument, replace all other diodes by a series combination of a voltage source and a (nonnegative) resistor, thus obtaining a new network, containing only one diode, for which the sets S_1 and S_2 still satisfy Kirchoff's and Ohm's laws. Suppose this remaining diode is characterized by the equation $i = f(v)$. The (now linear) network to which this diode is connected contains only independent sources and nonnegative resistors, and hence is characterized by one of the equations: $-i = gv + I_0$, $v = V_0$, where $g \geqq 0$, I_0, and V_0 are constants. Due to the strictly monotone increasing character of f, however, the graph of either of the above equations can intersect the graph of f in at most one point. Thus, the elements of S_1 and S_2 that specify the voltage across this diode must be equal. We can therefore conclude that the corresponding elements of S_1 and S_2 which specify the voltage across any diode are equal. That is, the diode voltages are unique for all diodes in the box at the top of Fig. 3.

We now consider the case in which some box at the bottom of Fig. 3 is equivalent to a current source. Let I_b denote the value of this current source (with reference direction chosen to be *out* of the base of the associated transistor). In this case, using the notation of Fig. 1(b), the variables v_1, i_1, v_2, and i_2, for the associated transistor, are constrained by the relationships:

$$i_1 = \frac{(1 - \alpha_{12}\alpha_{21})f_1(v_1) - \alpha_{12}I_b}{(1 - \alpha_{12})},$$
$$i_2 = \frac{(1 - \alpha_{12}\alpha_{21})f_2(v_2) - \alpha_{21}I_b}{(1 - \alpha_{21})}. \tag{18}$$

Thus, this transistor can be replaced by a pair of diodes (each in series with one of the resistors r_e, r_c) whose nonlinear conductance functions are specified by (18). We may now consider these diodes, these resistors, and the current source, all to be components of the box at the *top* of Fig. 3. We have thus shown, in summary, that when one (or more) of the boxes at the bottom of Fig. 3 is equivalent to a current source, the base-emitter and base-collector voltages of each transistor are still unique, since the network is then equivalent to a network of a type already considered.†

By use of the same type of argument that was applied to the case in which diodes are present in the box at the top of Fig. 3, the above results may, finally, be shown to be valid when diodes are present in the boxes at the bottom of Fig. 3.

The above results show the validity of the following statement concerning bistable networks: *One cannot synthesize a bistable network which consists of resistors, inductors, capacitors, diodes, independent voltage and current sources, and an arbitrary number of (Fig. 1b) transistors, and which has the structure of Fig. 3 when all capacitors are open-circuited and all inductors are short-circuited.*

APPENDIX

In this appendix we give the proof of a theorem which is used here and which is implied in Ref. 1 but is not stated explicitly there.

Theorem: If A is a real square matrix then $A \varepsilon P_0$ if and only if det $[A + D] \neq 0$ for every diagonal matrix $D > 0$.

† Here, of course, we use the proposition, proved above, that the voltage across each diode in the box at the top of Fig. 3 is unique.

Proof: (*if*) Suppose $A \not\varepsilon P_0$. 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 exists some value of $\zeta > 0$ such that $\det [\zeta I + A] = 0$. For this value of ζ let $D = \zeta I$.

If $\det A \geqq 0$ but, for some positive integer $k < n$, A has a $k \times k$ principal minor which is negative we may, without loss of generality, assume that 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 $\det A_1 < 0$. This is so because $\det [D + A]$ is not altered if any two rows and then the corresponding pair of columns are interchanged. Let $D^{(1)} = \text{diag}[d_1, \cdots, d_n]$ with $d_1 = \cdots = d_k = \xi$, where $\xi > 0$ is chosen so small that $\det[\xi I + A_1] < 0$. Then, with $d_{k+1} = \cdots = d_n = \zeta > 0$, we have

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

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

Thus, for $\zeta > 0$ chosen to be sufficiently large, $\det[D^{(1)} + A] < 0$. Now, if $D^{(2)} = \eta I$, for $\eta > 0$, then it is clear that for η chosen sufficiently large,

$$\det [D^{(2)} + A] = \eta^n \cdot \det \left[I + \frac{1}{\eta} A \right] > 0.$$

Thus, if

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

it is clear that there exists a value of ϵ, $0 < \epsilon < 1$, such that $\det [D(\epsilon) + A] = 0$.

(*only if*) By Theorem 1 of Ref. 1, since $A \varepsilon P_0$ and $D > 0$, $[D + A] \varepsilon P$. Thus, $\det [D + A] \neq 0$. □

REFERENCES

1. Sandberg, I. W. and Willson, A. N., Jr., "Some Theorems on Properties of DC Equations of Nonlinear Networks," B.S.T.J., *48,* No. 1 (January 1969), pp. 1–34.
2. 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.
3. Fiedler, M. and Pták, V., "Some Generalizations of Positive Definiteness and Monotonicity," Numer. Math., *9,* No. 2 (1966), pp. 163–172.
4. Sandberg, I. W., "On the Theory of Linear Multi-Loop Feedback Systems," B.S.T.J., *42,* No. 2 (March 1963), pp. 355–382.

Measuring Frequency Characteristics of Linear Two-Port Networks Automatically

By JAMES G. EVANS

(Manuscript received November 2, 1968)

This paper presents a new automatic technique for complete linear characterization of transistors and general two-port devices from standard insertion and bridging measurements. This technique includes a calibration sequence and mathematical transformation to provide parameters independent of actual test set impedances, as well as a special hardware design which allows for convenient self-measurement of the test set impedances. Knowledge of these impedances is used to reduce the measured quantities to arbitrary device parameters referenced entirely to a set of calibration standards. This independence of the parameters from the measuring set impedances allows for considerable reduction in the design constraints on the test set impedances and device connecting jigs.

I. INTRODUCTION

In implementing a linear two-port device characterizing facility on the computer operated transmission measuring set, several factors had to be considered.[1] First, the advantages of automated measurements could be retained only if the switching required to obtain four independent measurement configurations were done automatically. Second, the implementation must be broadband to take advantage of the 50 Hz to 250 MHz frequency range of the measuring set. Finally, the measurement method must be inherently capable of utilizing the high accuracy of the measuring set. An implementation was chosen which uses standard insertion and bridging measurements.[2–4] This choice is particularly compatible with the above factors.

Insertion and bridging measurements are made with the unknown terminated in a nominal impedance environment, in the present instance 50 ohms. At this impedance level, broadband, solenoid-operated coaxial switches are available which introduce only minor reflections in 50 ohm transmission circuits. In the computer operated transmission measuring set such switches are extensively used to provide automatic commuting of the unknown among the four independent configurations in which measurements are made. By arranging relay switching so that dc bias can be continuously maintained when measuring active unknowns such as transistors, multiple warm-up periods are eliminated and thermal equilibrium must be reached only once. Another advantage of the 50 ohm environment surrounding the unknown is that transistors and other active devices tend to be stable when terminated resistively.

If the terminal impedances deviate from the 50 ohm nominal, measurement data which assumes 50 ohm impedances will be in error. In the past, these errors were minimized to the best degree practical by controlling the impedance environment around the unknown. Even so, the lack of ideal circuit elements meant that significant errors remain in linear characterization data. In the computer operated transmission measuring set, the circuit elements are even less ideal because of impedance deviations resulting from the large number of coaxial relays used. The impedance control problem is further aggravated by the difficulty of designing low reflection dc bias networks to operate over several decades of frequency. For the latter reason the frequency range of measurement for devices requiring dc bias is confined to between 50 kHz and 250 MHz. The total measurement errors that could conceivably result from the residual impedance deviations would prevent meeting our accuracy targets.

A solution to this problem, which represents an advance over past practice, was to endow the measurement facility with the capability to self-measure the source and load impedance deviations around the unknown, thereby permitting measurement data to be corrected for the residual mistermination.

The effect achieved in the execution of this technique is to refer the corrected data to a set of calibration standards. It is not necessary to have carefully controlled terminal impedances, and high accuracy characterizations are obtainable using device-connecting jigs with poor terminal impedances.

A further advantage of the new technique lies in the greater analytical ease of converting measured data to two-port characterization sets of most direct interest to the designer. In past measuring arrangements, measured insertion ratios $e^{\phi_{21}}$ and $e^{\phi_{12}}$ are related to relevant parameters by the equations[3,4]

$$e^{\phi_{21}} = \frac{1 - s_{22}\rho_L - s_{11}\rho_g - \rho_g\rho_L(s_{12}s_{21} - s_{11}s_{22})}{s_{21}(1 - \rho_g\rho_L)}$$

and

$$e^{\phi_{12}} = \frac{1 - s_{11}\rho_L - s_{22}\rho_g - \rho_g\rho_L(s_{12}s_{21} - s_{11}s_{22})}{s_{12}(1 - \rho_g\rho_L)},$$

where the generator and load and reflection coefficients, ρ_g and ρ_L, and the scattering parameters are referred to the nominal design impedance, generally 50 or 75 ohms in earlier cases. Similarly, expressions can be derived for the measured input and output reflection coefficients in terms of all four of the desired parameters and the terminating reflection coefficients. Even if the test set reflection coefficients could be determined by independent measurement, the desired parameters cannot be obtained without recourse to a difficult mathematical inversion or a lengthy iterative calculation upon four coupled equations.

One of the important attributes of the new approach discussed in this paper is that the desired S parameters are explicitly dependent on known quantities and hence easily evaluated. This is made possible by initially finding the scattering parameters of the unknown, referred to the actual test set source and load impedances, as described in Section II.

II. MEASUREMENT AND IMPLEMENTATION

2.1 S *Parameter Representation*

The insertion and bridging measurement data are closely related to scattering parameters. (See Appendix A.) It should be recalled that there are several types of scattering parameters.[5-7] The calculations which follow deal exclusively with voltage scattering parameters. In the present context, these parameters are defined with respect to the terminal impedances which actually prevail in the test set. The voltage S parameters along with their normalizing impedances can be transformed to any other parameter representation by well known transformations.[5]

2.2 *Measurement of* s_{12} *and* s_{21}

The transmission S parameters, s_{12} and s_{21}, are obtained almost directly from the insertion measurement illustrated in Fig. 1. With

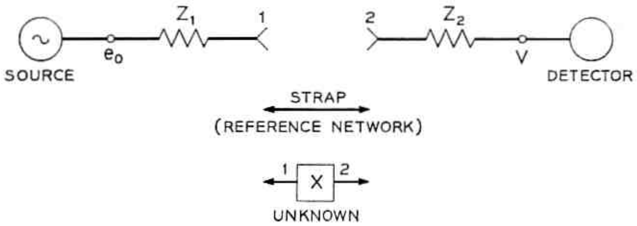


Fig. 1 — Model of S_{12}, S_{21} measurement. Circuit shown for S_{21} measurement; detector and source interchanged for S_{12} measurement.

the unknown inserted in the measuring network the detected voltage V is directly proportional to s_{21} of the unknown defined with respect to the normalizing impedances Z_1 and Z_2. (See Appendix A.)

$$\{s_{21}\}_{Z_1,Z_2} = \{R_{21}\cdot V\}_X \, . \tag{1}$$

The constant of proportionality R_{21} can be determined by inserting a reference network with known s_{R21}

$$\{s_{21}\}_{Z_1,Z_2} = \{s_{R21}\}_{Z_1,Z_2} \frac{V_X}{V_R}. \tag{2}$$

Typically the reference network is a coaxial line of known electrical length. When ports 1 and 2 can be directly connected the line is of zero length for which s_{R21} takes the simple form

$$\{s_{R21}\}_{Z_1,Z_2} = \frac{2\cdot Z_2}{Z_1 + Z_2}. \tag{3}$$

This result can be seen directly or derived from equation (31) in Appendix C. The results for a reference line of finite length are derived as an example in Appendix C.

The hardware implementation is such that the impedances seen to the left and right of the unknown are essentially the same for s_{12} and s_{21} measurement. This invariance of the terminal impedances, with respect to interchange of the source and detector, is accomplished by using a pair of judiciously located 20 dB pads. The hardware details are described in Section 2.5. The measurement of s_{12} is similar to that of s_{21}. For this case,

$$\{s_{12}\}_{Z_1,Z_2} = \{s_{R12}\}_{Z_1,Z_2} \frac{V_X}{V_R}. \tag{4}$$

2.3 *Measurement of* s_{11} *and* s_{22}

The quantity s_{11} of an unknown device is obtained indirectly by a bridging measurement. For this measurement the source and detector are directly connected as illustrated in Fig. 2. The device to be measured, as well as the calibration standards, are successively bridged across the source-detector interconnection. The implementation is such that the impedance seen at terminal 2 is Z_2, the same impedance as for the previously described s_{12} and s_{21} measurement.

The quantity s_{11} is determined by making four measurements of V. Three consist of calibration measurements using open, short, and reference impedance standards. The fourth measurement is made with the unknown connected. These measurements can be combined to obtain a reflection coefficient

$$\{s_{11}\}_{Z_R,Z_2}$$

closely related to the desired

$$\{s_{11}\}_{Z_1,Z_2}\,.$$

$$\{s_{11}\}_{Z_R,Z_2} = \frac{(V_X - V_R)(V_\infty - V_0)}{(V_\infty - V_R)(V_X - V_0) + (V_R - V_0)(V_\infty - V_X)}. \tag{5}$$

The term

$$\{s_{11}\}_{Z_R,Z_2}$$

does not have the desired impedance normalization on port 1. The desired result is obtained through the transformation

$$\{s_{11}\}_{Z_1,Z_2} = \left\{\frac{s_{11} - \Gamma_1}{1 - \Gamma_1\cdot s_{11}}\right\}_{Z_R,Z_2} \tag{6}$$

where

$$\Gamma_1 = \frac{Z_1 - Z_R}{Z_1 + Z_R} \tag{7}$$

is obtained from an additional measurement described in Section 2.4. Notice that Γ_1 is the reflection coefficient of Z_1 normalized with respect to Z_R. Hence, once Z_1 becomes known, all of the information is available to compute the value of $\{s_{11}\}_{Z_1,Z_2}$. The quantity s_{22} is determined in an analogous manner. In this case the hardware implementation is such that the impedance terminating port 1 of the unknown is Z_1, the same impedance as in the s_{12} and s_{21} measurements; it is seen that the

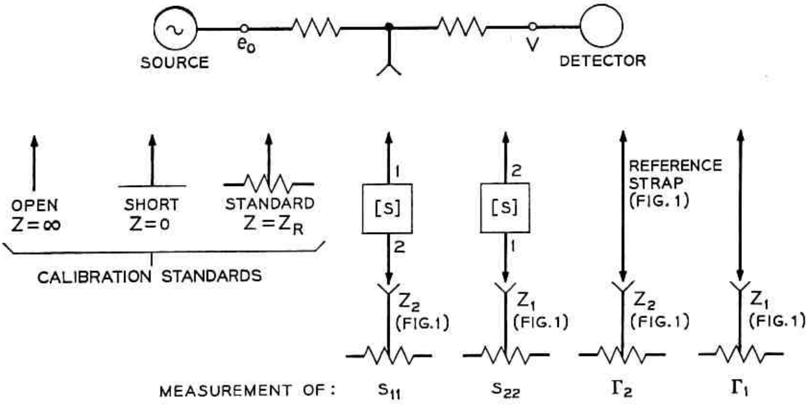


Fig. 2 — Model of S_{11}, S_{22}, Γ_1, and Γ_2 measurement.

measured data this time yield s_{22} with respect to Z_1 at port 1 and Z_R at port 2. The term $\{s_{22}\}_{Z_1,Z_2}$ is then determined as before, as a function of the four measured detector voltages and Γ_2, where

$$\Gamma_2 = \frac{Z_2 - Z_R}{Z_2 + Z_R}. \tag{8}$$

(Appendix B shows that the equations of Section 2.3 are general and apply for any linear network interconnecting the source, the detector, and the unknown.)

The above discussion has described a procedure for determining the four voltage scattering parameters $\{s\}_{Z_1,Z_2}$. This set can be transformed to a more useful parameter representation only if Z_1 and Z_2 or equivalently Γ_1 and Γ_2 can be determined. Section 2.4 discusses the procedure for determining Γ_1 and Γ_2.

2.4 *Measurement of* Γ_1 *and* Γ_2

The measurement procedure for determining Γ_2 is similar to that for evaluating s_{11}. With the reference transmission strap used in the s_{12} and s_{21} measurements inserted in the bridging configuration of Fig. 2 and terminated in Z_2, the detected voltage is V_{Γ_2}. The reflection coefficient computed from equation (5) with V_X replaced by V_{Γ_2} is the reflection coefficient of Z_2, with respect to Z_R, as viewed through the reference strap, which has known transforming properties. When ports 1 and 2 can be directly connected the line section is of zero length for which the transformation is unity. In this event the reflection coefficient computed from equation (5) is equal to Γ_2 . The term Γ_1 is evaluated in a similar manner using the s_{22} bridging configuration.

Notice that Z_1 and Z_2 can be determined on a broadband basis simply with two additional calibration measurements. It is important to realize that this is possible only because of the particular physical embodiment which results in the network being terminated in Z_2 on port 2 during the s_{12}, s_{21}, and s_{11} measurements and in Z_1 on port 1 during the s_{12}, s_{21}, and s_{22} measurements. Section 2.5 describes the hardware arrangement. Notice that for an arrangment in which the terminal impedances remain invariant under all four S measurement conditions these two additional measurements are redundant. In this case, the terminal impedances can be determined from the open, short, and standard impedance measurements. (See Ref. 8.)

2.5 *Physical Embodiment*

Figure 3 is a simplified schematic diagram of the 2-port linear characterization facility. The components L_1, C_1, L_2, and C_2 comprise the bias networks necessary for supplying dc bias to devices such as transistors. The attenuators P_1 and P_2 have an insertion loss of 20 dB. These attenuators play a critical role in maintaining the terminal impedance constancy described earlier.

The coaxial switch closures and the circuit paths of a s_{21} measurement are specifically shown in Fig. 4. The terminal impedances seen to the left and right of ports 1 and 2 are Z_1 and Z_2, respectively. A simple examination of Fig. 3 will reveal that the switch closures and paths inside P_1 and P_2 are identical for the s_{12} and s_{21} measurements. The switching necessary to convert to the s_{12} measurement changes the reflection coefficient seen looking to the left of P_1 and right of P_2 by less than 0.1. These changes are attenuated by P_1 and P_2 so that the changes in Γ_1 and Γ_2 seen at the terminals 1 and 2, respectively, are less than 0.001. Within the bounds of neglecting a possible 0.001 change, Γ_1 and Γ_2, and therefore Z_1 and Z_2, are invariant under the change from the s_{21} to the s_{12} measurement.

Figure 5 illustrates the switch closures and circuit paths of an s_{22} measurement. Attenuator P_2 has been eliminated from this measurement to prevent a loss in measurement resolution. Notice that all switch closures and circuit paths between terminal 1 and attenuator P_1 are the same as in the s_{12} and s_{21} measurements. The reflection coefficient seen to the left of P_1 has changed less than 0.1 in switching into this measurement mode. Therefore, the reflection

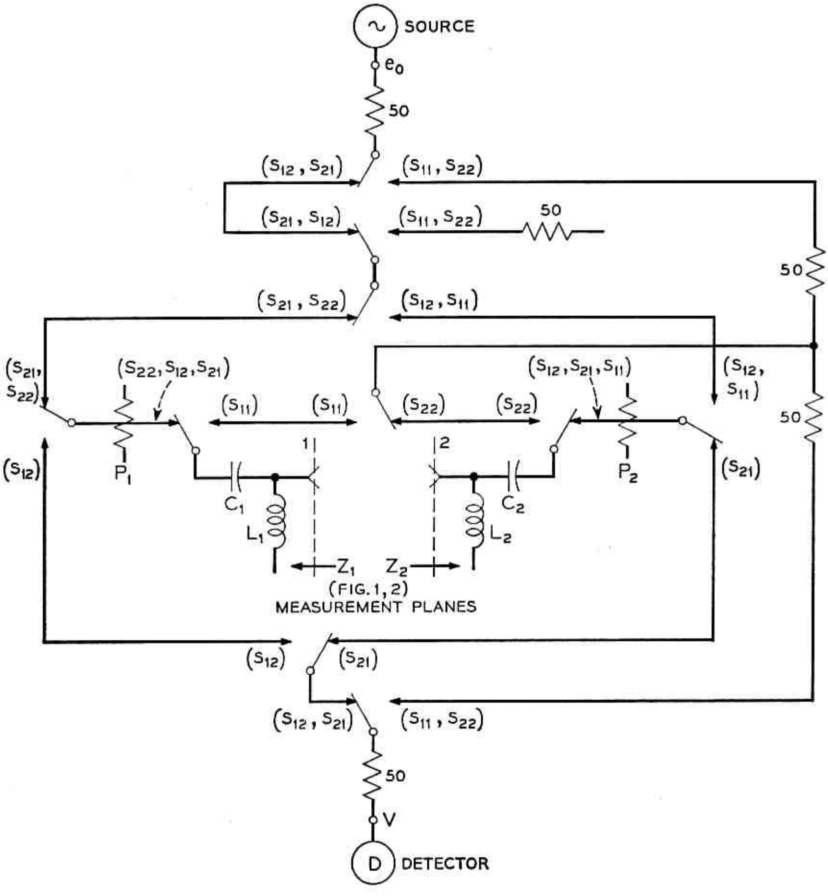


Fig. 3 — Simplified schematic of transistors measurement unit. Signal routing required to set up measurement paths for the determination of S_{12}, S_{21}, S_{11}, or S_{22} are shown for each switch.

coefficient seen looking into terminal 1 cannot differ from Γ_1 by more than 0.001, a negligible amount. A similar analysis of the s_{11} measurement mode reveals that the reflection coefficient seen looking into terminal 2 cannot differ from Γ_2 by more than 0.001.

2.6 *Other Parameter Representations*

The set $\{s\}_{Z_1, Z_2}$, $Z_1(\Gamma_1)$ and $Z_2(\Gamma_2)$ is a well defined voltage scattering parameter representation of the linear characteristics of the 2-port unknown. This representation has no practical application, but it can be easily converted to a more useful representation by well known transformations (Ref. 6). One particularly useful and easily obtained parameter set is the voltage scattering parameters normalized to 50 ohms (Z_R = 50 ohms). This set is obtained from the measured set by use of the transformations of Appendix C:

$$\{s_{11}\}_{Z_R,Z_R} = \left\{\frac{\Gamma_1 + s_{11} + \Gamma_1\Gamma_2 s_{22} + \Gamma_2\,\Delta s}{1 + \Gamma_1 s_{11} + \Gamma_2 s_{22} + \Gamma_1\Gamma_2\,\Delta s}\right\}_{Z_1,Z_2} \tag{9}$$

$$\{s_{12}\}_{Z_R,Z_R} = \left\{\frac{s_{12}(1 - \Gamma_1)(1 + \Gamma_2)}{1 + \Gamma_1 s_{11} + \Gamma_2 s_{22} + \Gamma_1\Gamma_2\,\Delta s}\right\}_{Z_1,Z_2} \tag{10}$$

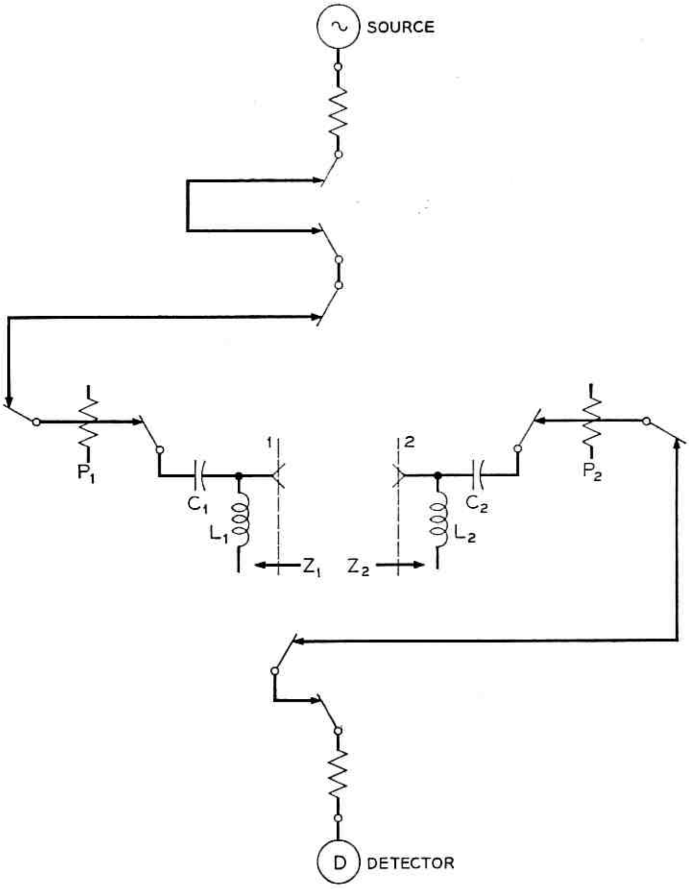


Fig. 4 — Path through transistor measurement unit for S_{21} evaluation.

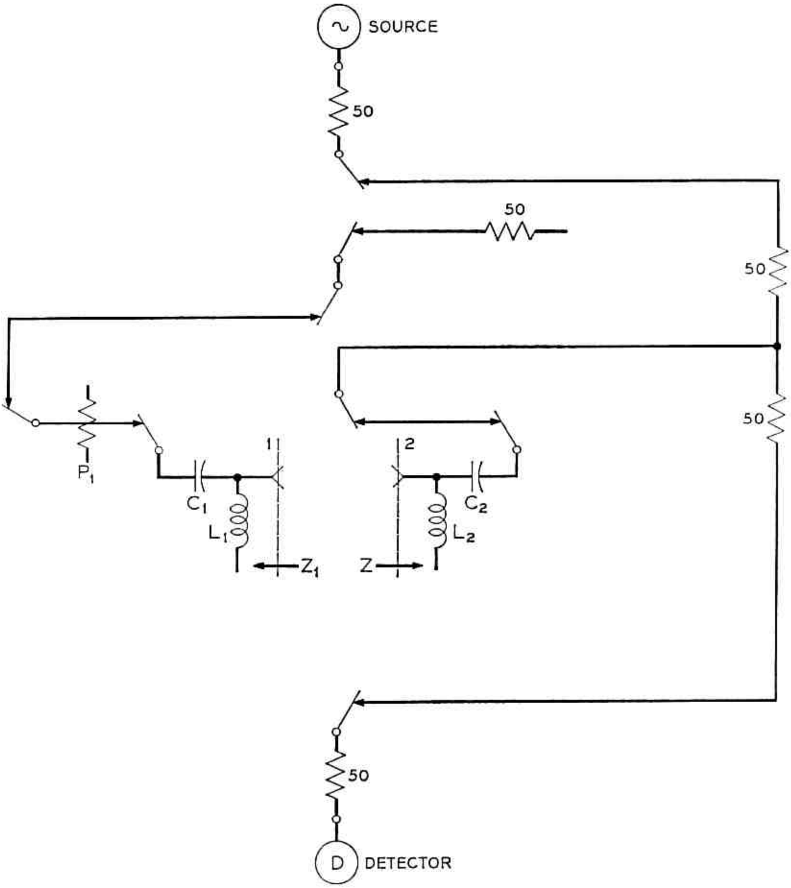


Fig. 5 — Path through transistor measurement unit for S_{22} evaluation.

$$\{\Delta s\}_{Z_1,Z_2} = \{s_{11}s_{22} - s_{12}s_{21}\}_{Z_1,Z_2} \; . \tag{11}$$

The transformations for s_{22} and s_{21} are found by transposing subscripts in the above equations. It is useful to realize that since the transformed S parameters are normalized to equal real impedances they are numerically equal to the current and power scattering parameters with the same normalization.

The transformed parameters are independent of Z_1 and Z_2, depending only upon the open, short, Z_R, reference network calibration standards and the loss and phase measurement accuracy of the test set. This independence from Z_1 and Z_2 allows for useful freedom in the design of the measuring apparatus terminal impedances.

III. SPECIAL CALIBRATION FEATURES

In Section II a measurement technique has been presented in which the linear characteristics of a 2-port unknown are determined relative to four calibration standards. An idealized set consisting of an open, short, standard impedance and a zero length line were treated for mathematical simplicity. The calibration standards used with the automated facility deviate considerably from this idealized set over the broad frequency range of interest. A failure to compensate for these deviations would adversely affect the accuracy of linear characterization. Compensation is accomplished by modeling the deviations as a function of frequency and then computationally accounting for them in the data reduction program.

3.1 *Compensation for a Transmission Reference Line of Nonzero Length*

Physical constraints often make it impossible to directly interconnect ports 1 and 2 for the measurements needed for the determination of s_{12}, s_{21}, Γ_1 and Γ_2. Interconnection is achieved in these cases by using a short transmission line with a characteristic impedance equal to Z_R and an electrical length equal to θ. For this network

$$\{s_{R12}\}_{Z_R, Z_R} = \{s_{R21}\}_{Z_R, Z_R} = e^{-i\theta},$$

where θ is computed from the line constants and length.

For determining Γ_1 and Γ_2 this line has a particularly simple transforming property. Γ_1 and Γ_2 when viewed through the line appear as $\Gamma_1 e^{-i2\theta}$ and $\Gamma_2 e^{-i2\theta}$, respectively.

In the actual measurements, determining s_{12} and s_{21} of the unknown require s_{R12} and s_{R21} of the line normalized with respect to Z_1 and Z_2. These quantities are obtained by transforming from the (Z_R, Z_R) impedance normalization to the (Z_1, Z_2) impedance normalization as illustrated in Appendix C. The data reduction program allows for reference lines of arbitrary length.

3.2 *Compensation for Nonideal Bridging Calibration Standards*

Typically high quality standard impedance (Z_R) terminations are available whose deviations from nominal are negligible. This is not true for the open and short calibration standards. Mechanical considerations sometimes require that the short and open reference planes be displaced from the measurement plane by a section of transmission line. In addition, the "open" differs from ideal by a fringing capacitance. For small reactive perturbations or arbitrary displacements in a transmission line of characteristic impedance Z_R, the actual open and short circuit reflection coefficients are of the simple form

$$\Gamma_\infty \big|_{Z_R} = e^{-2j\tau_\infty \omega} \tag{12}$$

$$\Gamma_0 \big|_{Z_R} = -e^{-2j\tau_0 \omega} \tag{13}$$

where τ_∞ and τ_0 are the time delays for the lengths of line involved, including a correction for fringing capacitance at the end. The linear dependence of the reflection phase angles on frequency facilitates broadband computational correction. In the data reduction program a more general bridging equation than equation (5) (see Appendix B) is programmed to allow for calibration standards of the above form.

3.3 *Measurement Plane Translations*

For some unknowns it is desirable to define the reference planes of the S parameters translated down 50 ohm transmission lines from the measurement planes. Examples are the air line measurements in Section 5.1, and the case of measuring an integrated circuit connected to the test set by 50 ohm microstrip transmission lines of significant electrical length when information about the chip alone is sought. (An alternative approach to characterization would be to develop integrated circuit standards so that calibration could be performed at the chip interface.) Analysis shows that the presence of transmission lines of electrical length α are accounted for within the previously developed mathematical framework by entering the translated angles $-\omega\tau_\infty + 2\alpha$, $-\omega\tau_0 + 2\alpha$ and $\theta - 2\alpha$ instead of the physical angles $-\omega\tau_\infty$, $-\omega\tau_0$ and θ into the data reduction program. The alternative approach, requiring additional programming, would "remove" the transmission lines by appropriate matrix manipulations.

IV. STATEMENT OF ERRORS IN $s_{50,50}$ PARAMETERS

This section gives the results of an approximate worst case error analysis for $s_{50,50}$ parameters. The analysis was performed on these parameters because of the mathematical simplifications resulting from their similarity to the measured quantities. The results are derived by assuming that Γ_1, Γ_2, and the fundamental error terms are small compared with unity. This approximation allows for simplification of the equations presented in the earlier sections. The fundamental error terms are added on a worst case basis to obtain overall error bounds. The lengthy analysis has been omitted for brevity.

4.1 *Bound for the* s_{11} *and* s_{22} *Measurements*

The errors in the s_{11} and s_{22} measurement arise from two principal sources, those associated with the bridging technique and those from interaction with the termination of the unknown. For an unknown with $| s_{12} \cdot s_{21} | \ll 1$ the latter error source is negligible and the bound for errors in the determination of s_{11} is,

$$| \Delta s_{11} |_{50,50} < \{0.0023 + 0.0023 \mid 1 - s_{11}^2 \mid + 0.0013 \mid s_{11} \mid \mid 1 + s_{11} \mid + 0.0013 \mid s_{11} \mid \mid 1 - s_{11} \mid + \mid \Gamma_s \mid \mid 1 - s_{11}^2 \mid\}_{50,50} \qquad (14)$$

Implicit in equation (14) is the assumption that the uncertainties in the phase angles of the open and short circuit standards are less than 0.02 degree.

$$\Gamma_s = \frac{Z_R - 50}{Z_R + 50}. \qquad (15)$$

Γ_s is the reflection coefficient of Z_R with respect to 50 ohms, which independent measurements have shown to be less than 0.005. Notice that certain terms in equation (14) disappear when s_{11} equals 0, -1, and 1. This reduction of the error bound occurs when the bridging measurement of the unknown reduces to a differential comparison of the unknown with either the 50 ohm, short, or open standard.

When the product $s_{12} \cdot s_{21}$ is not negligible, errors that occur in determining the reflection coefficient of the termination, Γ_2 for the s_{11} measurement, are transformed through the unknown to increase the error of the s_{11} determination. If $\Delta\Gamma_2$ is the error in determining Γ_2, the term

$$| \Delta\Gamma_2 | \, | s_{12} \cdot s_{21} |_{50,50} \qquad (16)$$

must be added to equation (14) to account for this second error source. Since Γ_2 is determined from a bridging measurement, the bound on $\Delta\Gamma_2$ can be computed directly from equation (14) for $\Gamma_2 \ll 1$.

$$| \Delta\Gamma_2 | < 0.005 + | \Gamma_s | < 0.01. \qquad (17)$$

If Γ_2 were not determined by measurement, then $\Delta\Gamma_2$ in equation (16) would have to be replaced by the worst case estimate of the value of Γ_2. From Fig. 6, Γ_2 is seen to be as large as 0.08. The error term of equation (16) is eight times larger in this case.

The relationship for Δs_{22} is found by changing subscripts.

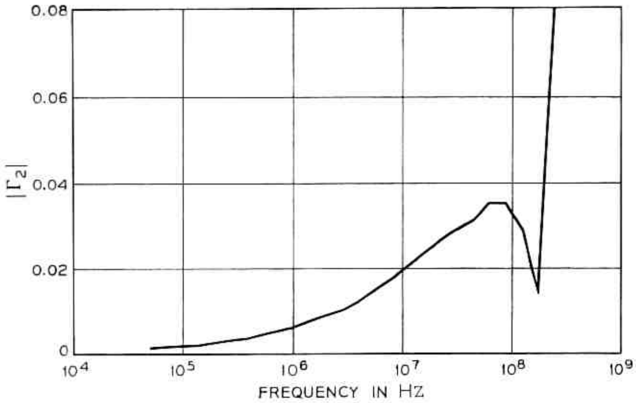


Fig. 6 — Test set port 2 reflection coefficient defined with respect to 50 ohms.

4.2 *Error Bound for the* s_{12} *and* s_{21} *Measurement*

The worst case fractional error in the determination of s_{12} is

$$\left|\frac{\Delta s_{12}}{s_{12}}\right|_{50,50} < \{0.0013 + 0.0013/|s_{12}| + |\Delta\theta| + |\Delta\Gamma_1|\cdot|s_{11}| + |\Delta\Gamma_2|\cdot|s_{22}|\}_{50,50} \tag{18}$$

The term that increases as $|s_{12}|$ decreases shows the intuitively appealing result that the relative error bound increases as the signal-to-noise ratio decreases. The term $\Delta\theta$ is the uncertainty in the electrical length of the reference network (zero line). The value of $\Delta\theta$ is less than 0.001 radians for the typical reference transmission line network. When direct interconnection of the measurement ports is possible, θ and therefore $\Delta\theta$, equals zero.

The terms $\Delta\Gamma_1$ and $\Delta\Gamma_2$ arise from the uncertainties in the knowledge of the terminating reflection coefficients, Γ_1 and Γ_2. The terms $\Delta\Gamma_1$ and $\Delta\Gamma_2$ are less than 0.01, as described in Section 4.1.

If Γ_1 and Γ_2 are not determined by measurement then $\Delta\Gamma_1$ and $\Delta\Gamma_2$ in equation (18) must be replaced by worst case values for Γ_1 and Γ_2. From Figs. 6 and 7, Γ_1 and Γ_2 are as large as 0.08 resulting in an eight fold increase in the mistermination terms of equation (18).

V. MEASUREMENTS TO CONFIRM ACCURACY

The two-port properties of a precision air-line, a precision attenuator, and a common base transistor were measured. These unknowns are useful for demonstrating the accuracy of small signal characterization over a wide range of test parameter magnitudes.

5.1 *Characterization of a Precision* 30-cm *Air Line*

A General Radio 900-L30 precision 14 mm air line was measured on the automated facility and the data processed to yield scattering parameter data. The calibration standards consisted of a General Radio 90-W50 coaxial 50 ohm standard, a General Radio 900-WN coaxial short circuit, a coaxial open circuit, and a zero length line. The test set measurement ports (General Radio 900) were at the ends of flexible cable, thus allowing direct interconnection for the reference insertion ("zero-line") measurement. The open circuit standard, consisting of an unterminated General Radio 900 connector, was corrected for 0.16 pF of fringing capacitance by the techniques of Section 3.2.

The $s_{50,50}$ ohm parameters of the line were computed. This matrix is symmetrical and therefore only s_{11} and s_{12} data are presented in Figs. 8 through 11. The results for an ideal air line are $s_{11} = 0$ and $s_{12} = e^{-j\omega\tau}$, where $\omega\tau$ equals 90° at 250 MHz. To expose the errors of measurement, the reference plane translation technique of Section 3.3 was used to remove the linear phase component from all the S parameters. Alternatively, the translation is equivalent to multiplying each matrix element by $e^{j\omega\tau}$. For an ideal air line the resulting $s_{50,50}$ parameters are $s_{11} = 0$ and $s_{12} = 1$. The data in Figs. 8 through 11 indicate how the actual air line deviates from these ideal values.

The deviations can be accounted for by skin effect losses. The high

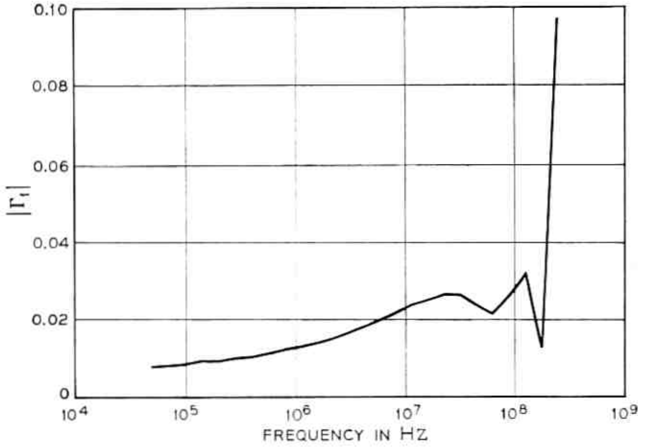


Fig. 7 — Test set port 1 reflection coefficient defined with respect to 50 ohms.

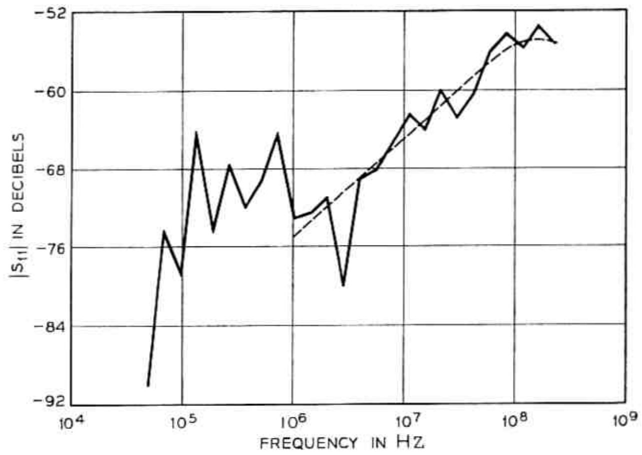


Fig. 8 — Magnitude of $S11_{50,50}$ for a precision 30 cm airline with the linear phase component subtracted.

frequency (1 MHz or greater for this example) $s_{50,50}$ parameters for a transmission line deviating from the ideal because of skin losses are derived in Appendix D. Multiplication of these parameters by $e^{j\omega\tau}$ converts them to a form compatible with the figures.

$$\{s_{11}\}_{50,50} = \lambda(z\omega\tau)^{\frac{1}{2}}\left(\frac{\sin\omega\tau}{\omega\tau}\right)\exp\left(j^{\pi/4}\right) \tag{19}$$

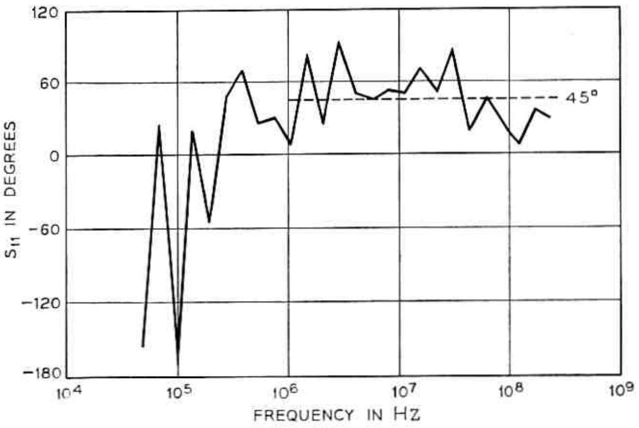


Fig. 9 — Phase of $S11_{50,50}$ for a precision 30 cm airline with the linear phase component subtracted.

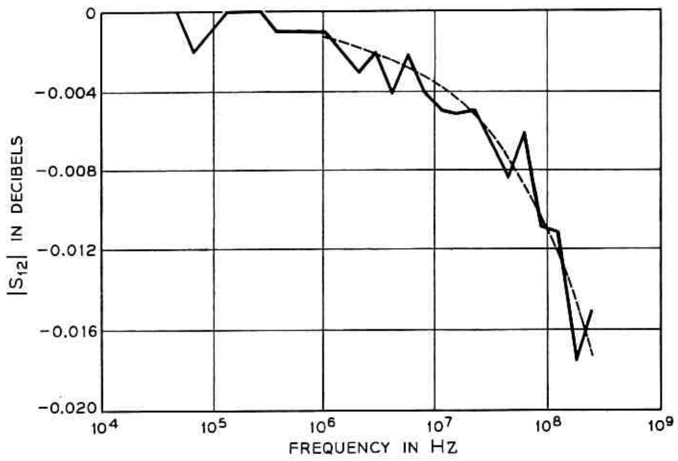


Fig. 10 — Magnitude of $S12_{50,50}$ for a precision 30 cm airline with the linear phase component subtracted.

$$\{s_{12}\}_{50,50} = \exp\,[-\lambda(\omega\tau)^{\frac{1}{2}}]\,\exp\,[-j\lambda(\omega\tau)^{\frac{1}{2}}] \tag{20}$$

where λ is a frequency independent, skin effect parameter dependent on surface conductivity as well as other line parameters. The value of λ was determined to be 0.0016 by fitting the magnitude of s_{12} from equation (20) to the measured results. The smooth curve in Fig. 10 shows that the fit to the magnitude of s_{12} is typically better than 0.002 dB. Using this value of λ, the phase of s_{12} and the magnitude of s_{11} were computed. The results are plotted in the Figs. 8 and 11.

The deviations of the measured results from the theoretical skin loss curves are estimators of characterization accuracy. Most of the deviations result from the sensitivity limits of 0.001 dB and 0.01°. The observed deviations are an order of magnitude smaller than the worst case errors predicted by the equations of Sections 4.2 and 4.3. If the mistermination corrections were not performed, the errors in s_{11} and s_{22} would be substantially larger. For example s_{11} would be virtually equal to Γ_2, the reflection coefficient of port 2. The values of Γ_2 are plotted in Fig. 7, showing that the resulting error in s_{11} could be as large as 0.08.

5.2 *Characterization of a Precision Attenuator*

A General Radio 900-G6 precision 14 mm 6 dB attenuator was measured and the data processed to obtain $s_{50,50}$ parameters. The true value

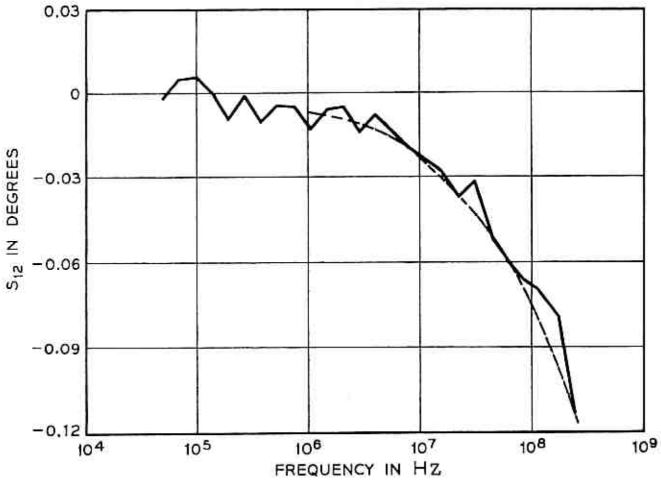


Fig. 11 — Phase of $S12_{50,50}$ for a precision 30 cm airline with the linear phase component subtracted.

of these parameters are not sufficiently well known for the attenuator to be used as a measurement standard. However, its measurement is useful in verifying characterization accuracy by comparing s_{12} and s_{21} . The terms $\{s_{12}\}_{50,50}$ and $\{s_{21}\}_{50,50}$ must be equal for a reciprocal network; $\{s_{12}\}_{Z_1,Z_2}$ and $\{s_{21}\}_{Z_1,Z_2}$ are in general not equal since typically $Z_1 \neq Z_2$. Therefore, the agreement between the 50 ohm S parameters is a measure of the success to which $\{s_{12}\}_{Z_1,Z_2}$, $\{s_{21}\}_{Z_1,Z_2}$, Z_1 , and Z_2 have been determined.

The magnitudes of s_{12} and s_{21} are plotted in Fig. 12 for comparison. The midfrequency values for s_{12} and s_{21} are close to the dc measured value of -6.0151 dB. The attenuation bump below 10^4 Hz is from a poorer test set audio frequency signal-to-interference ratio. The agreement between s_{12} and s_{21} over most of the 400 Hz to 250 MHz range is better than a few thousands of a dB. In Fig. 13 the difference between the phase angles of these two parameters is plotted. The typical agreement is again excellent, better than several hundredths of a degree. The above differences are well within the 0.035 dB and 0.23° fractional error bounds on s_{12} and s_{21} computed from equation (18)

5.3 *Common Base Transistor Measurement*

A medium-power silicon transistor was measured in the common base configuration. The measurement data were then processed to obtain common emitter h parameters by way of illustrating the flexibility of the data reduction program. The parameter h_{21} (or β) with and without mistermination errors is shown in Figs. 14 and 15. Differences between the corrected and uncorrected data as large as 5 dB and 20° are readily apparent. Discrepancies of this magnitude result from the β^2 multiplication of errors which occurs when converting common base parameters to common emitter parameters. Notice that the discrepancies decrease as the magnitude of β decreases. The agreement between corrected common emitter β curves derived from measurements in the common base, common collector, or common emitter modes is typically better than 1 dB and 5°. (See Ref. 1, Fig. 25.)

VI. SUMMARY

Complete device characterization can be rapidly and accurately achieved by the measurement method described in this paper. Loss of accuracy caused by nonideal test set terminations is virtually eliminated by measuring the deviations from ideal. The errors that arise are now the result of the smaller inaccuracy in measuring the deviations rather than to the gross deviations themselves. The self-measurement of the termination deviations is done at any frequency by two extra calibration measurements.

All measurements refer to a set of calibration standards, thereby making the derived parameters independent of the impedance properties of the test set. This attribute should facilitate the measurement

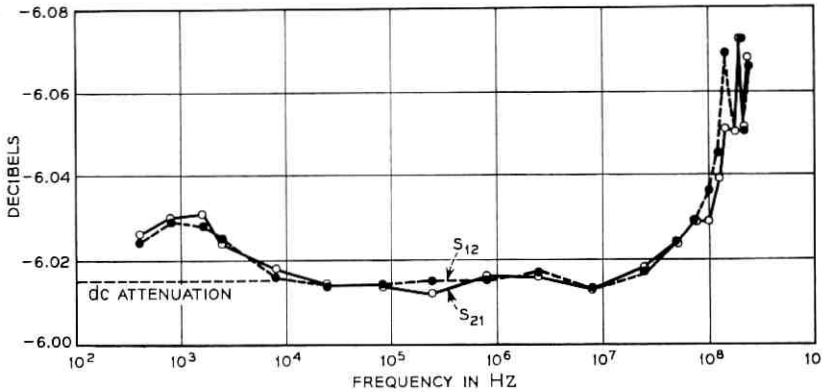


Fig. 12 — $|S_{12}|$ and $|S_{21}|$ of a precision 6 dB coaxial attenuator.

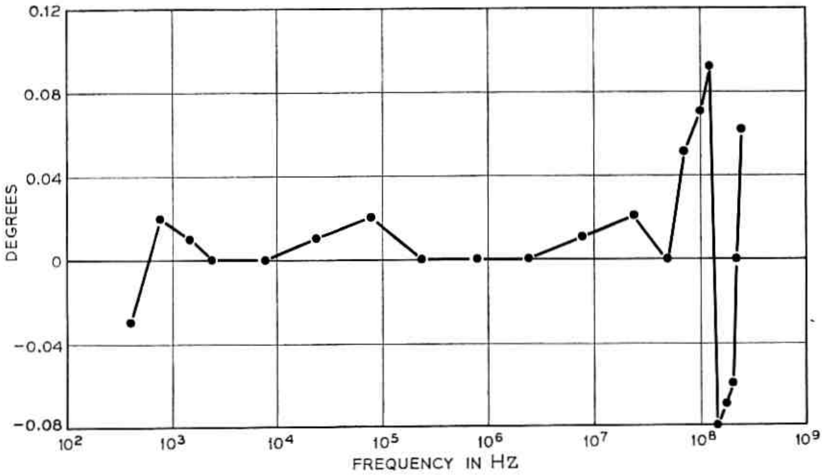


Fig. 13 — Angle of S_{21} minus the angle of S_{12} for the precision 6 dB coaxial attenuator.

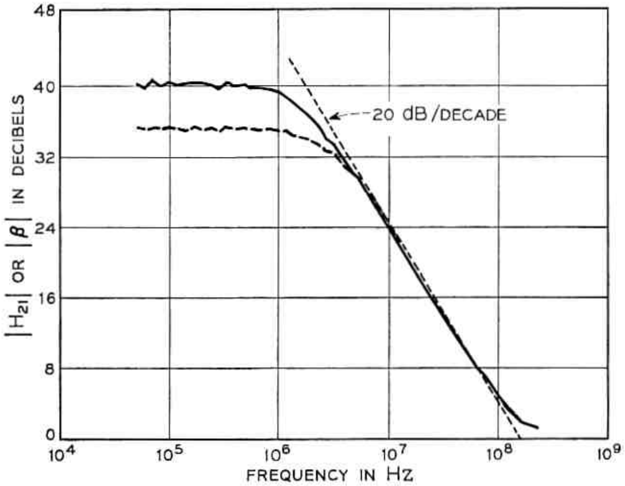


Fig. 14 — Common emitter $h21$ parameter computed from a common base measurement configuration. Mistermination errors are removed from the dashed curve and not from the solid curve.

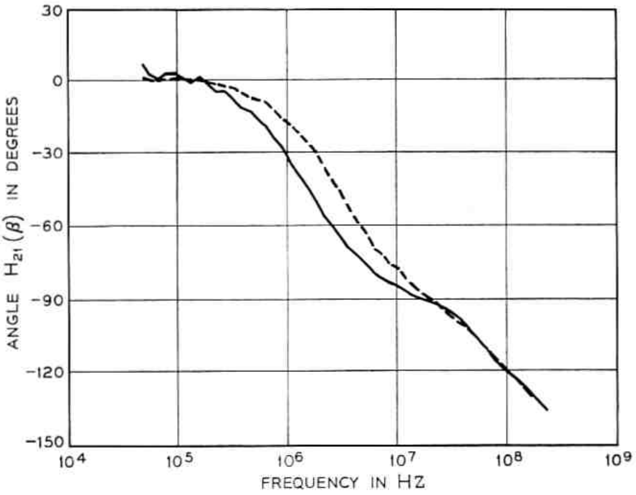


Fig. 15 — Common emitter *h*21 parameter computed from a common base measurement configuration. Mistermination errors are removed from the dashed curve and not from the solid curve.

of integrated circuits, for only the integrated calibration standards and not the connecting jig determine the accuracy of measurement.

VII. ACKNOWLEDGMENTS

R. G. Conway is responsible for a considerable portion of the mechanical design and assembly. The bridging equation deviation is an extension of the work of G. F. Critchlow. The author is especially indebted to G. D. Haynie and D. Leed for helpful criticism and advice.

APPENDIX A

Voltage Scattering Parameters

A network can be characterized in terms of traveling waves at selected reference planes rather than in terms of currents and voltages. Voltage scattering parameters are one such traveling wave representation. These parameters relate the reflected voltages from a network to the incident voltages. The matrix notation for this relationship for a two-port network is

$$\begin{bmatrix} V_{1r} \\ V_{2r} \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \Bigg|_{Z_1, Z_2} \begin{bmatrix} V_{1i} \\ V_{2i} \end{bmatrix}. \tag{21}$$

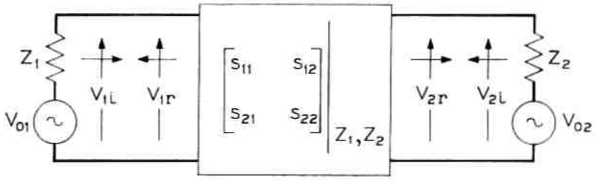


Fig. 16 — Scattering coefficient representation of a two-port network.

V_{1i} and V_{2i} are the incident voltage waves appearing at the port 1 and port 2 reference planes, respectively. V_{1r} and V_{2r} are the respective reflected voltage waves. The S parameters relating to the incident and reflected waves are defined with respect to the incident waves source impedances, Z_1 and Z_2.

The incident waves are related to the source potentials V_{01} and V_{02} (see Fig. 16) as follows,

$$V_{1i} = V_{01}/2 \qquad V_{2i} = V_{02}/2. \tag{22}$$

The conventional voltages appearing at ports 1 and 2 are

$$V_1 = V_{1i} + V_{1r} \qquad V_2 = V_{2i} + V_{2r} \, . \tag{23}$$

From equations (21), (22), and (23) it is easy to verify that

$$V_2 = \{(V_{01}/2)s_{21}\}_{Z_1, Z_2} \tag{24}$$

when $V_{02} = 0$. Therefore V_2 is directly proportional to $\{s_{21}\}_{Z_1, Z_2}$ when the network is inserted between a source of impedance Z_1 and a load of impedance Z_2. Also when $V_{02} = 0$ the reflection coefficient defined as V_{1r}/V_{1i} is equal to $\{s_{11}\}_{Z_1, Z_2}$.

APPENDIX B

The Bridging Technique

The bridging technique is one method of determining the input and output scattering parameters of a device. This technique requires an oscillator, a detector, three impedance standards and an arbitrary three-port linear network. The judicious selection of this network will lead to better measurement sensitivity.

The essentials of a bridging measurement are illustrated in Fig. 17. Z and E_{03} comprise the Thévenin's equivalent circuit of port 3, the measurement port. An analysis shows that the detected voltage V remains unchanged if the unknown impedance Z_X is replaced by the load and controlled source combination shown.

$$\Gamma = \frac{Z_X - Z}{Z_X + Z}. \tag{25}$$

Since the interconnecting network is linear,

$$V = A \cdot E + B \cdot \Gamma \cdot E_{03} \tag{26}$$

and

$$E_{03} = C \cdot E. \tag{27}$$

A, B, and C are system constants. Equation (26) is therefore seen to reduce to the form

$$V = a + b \cdot \Gamma \tag{28}$$

where $a = A \cdot E$ and $b = B \cdot C \cdot E$.
This equation has three unknowns, the two constants a and b, and the normalizing impedance Z. These constants are determined by three independent measurements made with Z_X (see Fig. 2) replaced with three calibration standards.

Three standards that are readily obtainable and which lead to computational simplicity are an open, short, and termination (R); the reflection coefficients defined with respect to Z are 1, $-$ 1, and Γ_R, respectively. A somewhat lengthy computation reveals that

$$\{\Gamma_X\}_R = \frac{(V_X - V_R)(V_\infty - V_0)}{(V_\infty - V_R)(V_X - V_0) + (V_R - V_0)(V_\infty - V_X)}. \tag{29}$$

$\{\Gamma_X\}_R$ is the reflection coefficient of Z_X normalized with respect to R.

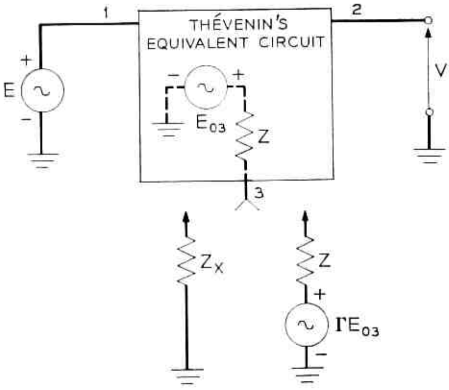


Fig. 17 — Simplified schematic of bridging measurement.

APPENDIX C

Change of the S *Parameter Impedance Normalization*

The voltage S parameters will describe the traveling wave properties of a network terminated in a particular impedance environment. In a different environment the description is no longer valid. Therefore, for example, it is not possible to measure the S parameters of a network in a 50 ohm test set and then use these parameters directly to describe the network performance in a 75 ohm system. It is possible to transform S parameters with one impedance normalization to those of another impedance normalization. These transformations are:

$$\{s_{11}\}_{Z_1,Z_2} = \left\{\frac{\Gamma_1 + s_{11} + \Gamma_1\Gamma_2 s_{22} + \Gamma_2\,\Delta s}{1 + \Gamma_1 s_{11} + \Gamma_2 s_{22} + \Gamma_1\Gamma_2\,\Delta s}\right\}_{Z_{01},Z_{02}} \tag{30}$$

$$\{s_{12}\}_{Z_1,Z_2} = \left\{\frac{s_{12}(1 - \Gamma_1)(1 + \Gamma_2)}{1 + \Gamma_1 s_{11} + \Gamma_2 s_{22} + \Gamma_1\Gamma_2\,\Delta s}\right\}_{Z_{01},Z_{02}} \tag{31}$$

$$\{\Delta s\}_{Z_{01},Z_{02}} = \{s_{11}s_{22} - s_{12}s_{21}\}_{Z_{01},Z_{02}} \tag{32}$$

$$\Gamma_1 = \frac{Z_{01} - Z_1}{Z_{01} + Z_1} \qquad \Gamma_2 = \frac{Z_{02} - Z_2}{Z_{02} - Z_2} \tag{33}$$

where s_{21} and s_{22} are found by transposing subscripts.

These transformations are useful in computing s_{12} of a uniform transmission line normalized with respect to two arbitrary impedances Z_1 and Z_2. The voltage S parameters of a uniform transmission line, normalized with respect to the characteristic impedance of the line, are $s_{11} = s_{22} = 0$ and $s_{12} = s_{21} = e^{-j\theta}$. Then from equation (31),

$$\{s_{12}\}_{Z_1,Z_2} = \frac{e^{-j\theta}(1 - \Gamma_1)(1 + \Gamma_2)}{1 - \Gamma_1\Gamma_2 e^{-j2\theta}}. \tag{34}$$

APPENDIX D

S *Parameters of a Nonideal Transmission Line*

The primary deviation of a physical uniform airline from an ideal airline is caused by the skin effect. The nonideal line is modeled as an ideal line with the added series skin effect resistance of $R_0\omega^{\frac{1}{2}}(1 + j)$ ohms per unit length. The characteristic impedance is

$$Z = Z_0\left[1 + \frac{R_0\omega^{\frac{1}{2}}}{\omega L_0}\right]^{\frac{1}{2}}. \tag{35}$$

The propagation constant is

$$\gamma = -j\omega C_0 Z. \tag{36}$$

The quantities used in the above expressions are defined as:

ω = angular frequency

L_0 = series inductance per unit length of an ideal line

C_0 = shunt capacitance per unit length of an ideal line.

The term $Z_0 = (L_0/C_0)^{\frac{1}{2}}$ = characteristic impedance of an ideal line. The voltage scattering parameters of the nonideal line of length l normalized to two impedances equal to Z are of the simple form

$$s_{11}\,|_{Z,Z} = s_{22}\,|_{Z,Z} = 0 \tag{37}$$

$$\{s_{12}\}_{Z,Z} = \{s_{21}\}_{Z,Z} = e^{\gamma l}. \tag{38}$$

The more useful $s|_{Z_0,Z_0}$ parameters are easily obtained using the transformations of Appendix C. From equation (33) one obtains

$$\Gamma_1 = \Gamma_2 = \frac{Z - Z_0}{Z + Z_0}. \tag{39}$$

For a practical airline $[R_0\omega^{\frac{1}{2}}]/L_0\omega \equiv 2z \ll 1$, allowing for the simplification of expressions (38) and (39) to

$$\Gamma_1 = \Gamma_2 \doteq \frac{z}{2}(1 - j) \tag{40}$$

and

$$\{s_{12}\}_{Z,Z} = \{s_{21}\}_{Z,Z} = e^{-\omega\tau z}e^{-j\omega\tau(1+z)}, \tag{41}$$

$\omega\tau = \omega C_0 Z_0 l$ is the electrical length of the ideal airline. The application of the Appendix C transformations to the $\{s\}_{Z,Z}$ parameters, assuming that the line is electrical short (that is, $z\omega\tau \ll 1$), yields the desired $\{s_{11}\}_{Z_0,Z_0}$ parameters.

For $1/\lambda^2 \gg \omega\tau \gg \lambda^2$

$$\{s_{11}\}_{Z_0,Z_0} = \{s_{22}\}_{Z_0,Z_0} \doteq \lambda(2\omega\tau)^{\frac{1}{2}} \frac{\sin \omega\tau}{\omega\tau} e^{-j\omega\tau + j(\pi/4)} \tag{42}$$

$$\{s_{12}\}_{Z_0,Z_0} = \{s_{21}\}_{Z_0,Z_0} \doteq e^{-\lambda(\omega\tau)^{\frac{1}{2}}}e^{-j\lambda(\omega\tau)^{\frac{1}{2}}}e^{-j\omega\tau} \tag{43}$$

$\lambda = z(\omega\tau)^{\frac{1}{2}}$ is a frequency independent constant of the nonideal line. (See also Ref. 9.)

REFERENCES

1. Geldart, W. J., Haynie, G. D., and Schleich, R. G., "A 50 Hz—250 MHz Computer-Operated Transmission Measuring Set," B.S.T.J., this issue, pp. 1339–1381.
2. Follingstad, Henry G., "Complete Linear Characterization of Transistors from Low Through Very High Frequencies," IRE Transactions on Instrumentation, *I–6,* No. 1 (March 1957), pp. 49–63.
3. Leed, D., and Kummer, O., "A Loss and Phase Set for Measuring Transistor Parameters and Two-Port Networks Between 5 and 250 MC," B.S.T.J., *40,* No. 3 (May 1961), pp. 841–883.
4. Leed, D., "An Insertion Loss, Phase and Delay Measuring Set for Characterizing Transistors and Two-Port Networks Between 0.25 and 4.2 GC," B.S.T.J., *45,* No. 3 (March 1966), pp. 397–440.
5. Kuvokawa, K., "Power Waves and the Scattering Matrix," IEEE Trans. on Microwave Theory and Techniques, *MTT-13,* No. 2 (March 1965), pp. 194–202.
6. Weinberg, Louis, "Fundamentals of Scattering Matrices," Electro-Technology, *80,* No. 1 (July 1967), pp. 55–72.
7. Carlin, H. J. and Giordano, A. B., *Network Theory,* Englewood Cliffs, New Jersey: Prentice-Hall Inc., 1964, pp. 237–255.
8. Evans, J. G., "Linear Two-Port Characterization Independent of Measuring Set Impedance Imperfections," Proc. IEEE, *59,* No. 4 (April 1968), pp. 754–755.
9. Zorzy, J., "Skin-Effect Corrections In Immittance and Scattering Coefficient Standards Employing Precision Air-Dielectric Coaxial Lines," IEEE Trans., on Instrumentation and Measurement, *IM-15,* No. 4 (December 1966), pp. 358–364.

A 50 Hz–250 MHz Computer-Operated Transmission Measuring Set

By W. J. GELDART, G. D. HAYNIE and R. G. SCHLEICH

(Manuscript received November 27, 1968)

A computer-operated transmission measuring set has been developed for the 50 Hz to 250 MHz frequency range. Use of the computer in this system has significantly effected the test set design and the measurements obtainable.

Compared with previously available transmission measuring sets, the computer-operated set increases speed more than 300 : 1. This speed, along with state-of-the-art accuracy and increases in operating range, flexibility, and convenience, enables the set to be used for types and quantities of measurements previously not practical. It has already been applied to laboratory and production testing with resulting improvements in the quality and reliability of manufactured product designs.

In addition to the directly measured quantities of insertion loss and phase, the set provides insertion delay, impedance, and two-port parameters as derived quantities. The two-port data conversion program provides H, Y, Z, G, T, S, ABCD *and* ABCD^{-1} *parameters with a number of useful options. Results of transmission measurements, impedance measurements, and two-port measurements are presented. Some of the error mechanisms and means of measuring them are discussed. Further development of centralized measuring facilities, with the computer operated set as a basic element, is discussed.*

I. INTRODUCTION

The development of communication systems for the Bell System has, in the past, required large numbers of transmission measurements. These measurements have been costly and time-consuming. With the trend toward more complex systems, the volume and accuracy of measurements must be increased. At the same time, the increased use of computers in modeling and design requires flexibility in the types of data obtained. For example, two-port characterization of devices such as transistors is becoming particularly important.

This paper describes a computer-operated transmission measuring set developed for laboratory use in the 50 Hz to 250 MHz frequency range (see Fig. 1). Compared with previously available sets, this set provides increased speed, operating range, accuracy, flexibility, and convenience, and the capacity for types and quantities of measurements previously not practical.

The computer-operated measuring set is a part of the centralized measuring facilities being developed for Bell Laboratories. With the addition of various appliques now being developed, the measurement centers will provide additional measurement facilities in the 50 Hz to 250 MHz range, with environmental control, and in the three microwave radio bands at 4, 6, and 11 GHz. The set is also being used for production testing by the Western Electric Company.

II. MEASUREMENTS SET CHARACTERISTICS

2.1 *General*

The small, general-purpose digital computer with fast and precise digital-analog components and with broadband analog components has had a significant effect on the measurements obtainable.

Increased operating speed results from the use of computer control, memory, and computation as well as fast measuring set components. The operating speed is 10 to 300 times faster than manual test sets (per measurement point), depending on the output media and the test frequency.

Computer control also makes wide operating ranges practical. By automatic control of test set level patterns, insertion loss can be measured over a range from −40 to +100 dB. By automatic control of signal sources and frequency dependent elements, operation over nearly seven decades is obtained. This frequency range previously required three separate test sets.

Accuracy is, of course, limited by accuracy of the test set standards. Computer operation, however, does play a significant part in the test set accuracy. First, test set errors can be comprehensively evaluated to a degree only practical with a high speed set. Second, the speed, memory, and computation capability permits the correction of data for known errors and the averaging of random errors. Correction of "zero line" errors is particularly important with the complex transmission paths used in the test set. Finally, the operating rules

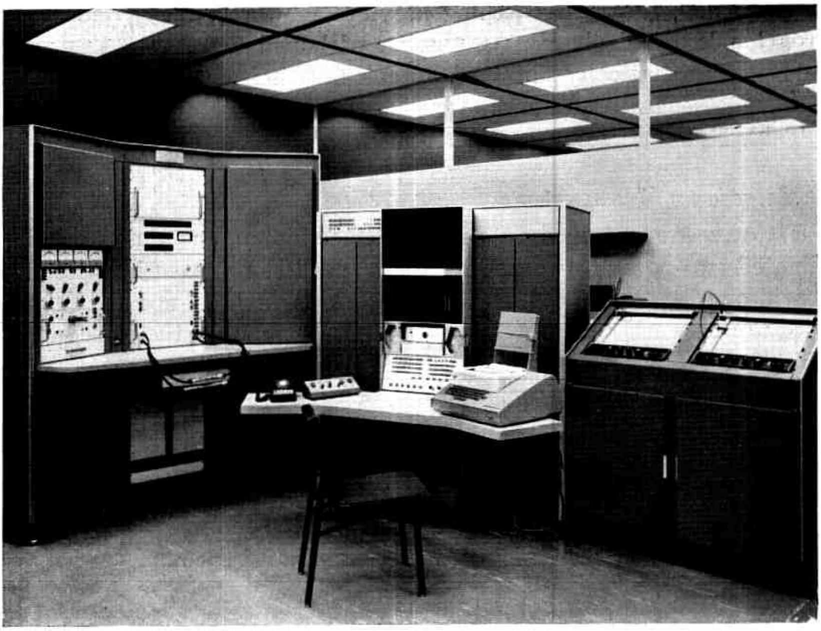


Fig. 1 — Computer-operated transmission measuring set.

are set up so as to maintain near optimum levels during the measurements. In this way, tolerable compression errors can be deliberately introduced to improve signal-to-noise ratio.

The computer-operated set provides flexibility in a number of ways.

(*i*) By using software rather than hardware control, operation of the set can be readily changed.

(*ii*) Special measurements can be made which include system maintenance tests and periodic measurement of test set error sources.

(*iii*) Measured data can be converted into more useful forms. For example, meaningful acceptance criteria, which may be too complex for manual application, can be used for GO, NO GO tests of measured networks.

The computer-operated set is convenient in the ease of setting up measurements, particularly repeated measurements, and in the ability to yield measured or derived parameters on various output media.

2.2 *Transmission Measurements*

The basic quantities measured by the transmission-measuring set are insertion loss (or gain) and insertion phase shift as a function of frequency. Internal switching provides for direct measurement of insertion loss and phase at a number of impedance levels. Insertion delay is automatically obtained by calculation from two phase measurements separated by an appropriate frequency interval.

Table I summarizes test set performance.

2.3 *Impedance Measurements*

Fixtures are provided to permit connection of one port (or terminated two-port) networks into the transmission measuring circuits. By measuring suitable impedance standards (three needed) along with the network and processing data on a computer, the impedance, reflection coefficient, or return loss can be obtained.[1] With a change of program, the test set computer can be used for processing the data. In this case, the derived parameters can be output on the test set output equipment.

Measurement accuracy is a function of the impedance measured but will be better than 1 percent over an impedance range of 10^4. Section 5.2 has a more detailed discussion of impedance measurement accuracy.

2.4 *Two-Port Characterization Measurements*

Coaxial terminals, coaxial fixtures, and dc bias supplies are provided which permit the linear characterization of two-port networks and of devices such as transistors. For unbiased networks, the fre-

Table I — Test Set Performance

Characteristic	Range	Best accuracy
Test frequency	50 Hz to 250 MHz, adjustable to 0.01–0.08 Hz	3 parts in 10^8
Gain measurements	0 to 40 dB	0.001 dB
Loss measurements	0 to 100 dB	0.001 dB*
Phase measurements	0 to 360°	0.01°
Delay measurements†	39.999 ms to 39.999 nsec, full scale	$55/\Delta f$ μsec

Impedance levels: 50, 75, 135 (balanced or unbalanced), 600 (balanced or unbalanced), arbitrary Z, and probe mode.

* Loss and phase accuracies decrease as loss or gain increases. For example, the random error in loss measurement is $0.001\ (1 + 0.2 \times 10^{L/20})$ dB for losses <40 dB and $0.01\ (1 + 0.2 \times 10^{(L-40)/20})$ dB for losses >40 dB.

† Computed from two phase measurements separated by Δf Hz.

quency range is 50 Hz to 250 MHz and for biased networks the range is 50 kHz to 250 MHz. Two-port characterization data, including calibration data, is processed by computer (such as the IBM 7094) to provide any of the standard two-port matrix representations such as H or Y parameters. Accuracy of the output parameters is parameter dependent but errors in S parameters will be less than 0.02 for most unknowns. Section 5.3 has a more detailed discussion of parameter accuracy.

2.5 *Test Set Input-Output*

Figure 2 is a simplified block diagram of the computer operated test set from the human operator-test set interface. Information required by the computer to control a particular set of measurements is contained in the computer memory, in switch positions on the operator control panel, or possibly on punched paper tape. The information in computer memory is inserted either with the tape reader or the typewriter. The operator control panel can be used both to set up and start the desired set of measurements or to modify the sequence after it has started.

Outputs are selected on the operator control panel. Visual readout is always present, and the typewriter, tape punch, and X–Y plotters can be independently selected. Plotting parameters are part of the input data, and output readings are scaled by the computer. Points are plotted to an accuracy of about 0.1 percent. If no output is selected

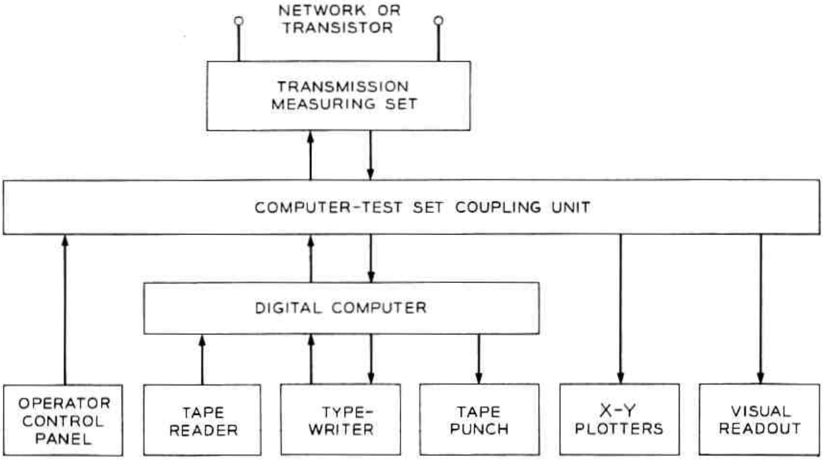


Fig. 2 — Operator-test set interface.

or if the output switches are turned off during a measurement sequence, the test set will provide continuous loss and phase measurements with the test frequency kept constant.

2.6 *Test Set Speed*

Measurement speed varies with a number of factors, particularly the number of measurements averaged to reduce random errors. However, Table II gives a useful summary of typical measurement times.

III. BLOCK DIAGRAM DESCRIPTIONS

3.1 *Overall Description*

Figure 3 is a simplified block diagram of the measuring set and connections to the computer. At this interface, control and readout of the measuring set is entirely digital.

Under computer control, the signal oscillator supplies the test signal to the comparison unit for excitation of the circuit to be measured and to the reference path frequency converter. The local oscillator provides the proper frequency to the measurement path and reference path converters for translating the test frequency to a fixed intermediate frequency. The loss standard is adjusted to have a loss equal to that of the measured circuit using error signals provided by the loss detector. The phase meter provides a measurement of the phase difference between the measurement path and reference path inputs. With suitable switching, the difference between two readings provides the desired phase measurement.

The measuring set block configuration is similar to others previously reported.[2, 3] However, when the elements in the configuration are realized with the components to be described and then controlled by a computer, the advantages cited in Section I result.

3.2 *Gain-Loss Measurements*

Figure 4 shows the loss measuring circuit with details added which are esential to the discussion.

3.2.1 *Comparison Circuit*

The comparison circuit rapidly interchanges the unknown path and standard path between the signal source and the heterodyne detector. This produces amplitude and phase variations in the signal at the switching rate which correspond to differences in transmission between the unknown and standard paths. The unknown path con-

TABLE II — TYPICAL MEASUREMENT TIMES

Type of output	Time per point (seconds)	
	(Freq. <2 kHz)	(Freq. >2 kHz)
Magnetic core	2	0.2
$X - Y$ plotters	2	0.2
Paper tape	3	1
Typewriter	9	7

tains either the switching unit for two port characterization measurements or a path connecting the network for insertion loss and phase measurements. Additional switches, not shown, provide connections for the various impedance levels.

The standard path contains either a low loss transmission line or a 40 dB pad. The 40 pad is always inserted in the standard path when the unknown path is being measured. When the standard path is being measured, the transmission line is inserted for losses less than 40 dB. For losses equal or greater than 40 dB, the 40 dB pad is inserted in the standard path and a 40 dB preamplifier is inserted in the level adjust unit.

Use of the 40 dB pad and preamplifier in this way has several

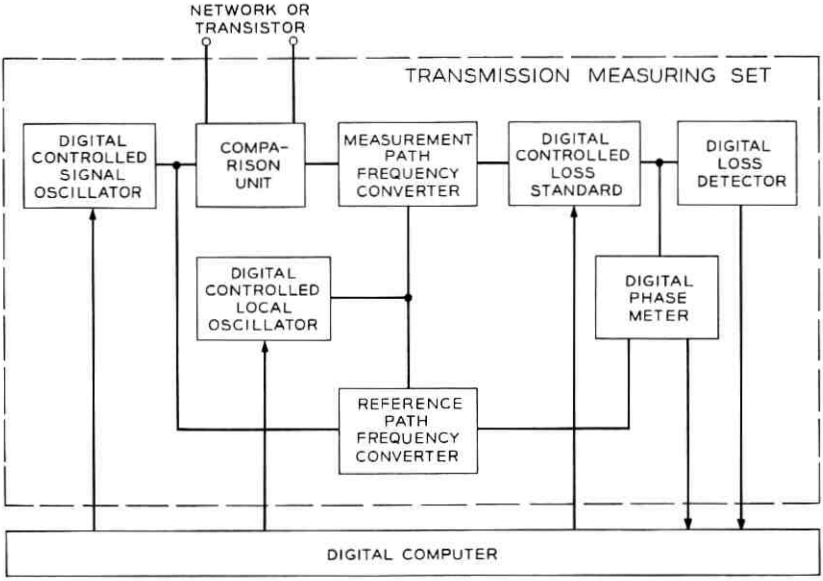

NETWORK OR TRANSISTOR
TRANSMISSION MEASURING SET
DIGITAL CONTROLLED SIGNAL OSCILLATOR
COMPARISON UNIT
MEASUREMENT PATH FREQUENCY CONVERTER
DIGITAL CONTROLLED LOSS STANDARD
DIGITAL LOSS DETECTOR
DIGITAL CONTROLLED LOCAL OSCILLATOR
DIGITAL PHASE METER
REFERENCE PATH FREQUENCY CONVERTER
DIGITAL COMPUTER

Fig. 3 — Simplified block diagram of measuring set.

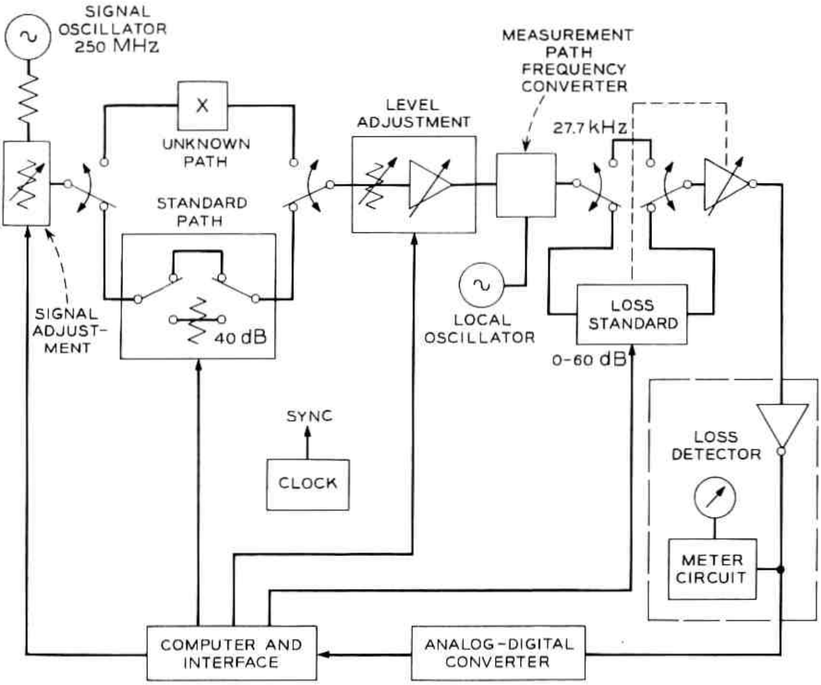


Fig. 4 — Loss measurement.

advantages: the crosstalk requirement in the comparison switches is reduced by 40 dB and the required dynamic range of the measurement path frequency converter is reduced by 40 dB. Use of the preamplifier also improves the system noise figure by about 20 dB.

Since the 40 dB pad has a significant frequency characteristic, some method of correction is necessary. This is conveniently accomplished by initiating an additional comparison cycle whenever the 40 dB pad is used. During this cycle, the two elements in the standard path are compared and the difference is measured on the IF loss standard. The 40 dB pad is, therefore, a transfer standard.

3.2.2 *Signal and System Level Adjustments*

The attenuators following the signal oscillator adjust the signal level into the measurement circuit. The attenuators are switched by the computer under manual or program control. In either case, if the unknown has enough gain to overload the test set detector, the signal level is automatically reduced and the signal level actually used is typed out.

The level-adjust circuit adjusts the level into the measurement path converter to minimize certain errors. When measured loss or gain is low, the level into the converter is maintained for best signal-to-noise ratio consistent with 0.001 dB linearity. At high loss, the level into the converter (not yet programmed) is increased to improve the signal-to-noise ratio without producing significant linearity errors.

3.2.3 *IF Loss Standard and Detector*

Amplitude and phase variations produced in the test signal by the comparison unit are linearly translated to the final IF of 27.777 kHz. Here the loss standard is switched in synchronism. For loss, the IF loss standard and the unknown are switched out of phase and for gain they are switched in phase.

The loss standard contains relay switched, precision attenuators ranging from 0 to 59.99 dB in 0.01 dB steps. Complementary gain is provided in the common output path so that the output is constant (±1 dB) when the standard is correctly balanced. Hence the loss detector and phase detector are operated at nearly constant levels.

The amplifier-detector in the loss detector has a logarithmic characteristic which provides a dc output proportional to the input amplitude in dB. During the balance sequence, a measure of the difference in loss between the loss standard and the measured network (loss unbalance) is obtained from the difference of two readings with the analog-digital converter. After the loss balance is completed, the analog-digital converter readings provide the 0.001 dB decade indication to the computer for automatic readout.

In order for the analog-digital converter to have ±20 dB balancing range and yet provide the ±0.001 dB indication, the equivalent of 16 bits is required. This was achieved with a 13 bit converter and a switched (5X) preamplifier. For loss standard balancing, the amplifier is out and the least significant bit is 0.005 dB. When a balance is achieved, the amplifier is switched in and the least significant bit is 0.001 dB.

3.3 *Phase Measurements*

Figure 5 is a block diagram of the phase measuring circuit. The phase changes produced by comparison switching are the changes between the unknown and standard paths at test frequency and between the loss standard and "strap" at intermediate frequency. The net change in phase is the change between the unknown and standard paths since

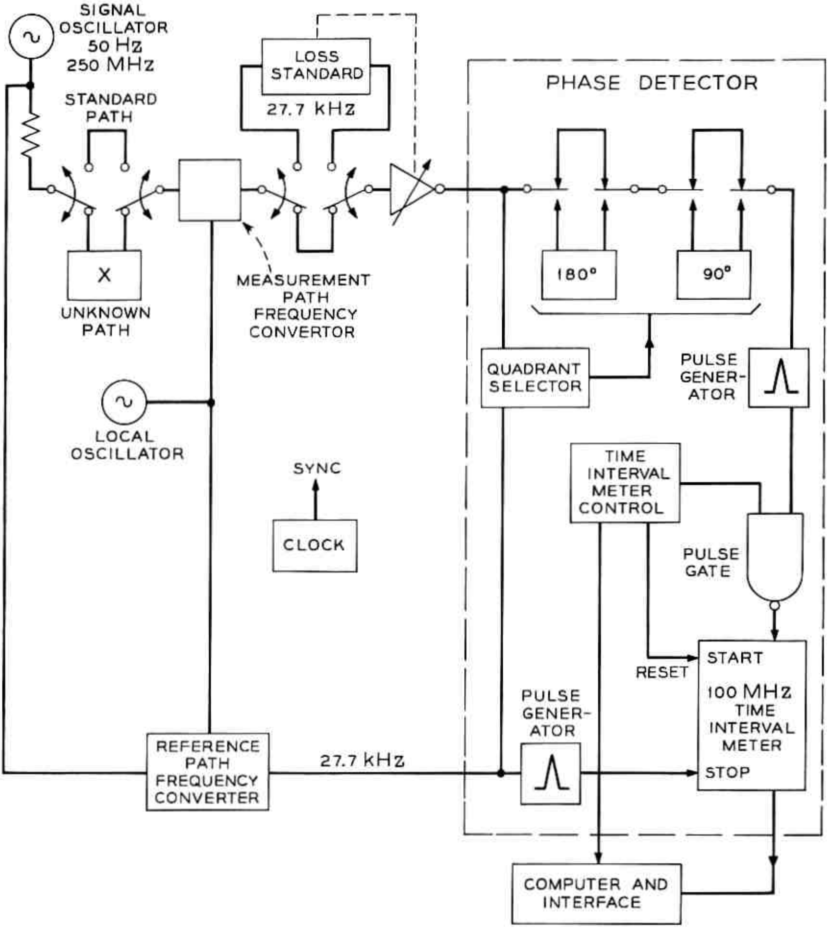


Fig. 5 — Phase measurement.

the two paths at intermediate frequency are adjusted to have phase differences less than ±0.002°. The reference path signal provides a constant phase reference during the switching cycle.

The phase measuring technique used cannot measure phase at exactly 0°. Phase equalization of the measurement path and the reference path over the 250 MHz frequency range is not possible with fixed networks. The quadrant select circuit together with the 180° and 90° networks select a quadrant such that the relative phase into the phase measuring circuit is within 180° ± 135° for both positions of the comparison switch. Operation of the quadrant select circuit does not add to the measurement time since it operates during an early portion of the switching cycle before transients have settled to the point where precision phase measurements can begin.

The pulse generators produce pulses on the positive zero crossing of the IF signals. The pulse from the measurement path generator starts the time interval meter measurement and the reference path pulse generator stops the time interval measurement. The time interval is measured by counting the number of pulses from a 100 MHz source that occur between the start and stop pulses. In a single period the time interval meter can resolve. 0.1°

$$\left(\text{that is, } \frac{27.777 \text{ kHz}}{100 \text{ MHz}} = \frac{1}{3600} \text{ period}\right).$$

The opportunity for increasing the resolution exists in the system. (See Ref. 3 and the Appendix.) By locking the 100 MHz source to the 1 MHz crystal source in the frequency synthesizer, by a proper choice of the intermediate frequency, and by taking 100 readings, resolution can be increased by a factor of up to 100. Because of compromises in this system, resolution was increased by a factor of about 20, to 0.005°.

The time interval meter control provides gating and reset signals so that 100 readings can be taken and provides timing signals to the computer. Level dependence of the phase meter is not a problem since the loss balance made before phase measurements is within 0.01 dB and the level-to-phase conversion of the pulse generators is less than 0.03° per dB.

3.4 *Signal Frequency Generation and Conversion*

Automatic control of signal sources and frequency converters provides operation over nearly seven decades. Fig. 6 is a simplified block diagram of the elements, including switches, for signal generation and conversion.

3.4.1 *Signal Oscillator and Local Oscillator*

The signal oscillator provides a sinusoidal (harmonics less than 40 dB) test signal to the measurement and reference paths, and the local oscillator provides the signal to tune the heterodyne detector. Each oscillator is composed of a frequency synthesizer and a frequency multiplier (including filters) which together produce output frequencies from 50 Hz to 250 MHz in response to digital signals. Be-

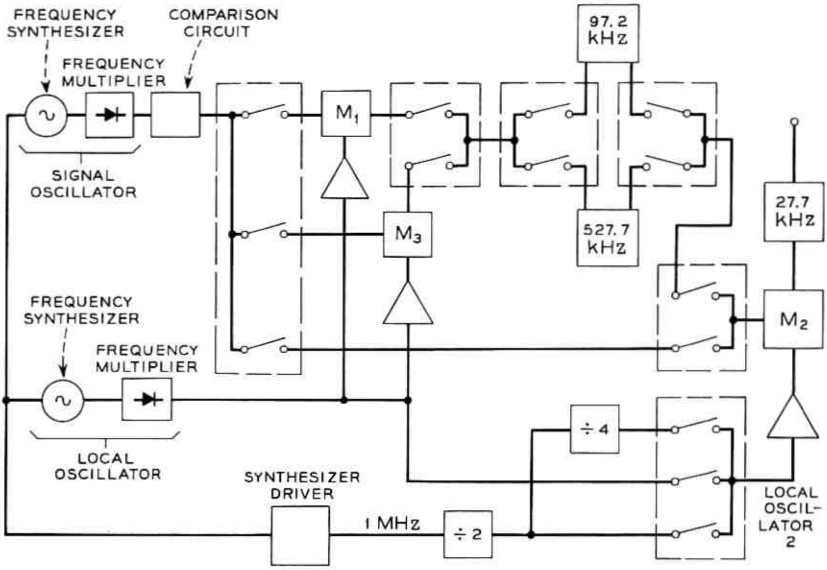


Fig. 6 — Frequency generation and conversion.

Signal frequency range	Frequency multiplier	Modulators used	Local oscillator frequency synthesizer offset (kHz)	Local oscillator 2 input (kHz)
0.05–2 kHz	X1	M1, M2	97.223	125
2–100 kHz	X1	M1, M2	527.777	500
0.1–5 MHz	X1	M2	27.777	Local oscillator
5–50 MHz	X1	M3, M2	527.777	500
50–100 MHz	X2	M3, M2	527.777/2	500
100–200 MHz	X4	M3, M2	527.777/4	500
200–250 MHz	X8	M3, M2	527.777/8	500

low 50 MHz, frequencies are set to a precision of 0.01 Hz within 1 ms after signalling. Above 50 MHz, frequencies are set to a precision ranging from 0.02 to 0.08 Hz within 10 ms. Absolute accuracies of the output frequency, frequency changes of either oscillator, and frequency differences between the two oscillators are all within 3 parts in 10^8. The accuracy of these oscillators eliminate test frequency uncertainty as a source of measurement error and permit the use of very narrow detection bandwidths to reduce errors caused by noise. In the test set operation the oscillators are tuned while the previous data point is being read out so the 10 ms tuning period is negligible.

3.4.2 *Frequency Converters*

The frequency converters must provide linear translation of the 50 Hz to 250 MHz test signal to a fixed intermediate frequency of 27.777 kHz and maintain a satisfactory signal-to-interference ratio at the output.

Since level differences produced by the comparison unit (up to 60 dB) are transmitted through the converters, all elements in the converters including amplifiers and filters must be linear within 0.001 dB. To measure 20 dB loss to ±0.001 dB, the converter noise must be 100 dB and spurious products 80 dB below the maximum linear output level.

In order to achieve this performance, four bands were used. When two stages of conversion are used, the second local oscillator frequency is derived from the 1 MHz standard in the frequency synthesizer. This provides a final IF accurate to 3 parts in 10^8 and a precision within ±0.04 Hz, the precision required for the phase measuring circuit.

3.5 *Impedance Measurements*

Impedance and return loss measurements on one port (or terminated two port) networks can be obtained by making the appropriate connections, making the required sequence of measurements, and processing the measurement data. The frequency range for these measurements is from 50 Hz to 250 MHz.

The three practical connections used for impedance measurements are shown in Fig. 7. The connections are implemented with such networks as a coaxial tee, connector boxes furnished with the test set, or other means such as hybrid networks.

The measured transmission obtained when the network is connected has been shown to be a bilinear function of the network impedance.[4] If four measurements are made, one with the measured network and the other three with "known" impedance standards, the impedance of the network can be determined in terms of the impedance standards.

The equations relating the measurements and the standards are:

$$T_x = \frac{\text{Detector voltage, switches in upper path}}{\text{Detector voltage, switches in lower path}}; \quad Z_x \text{ connected}$$

$$T_x = \frac{T_\infty Z_x + T_0 Z_i}{Z_x + Z_i} = \frac{T_\infty + T_0}{2} + \frac{T_\infty - T_0}{2} \rho_x$$

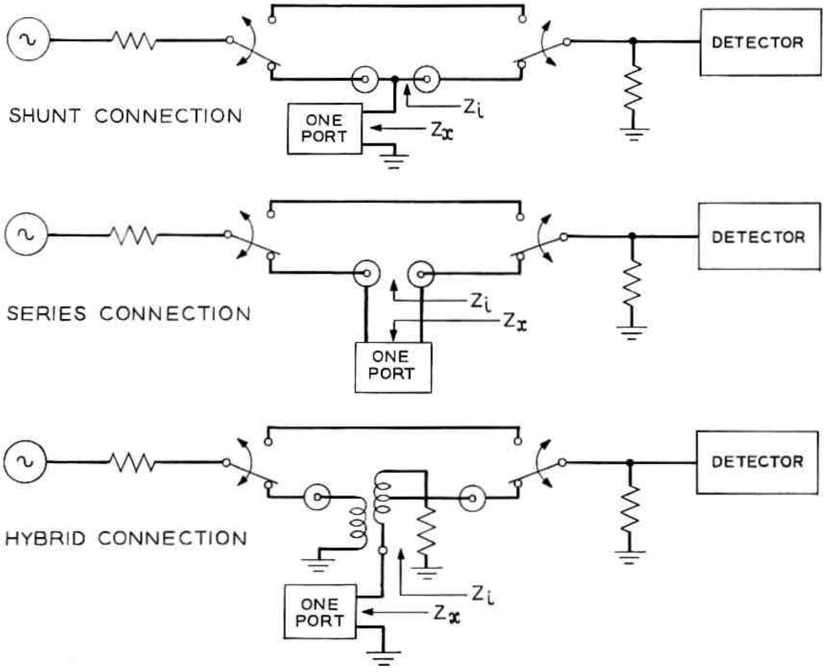


Fig. 7 — Impedance measurements.

Where:

$$T_\infty = T_x \Big|_{Z_x \to \infty}, \qquad T_0 = T_x \Big|_{Z_x = 0}, \qquad \rho_x = \frac{Z_x - Z_i}{Z_x + Z_i}.$$

If an impedance standard Z_s is used and $T_s = T_x |_{Z_x = Z_s}$

$$Z_x = Z_s \frac{(T_0 - T_x)(T_s - T_\infty)}{(T_x - T_\infty)(T_0 - T_s)}$$

and

$$\rho = \frac{Z_x - Z_s}{Z_x + Z_s} = \frac{(T_x - T_s)(T_\infty - T_0)}{(T_\infty - T_s)(T_x - T_0) + (T_s - T_0)(T_\infty - T_x)}$$

Accuracy of the method depends on the accuracy of the standards and on the accuracy of the individual transmission measurements. As indicated in these equations, the expressions for impedance and reflection coefficient involves differences in measurements, and the error in the computed result is a function of impedance as well.

Processing of the impedance measurement data can be done on the system computer (PDP-7) by replacing the operating program with one of the available conversion programs. These programs permit computation of real and imaginary components of impedance and the reactance for a series or parallel equivalent (two element) circuits. Angle and magnitude of reflection coefficient with respect to an arbitrary reference impedance can also be computed. Output from the program can either be on tele-typewriter or X-Y plotters.

Some impedance measurement results are discussed in Section 6.2 and illustrated in Figs. 21 through 23.

3.6 *Two-Port Measurements*

3.6.1 *General*

Linear characterization of two-port networks can be obtained by connecting the device (for example, the transistor) to the appropriate jig or coaxial cable, making the required sequence of measurements, and processing the measurement data. Devices not requiring bias can be measured from 50 Hz to 250 MHz and devices requiring bias can be measured from 50 kHz to 250 MHz. The bias can be voltage-regulated from 0 to 150 volts or current regulated from 0 to 1 ampere. Transistor case temperatures can be controlled from 0° to 95° C for dissipation up to 10 watts.

To the extent that the imperfections in the calibration standards and the capacitance added by the temperature control unit are accurately known, it is possible to reduce the measurement data to device parameters which are independent of the measurement environment.[5, 6] The data reduction program provides H, Y, Z, G, T, S, ABCD and $(\text{ABCD})^{-1}$ parameters with any one of the three device terminals grounded. The terminal grounded in the output parameters is not necessarily the same terminal which was grounded during the measurements.

3.6.2 *Measurement and Calibration Techniques*

The technique which is used to characterize two-port networks is obtained from four transmission measurements which are a combination of two ordinary transmission measurements and two impedance measurements (shunt connection) as described in Section 3.5. Figure 8 shows the four connections needed for the measurements of forward and reverse gain and of forward and reverse bridging loss.

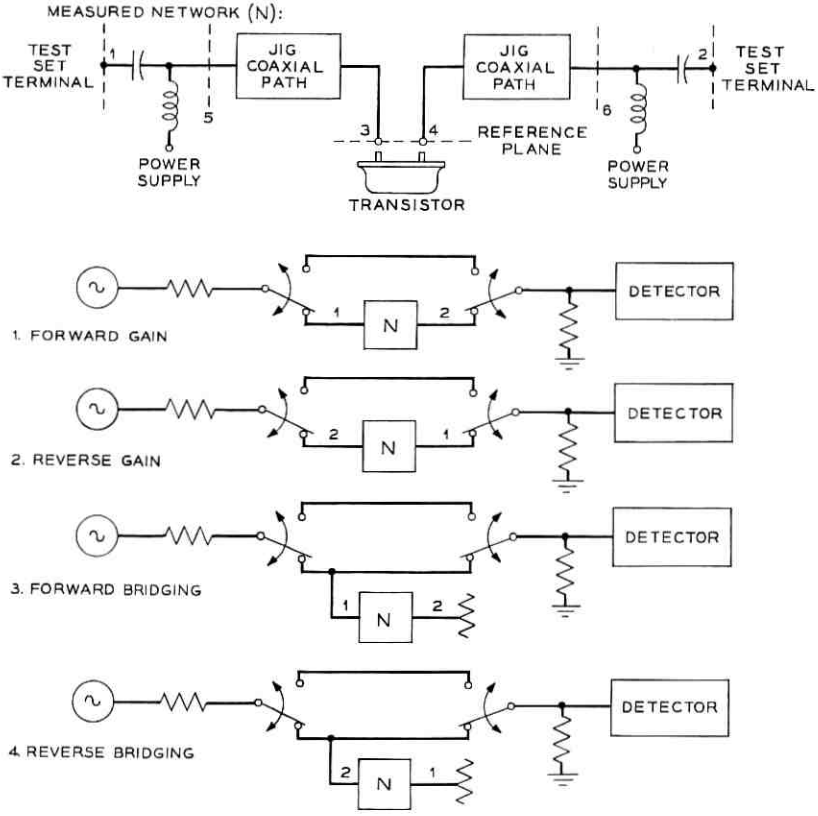


Fig. 8 — Transistor measurement technique.

In automating this sequence of four measurements, the constraint was applied that dc bias paths through the measured device must not be interrupted during the sequence. This means that the minimum network which can be measured when characterizing a transistor is that shown between ports 1 and 2. Even with an accurate test set, accurate device measurements can be made only if the complete network between the test set terminals is characterized or if the reference plane (ports 3 and 4) can be characterized.

Each approach has been used in a previous implementation. In the first case, available 7 mm coaxial standards were used to calibrate ports 5 and 6 (Fig. 8).[1] Then four jigs containing an open circuit, short circuit, reference impedance, and the measured device were successively plugged in to obtain the necessary measurements. The disadvantage of this technique is that symmetry in the jigs must be assumed and that the device bias must be set four times.

In the second case, 21 mm coaxial standards are used to calibrate ports 1 and 2 (Fig. 8).[7] Other coaxial standards (0.054-inch bore) are used to characterize the network between ports 1 and 3, and 4 and 2 at specified frequencies. When the device is inserted, the total network between ports 1 and 2 is measured and the device characteristics are extracted by calculation. The disadvantages of this technique are that calibration data is only available at the specified frequencies and that the network must be manually connected for each of the four measurements.

3.6.3 *Implementation of Automatic Two Port Measurements*

In order to automatically switch the network into the four measurement configurations, considerable coaxial relay switching was required. The return losses that these relays present are low enough so that, when "seen" from the measurement ports 3 and 4, appreciable errors will occur if corrections are not made. It is also desirable to be able to measure at any frequency in the range thus ruling out calibration at a fixed number of points.

It was decided to fabricate "small bore" standards to calibrate ports 3 and 4. Short circuit and 50 ohm standards were developed and the open circuit "standard" is obtained by open circuiting the ports and correcting for fringing capacitance. A 50 ohm "strap" was developed to give "zero line" measurements for the forward and reverse gain measurements. The "strap" also is used to measure the impedance into port 3 with port 4 and vice-versa. Table III lists the measurement and calibration sequence for characterization of a device.

The same technique is used for coaxial unknowns (for example, 14 mm connectors). Open, short, and 50 ohm standards are commercially available and the strap measurement is simply made by connecting the cables from the two terminals together.

3.6.4 *Data Reduction*

Reduction of the data obtained in the measurement sequence just described is a sizable data reduction problem. At present, measurement data is processed on the IBM 7094 computer.

The data reduction program first computes scattering parameters referred to the physical test set impedances. The next step is to trans-

TABLE III — TRANSISTOR MEASUREMENT AND CALIBRATION SEQUENCE

REQUIRED ELEMENTS

Transistor sample and jig
"Small bore" coaxial standards:
  50 ohm termination
  Strap
  Short
  Open

MEASUREMENT PROCEDURE

Test set calibration

Insert	Measure	Number of measurements
Strap	FWD & REV gain, FWD & REV bridging	4
50 ohm in port 1 50 ohm in port 2 Short in port 1 Short in port 2 Open in port 1 Open in port 2	Bridging*	2 2 2 2 2 2
Transistor measurement		
Insert transistor, set bias	FWD & REV gain FWD & REV bridging*	6

* Bridging measurements made with and without quarter wave networks inserted.

form to scattering parameters referred to 50 ohms. This sequence is considerably simpler than initially calculating 50 ohm scattering parameters. The 50 ohm S-parameters can then be transformed into H, Y, Z, G, T, S, ABCD or (ABCD)$^{-1}$ parameters.

A number of useful options are also included. The measurement plane for a measured device can be translated through an arbitrary length of 50 ohm transmission line. Capacitance in shunt with any pair of terminals (for example, the transistor temperature control unit adds about 3 pF from case to ground when used) can be removed by computation and either of the noncommon terminals in the measurement can be made the common terminal in the output data.

The derived parameters can be output in cartesian or polar coordinates or in magnitude (dB) and angle. The output data forms include both tabulated and plotted data. Some two-port measurement results are discussed in Section 6.3 and illustrated in Figs. 24 and 25.

IV. OVERALL CONTROL DESCRIPTION

4.1 *Operating Description*

As Fig. 2 indicates, the primary means of input are the operator control panel (mode selector), the tape reader, and the typewriter. Inputs are also made from the computer console when loading tapes and in special instances such as maintenance tests. Output is obtained from the visual readout, X-Y plotters, tape punch, and typewriter.

In some of the operating modes, information previously entered via the typewriter or tape reader is stored in computer memory and can be used to make or repeat a measurement sequence. In other modes, the information is not stored and must be reinserted for each measurement sequence.

4.1.1 *Operator Control Panel*

The operator control panel, or mode selector, is shown on Fig. 9. Switch positions provide for selecting the desired measurement and readouts. The panel also provides for operator interaction during a measurement.

The frequency selection switch selects the method by which test

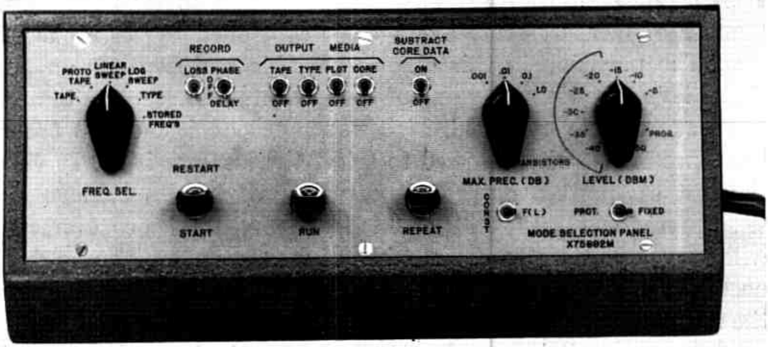


Fig. 9 — Operator control panel.

frequencies are chosen. The frequencies may be a sequence read from punched tape, a linear or logarithmic sequence generated by the computer, a list of frequencies internally stored, or single values entered on the typewriter.

The RECORD switches are used to select the parameters to be measured. This provides independent selection of loss-gain measurements and phase or delay measurements. In the usual case the measured values are displayed on the output media. In the PROTOTAPE mode, the displayed values are the measured values minus the values stored in the paper tape. This, of course, provides the means to compare a measured network with either a previously measured physical prototype or a "mathematical" prototype.

Visual readout is always provided. The OUTPUT MEDIA switches are used to select other desired readouts. The TAPE, TYPE, and PLOT switches provide output on the punched tape, typewriter, and X-Y plotters, respectively. The CORE switch provides output to the computer memory. On any measurement run, the contents of the core can be subtracted from the run being made by operating the SUBTRACT CORE DATA switch. This is a very useful option. Some of its uses are illustrated in Sections 5 and 6 on measurement results and on measurement of test set errors. If none of the output media switches are operated, a state is produced where continuous measurements are made at the prevailing test frequency.

Precision of the output data is controlled by two switches. The switch labeled CONST-F(L) determines whether the measurement precision is to be variable or constant. In the F(L) position, the output loss and phase data are the average of from 1 to 1024 readings. The number of measurements averaged is controlled by an input parameter. The precision in this case is variable and depends on the test signal level, the loss (or gain) of the measured device, and the number of measurements averaged. If CONST precision is desired, the maximum precision switch is used to select the loss precision. In this case the test set must average enough measurements to achieve the desired precision (1, 0.1, 0.01, or 0.001 dB) and total measurement time will vary according to the number of measurements averaged. If the necessary averaging exceeds the allowed limit of 1024 measurements, the estimated precision will be typed out.

The LEVEL switch provides manual control of the signal level into the device being tested. A programmed position is also provided to permit program control of signal levels.

The START-RESTART, RUN, and REPEAT switches provide control over the program. The switches permit the measurement run to be stopped, to be resumed, to return to the beginning with or without new parameter entry, and to repeat the measurement run one or more times.

4.1.2 *Parameter Insertion*

The parameters required for a measurement depend on the settings of the mode selector switches. For example, if a linear frequency sweep is selected, the minimum frequency, frequency increment, and maximum frequency must be stored in the computer. If plotting is to be done, scaling parameters must be stored.

The initial input of a set of parameters is usually made with the typewriter. If a parameter change is being considered, the parameter code is typed in and the current value is automatically typed out. The parameter is either changed by typing in a new value or retained by typing the appropriate symbol. Typing the list request code will cause the entire parameter list and current values to be typed out.

A parameter list can be retained for later use by typing in a "dump" code. The list is then stored on punched paper tape. When the measurement is repeated, the list can be read in on the paper tape reader by typing in a "read tape" code. This appreciably reduces the measurement set-up time.

4.2 *Program Description*

The stored program in the computer provides the necessary control for data input, processing, and output; for operation of the transmission measuring set; and for dialogue with the operator. The dialogue occurs when parameters are being entered, when networks are being connected for two-port measurements, and when trouble indications occur.

4.2.1 *Data Input, Processing, and Output*

Figure 10 is a block diagram showing the input-output options available for data used in the operation of the set. Measurement parameters are entered with the typewriter or tape reader. In a later stage of development, measurement parameters will also be entered from punched cards or magnetic tape.*

The test frequencies are generated by the program from stored measurement parameters, stored test frequencies, or from specific frequencies entered by typewriter, card reader, or tape reader.

* Dashed lines indicate not yet operational.

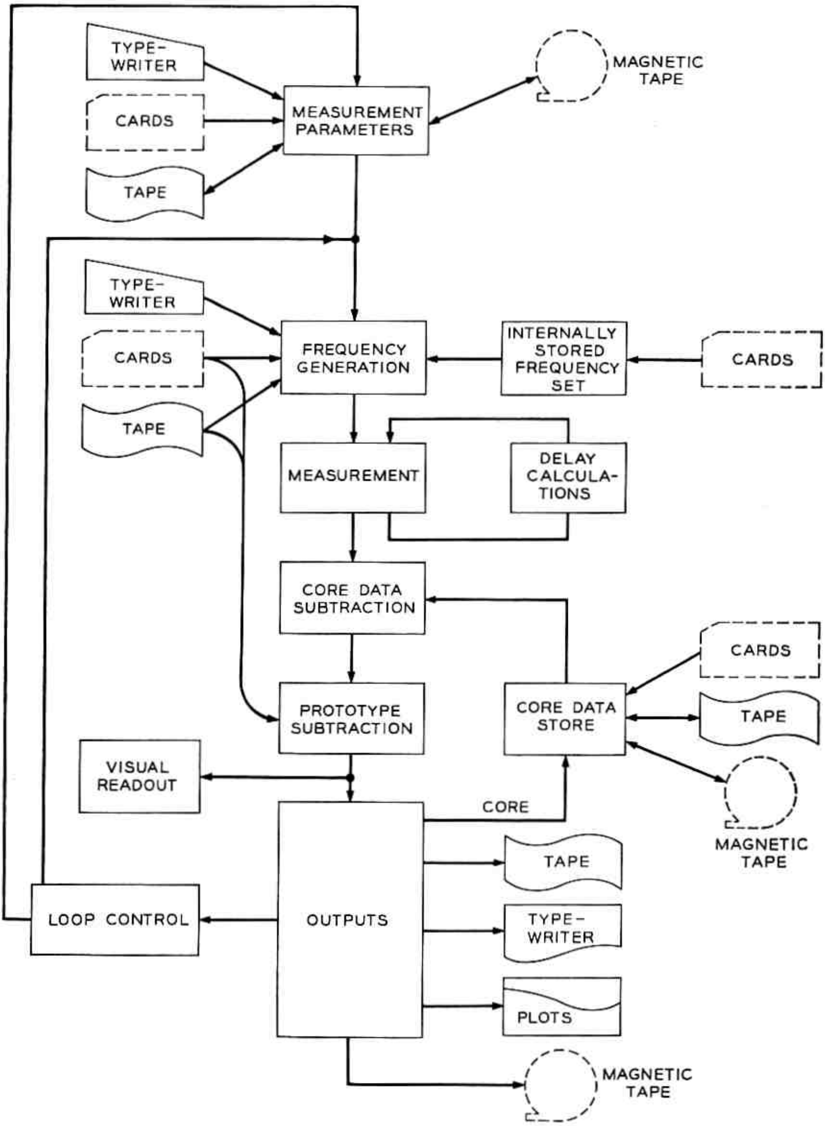


Fig. 10 — Input-output options.

After each measurement, which may include delay calculations, core data is subtracted if indicated by the mode selector. Prototype values on tape or cards can also be subtracted before data is output.

Output is always obtained visually and can also be obtained in core and on paper tape, typewriter, X-Y plotters, and magnetic tape. After output, the loop control returns the program to obtain new parameters or set a new test frequency.

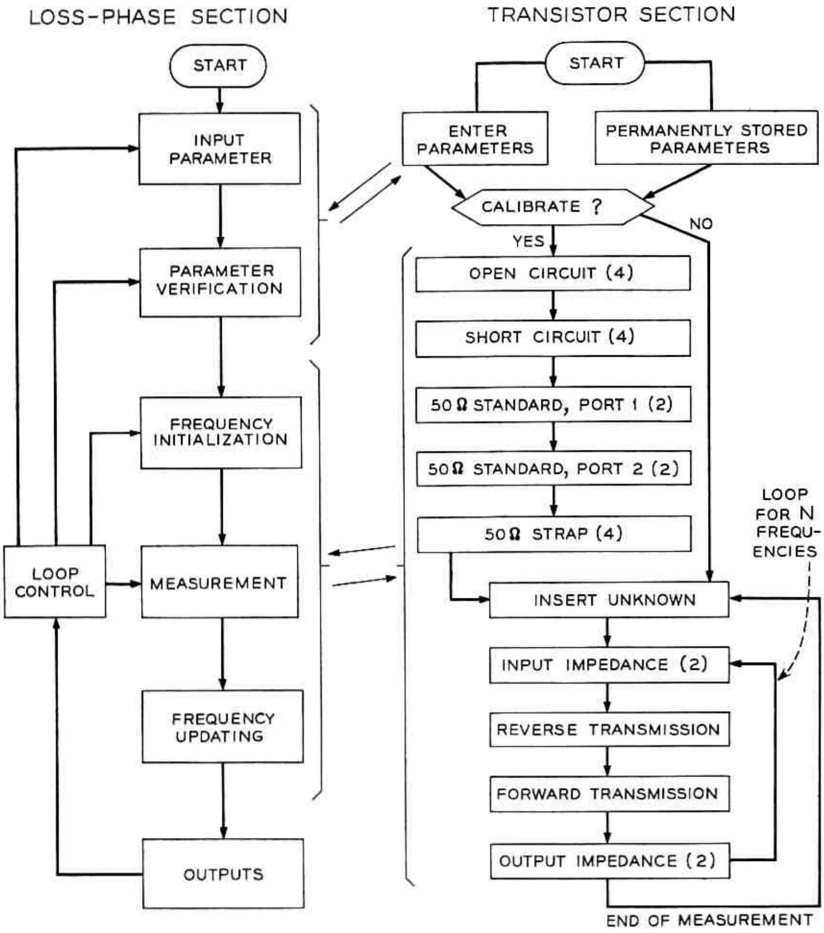


Fig. 11 — Control program.

4.2.2 *Control Program for Test Set Operation*

Figure 11 is a flow chart of the test set control program. Selection between the loss-phase section and the transistor section is made by choice of program starting address. The loss-phase section provides the subroutines for parameter entry, measurement, and data output. Notice that the frequency updating (change to a new frequency) occurs before data output. This gives a maximum time for the measured circuit to reach steady state before the next measurement.

The transistor section of the program provides the necessary control of switching, generates instructions for the operator, and provides the necessary connections to the subroutines in the loss-phase section. After entering the transistor section, there is the option of using permanently stored parameters or manually entering parameters. The next option is that of calibrating or not. Calibration is required if maximum accuracy is needed but adds 16 measurements at each frequency. With either option, the typewriter types instructions to indicate the standard or unknown network to be inserted and the particular connection to be made. The computer also makes the necessary connections in the measuring circuit. When the unknown is connected, six measurements are made (only four independent) and the four which provide the most accurate data are saved. Data is normally output on punched tape for subsequent processing to produce corrected, two-port parameters.

V. MEASUREMENT ERRORS

Measurement errors are evaluated in two complementary ways. The first is to directly evaluate each error-producing element in the system. The second is to measure networks with predictable properties. If the second group of measurements gives satisfactory results, confidence is gained that all of the important elements were evaluated in the first group.

5.1 *Test Set Errors*

The error sources in the test set that must be considered are: spurious signals, amplitude compression, mistermination errors, errors in standards, errors in detectors, circuit drifts, switching transients, and quantization errors. In order to measure these error sources efficiently and to approach the ideal of complete self-testing as closely as possible, measurements are made automatically on the test set wherever practical.

One example of automatic measurement of error sources is in the measurement of spurious signals. Any spurious signals present at the inputs of the loss detector or phase detector and some types of spurious signals at the inputs to nonlinear system elements can cause measurement error. Spurious signals considered are crosstalk, noise, modulation products, power frequency pickup, and test signal harmonics.

By the introduction of several auxiliary circuit elements in the test set transmission paths (Fig. 12), spurious products such as crosstalk, noise, 60 Hz harmonic pickup, and IF carrier leak as small as 100 dB down can be automatically measured and plotted as a function of frequency and level configuration. With the connections, A–A′, the converter and IF circuits are evaluated; with the connections B–B′, the RF circuits are evaluated.

Figure 13 plots the detector signal to interference ratio for a particular system configuration. For the range where the ratio is above 80 dB, random and systematic errors for low losses will be less than 0.001 dB and 0.01° (neglecting other error causes). Techniques have also been devised to automatically measure the other listed error sources with the test set.

5.2 *Impedance Measuring Errors*

Accuracy of the impedance measurements derived from transmission methods (Fig. 7) depends upon test set accuracy and accuracy of the impedance standards.

An error is made by the test set when each of the four transmission

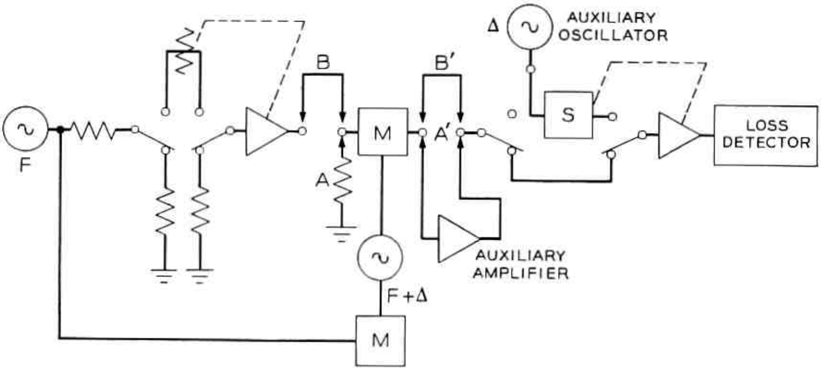


Fig. 12 — Measurement of spurious products.

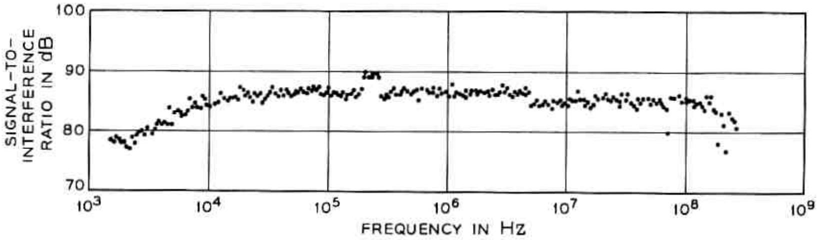


Fig. 13 — Test set signal to interference ratio.

measurements (that is, unknown and three standards) is made. These errors depend on the impedance of the unknown and are caused primarily by test set noise and finite measurement precision. The expression for the error in the calculated reflection coefficient resulting from test set errors is given in ref. 6. For the case of measuring a small reflection with a coaxial tee, the error, $\Delta\rho$, is typically less than 0.004.

The unknown is defined in terms of the standards used. Imperfections in the standards therefore cause errors unless the standards are adequately characterized. At high frequencies, characterization of the standards includes making a precise definition of the measurement.

When appropriate coaxial connectors are used to connect coaxial unknowns and standards, the reference plane and the measurement are well defined. In the case of unknowns with pig tail leads, the measurement is not well defined unless the unknown and the standards have precisely the same geometry so that field patterns are identical in the four measurements. If the geometries are not uniform, the unknown and standards can be placed in a shield but then the resulting parasitics must be estimated.

5.3 *Two-Port Measurement Errors*

Errors in determining two-port parameters include errors in measuring impedance as well as errors in measuring transmission. In the data reduction program used, two-port measurement data is always reduced first to s parameters (50 ohm reference). The first order error expressions for the s parameters are given in Ref. 6.

A worst case error for the s parameters of a precision air line would be:

$$|\,{}^{\Delta s^{11}}(50,\,50)\,| = |\,{}^{\Delta s^{22}}(50,\,50)\,| \leqq 0.004$$

$$\left|\frac{{}^{\Delta s^{21}}(50,\,50)}{s^{21}(50,\,50)}\right| = \left|\frac{{}^{\Delta s^{12}}(50,\,50)}{s^{12}(50,\,50)}\right| \leqq 0.002.$$

Estimation of errors in the s parameters is tedious, even when using a worst case estimate. The most promising approach advanced so far is to calculate errors by modifying the data reduction program. In the modified program each measured value obtained in the measurement sequence would be perturbed a small amount to obtain the parameter sensitivities to the measurement. Then the known systematic errors and the calculated random error (a standard test set output) would be applied to the sensitivity factors already obtained. Proper summing of the error terms would then give a good estimate of the parameter error expectation.

Two-port measurement data obtained so far indicates errors that are perhaps one fourth of the worst case errors indicated by the formulas just referred to.

VI. MEASUREMENT RESULTS

As indicated previously, the basic quantities measured by the transmission measuring set are insertion loss (or gain) and insertion phase shift as a function of frequency. Insertion delay, impedance, and two-port parameters are quantities derived from transmission measurements.

6.1 *Transmission Measurements*

Ideally, transmission measurement accuracy would be confirmed with measurement standards whose properties were precisely known. No complete set of standards exist, but some networks are available where knowledge of their behavior gives reliable indications of the test set precision and accuracy.

6.1.1 *Precision Coaxial Attenuators*

The uniform transmission properties of precision coaxial attenuators over the low frequency range makes possible meaningful comparisons between dc and ac measurements. A 6 dB and 14 dB attenuator were measured individually and in tandem on the test set and compared with measurements made on a dc ratio bridge capable of 0.0001 dB accuracy.

Results of the measurements are given on Figs. 14 through 17. Figure 14 shows the discrepancy between two measurements of the same 6 dB attenuator taken 15 minutes apart. Figures 15 and 16 give the data for the individual measurements on the 6 dB and 14 dB attenuators. In midband, discrepancies between ac and dc loss values are less than those which occur with a 2°F change in ambient temperature. Midband

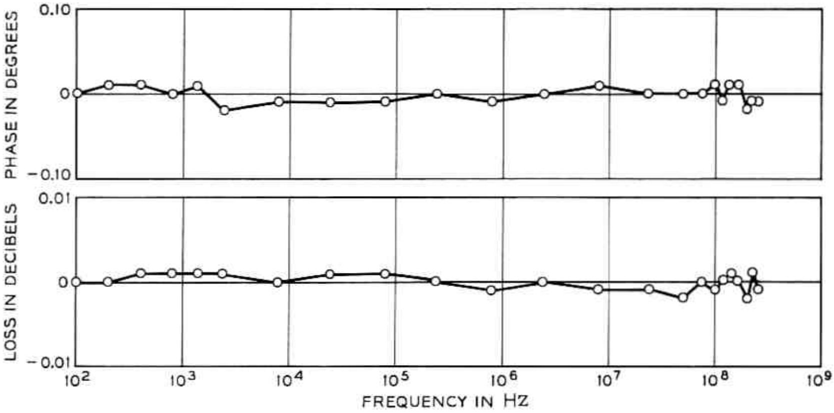


Fig. 14 — Repeatability of 6 dB pad measurements.

phase shift of the attenuators is within ±0.01° of zero. Figure 17 shows the discrepancy between the sum of the 14 dB and 6 dB measurements and the measurement of the two attenuators in tandem.

Errors are seen to increase at the low frequency end of the test set range. The error magnitudes correspond to signal-to-interference ratios measured over the same range by the method described in Section 5.1.2. The deviations at the high frequency end in Figs. 15 and 16 result primarily from changes in the insertion loss and phase of the attenuators.

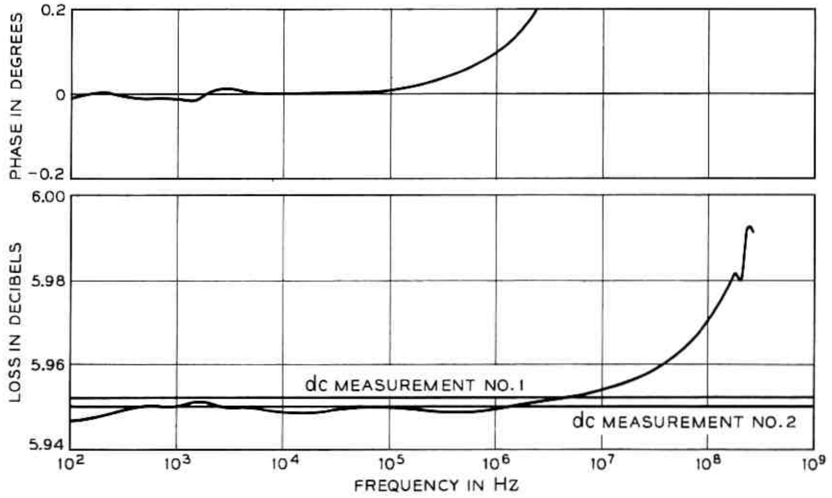


Fig. 15 — 6 dB coaxial pad measurement.

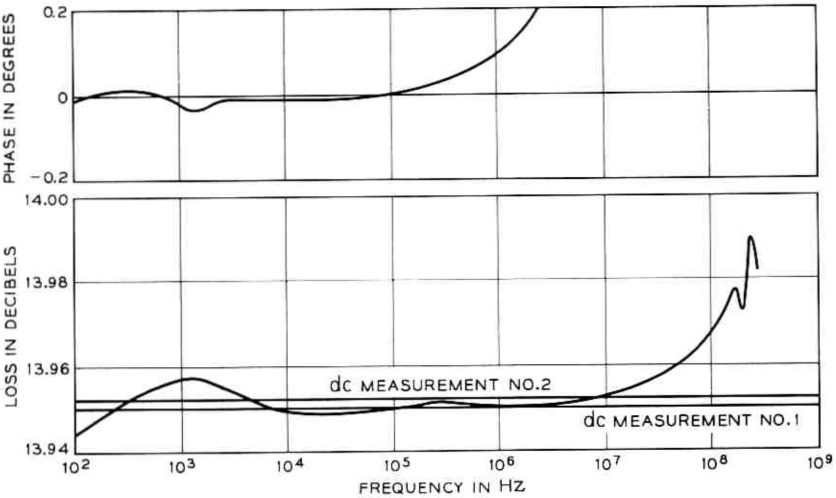


Fig. 16 — 14 dB coaxial pad measurement.

6.1.2 *Precision 30 cm Air Line*

A well-made precision air line is essentially ideal up to 250 MHz except for skin effect loss. The 30 cm line was first measured at a list of frequencies for comparison of measured loss and phase shift with theoretical values. Table IV gives this data. The discrepancies above 100 MHz are within the variation which could be caused by the ±0.02 cm tolerance in line length.

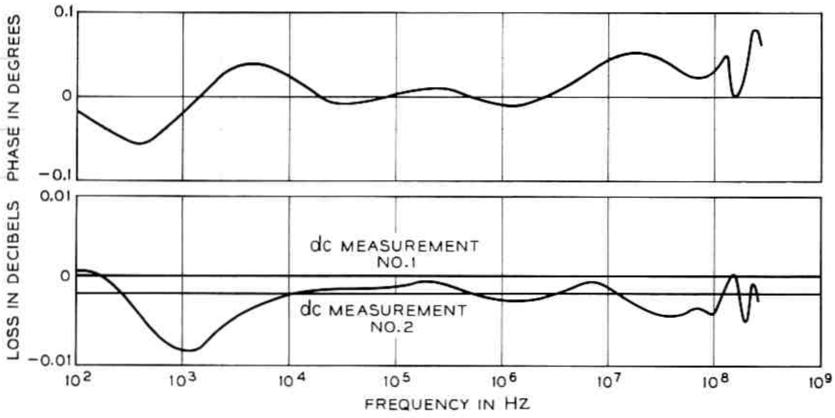


Fig. 17 — 14 dB and 6 dB coaxial pad addition.

Table IV — Phase Measurements on Precision 30 CM Air Line

Frequency (Hz)	Phase		$\theta_c - \theta_m$
	Measured θ_m (degrees)	Calculated* θ_c (degrees)	
27,777.77	−0.01	0.0115792	0.02
55,555.55	0.02	0.0222333	0
83,333.33	0.04	0.0327353	−0.01
111,111.11	0.05	0.0431584	−0.01
138,888.89	0.05	0.0535312	0
166,666.67	0.06	0.0638682	0
194,444.45	0.09	0.0741732	−0.02
222,222.23	0.07	0.0844666	0.01
250,000.01	0.08	0.0947376	0.01
277,777.79	0.12	0.1049939	−0.02
277,777.78	0.09	0.1049939	0.01
555,555.56	0.21	0.2070624	0
833,333.34	0.29	0.3086497	0.02
1,111,111.12	0.41	0.4099878	0
1,388,888.90	0.51	0.5111667	0
1,666,666.68	0.62	0.6122325	−0.01
1,944,444.46	0.70	0.7132126	0.01
2,222,222.24	0.80	0.8141248	0.01
2,500,000.02	0.91	0.9149817	0
2,777,777.80	1.02	1.0157921	0
2,777,777.78	1.02	1.0157921	0
5,555,555.56	2.02	2.0223333	0
8,333,333.34	3.02	3.0273524	0.01
11,111,111.12	4.04	4.0315841	−0.01
13,888,888.90	5.04	5.0353121	0
16,666,666.68	6.02	6.0386825	0.02
19,444,444.46	7.04	7.0417819	0
22,222,222.24	8.04	8.0446667	0
25,000,000.02	9.06	9.0473762	−0.01
27,777,777.80	10.05	10.0499389	0
27,777,777.78	10.06	10.0499389	−0.01
55,555,555.56	20.08	20.0706242	−0.01
83,333,333.34	30.09	30.0864966	0
111,111,111.12	40.09	40.0998777	0.01
138,888,888.90	50.13	50.1116667	−0.02
166,666,666.68	60.15	60.1223247	−0.03
194,444,444.46	70.17	70.1321258	−0.04
222,222,222.24	80.20	80.1412484	−0.06
250,000,000.02	90.22	90.1498166	−0.07

* $\theta_c = 90/250\ F_{\mathrm{MHz}} + 0.134\ (F_{\mathrm{MHz}}/200)^{\frac{1}{2}}$.

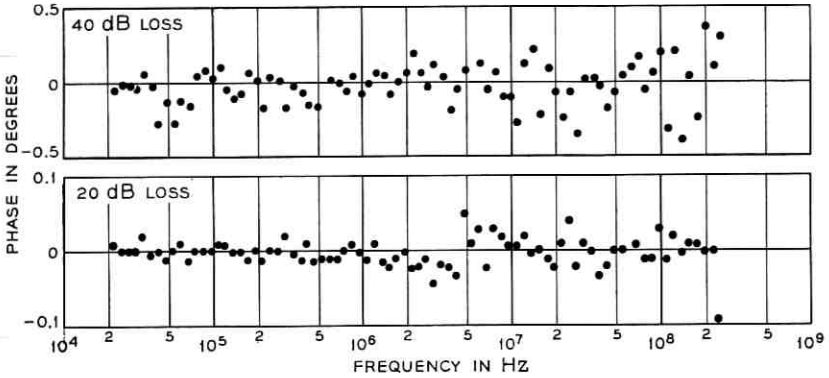


Fig. 18 — Effect of loss on fixed and random phase errors; —20 dBm loss level, 16 measurments averaged.

The same line was remeasured in tandem with 20 dB and 40 dB of loss. Figure 18 shows the effect of loss on the fixed and random phase errors. Figure 19 shows a curve of the measured delay of the line. The ±0.01 nsec variation in the measured value corresponds to a ±0.01° phase error with the 5 MHz frequency interval used.

6.1.3 *9 MHz Band Pass Crystal Filter*

Networks with especially "difficult" properties also provide useful information on the test set capabilities. Phase measurements in the passband of a 9 MHz filter with a 3 kHz bandwidth are especially sensitive to FM in the signal source.

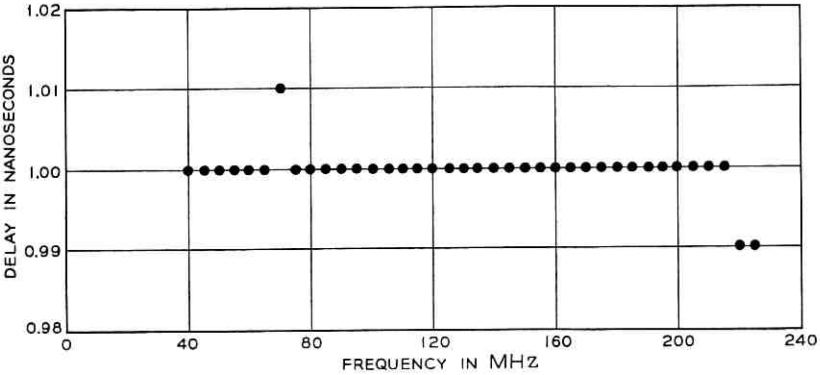


Fig. 19 — Measured delay of precision 30 cm air line.

The upper curve on Figure 20 shows the measured delay of the filter with an in-band delay of about 0.5 ms. The lower curve shows the difference of two successive delay measurements. Mid-band values repeat to ±0.001 ms, or ±0.02° with the 100 Hz frequency interval used.

6.2 *Impedance Measurements*

Resistors, capacitors, and inductors were measured to confirm the accuracy of impedance measurements. In the results given, the 50 ohm terminals and a 50 ohm impedance standard were used. The results were compared with measurements made on precision bridges.

Resistor measurements were made from 2 kHz to 10 MHz. For resistors (pigtail) of 100, 400, and 2500 ohms, the measurements are as shown in Fig. 21. The maximum deviation from the dc value of the resistors is less than 0.1 percent.

Capacitors with nominal values of 10, 100, 1000, and 10,000 pF were measured in the 2 kHz to 3 MHz range which provided a reactance range from 50 ohms to 500 k ohms. These results were compared with bridge measurements made at 100 kHz (less than 0.1 percent error) and the results are shown in Fig. 22.

Inductors with nominal values of 0.3, 1, 10, 100, and 1000 μh were

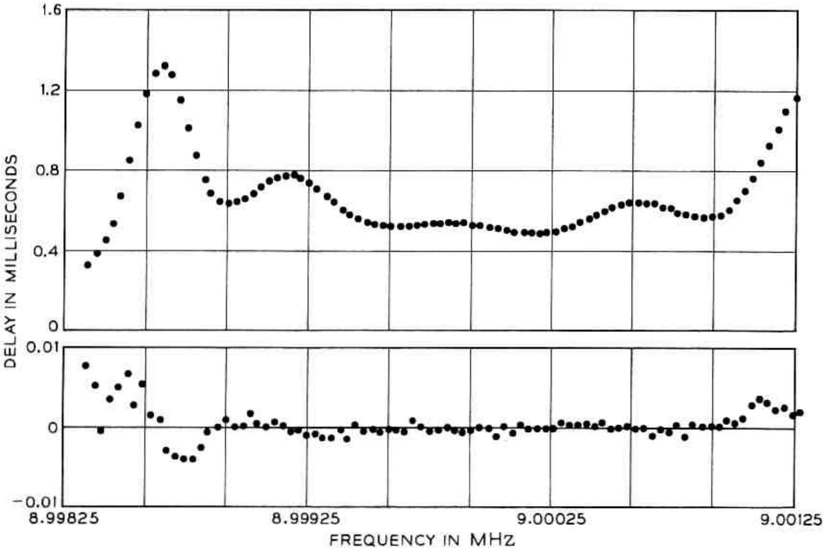


Fig. 20 — Measured delay of narrow band crystal filter.

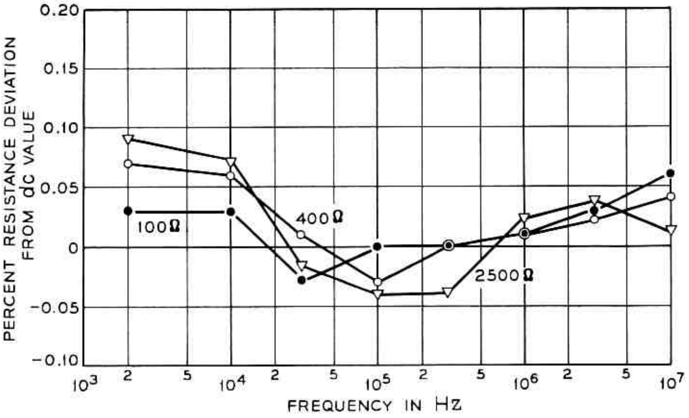


Fig. 21 — Deviations of measured resistance from dc values.

measured in the frequency range from 3 kHz to 10 MHz over an impedance range from 0.006 to 600 ohms. The results are shown in Fig. 23 with the indicated deviations being from bridge measurements (less than 0.05 percent error).

Typical agreement between impedance measurements on the set and the bridge is within 0.2 percent in favorable impedance ranges and the worst errors are less than 1 percent.

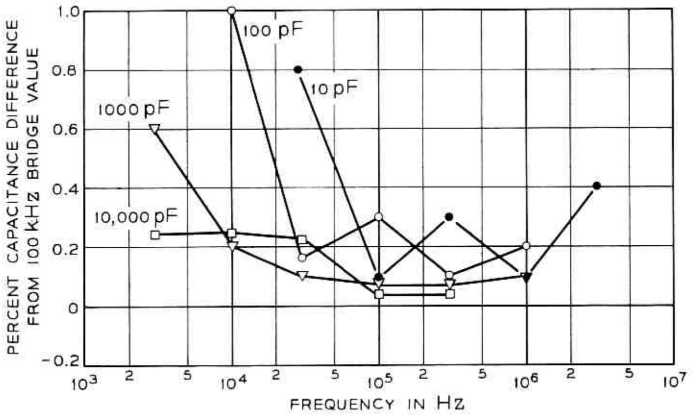


Fig. 22 — Deviations of measured capacitance from bridge measurements.

6.3 *Two-Port Measurements*

As an example of two-port characterization of a precise network, a precision 30 cm air line was measured.[6] Deviations of the measured characteristics from an ideal line result from measurement errors and skin effect. To emphasize these deviations, the characteristics of an ideal 30 cm line were mathematically removed from the measured data during processing. Figures 8 through 11 of Ref. 6 show the measured characteristics (normalized) compared with theoretical values. The S12 curves show the 0.001 dB and 0.01° resolution of the test set. Figure 24 shows curves of |S11| and |S12| for the same air line. The solid lines are the same as in the previous curves and the broken lines show the results that would be obtained if corrections were not made for test set misterminations.

Transistor characterization data is shown in Fig. 25. In this case, the transistor was measured in the common emitter, common base, and common collector modes and data from all three sets of measurements were transformed to common emitter *h* parameters. The curves shown are for magnitude and angle of *h*21 (beta).

VII. FURTHER DEVELOPMENT

The speed and flexibility of the computer operated transmission measuring set offers the opportunity for further development in a number of areas. Those being considered include increasing the frequency range, provision for measurements under environmental control and at remote locations, development of fault detection and location tests, and provision for interaction with a larger computer.

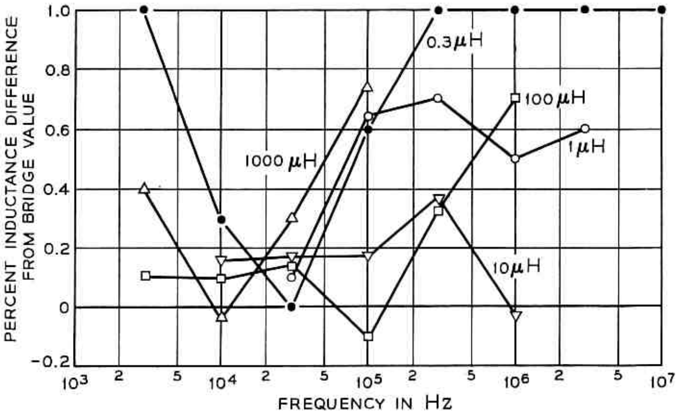


Fig. 23 — Deviations of measured inductance from bridge measurements.

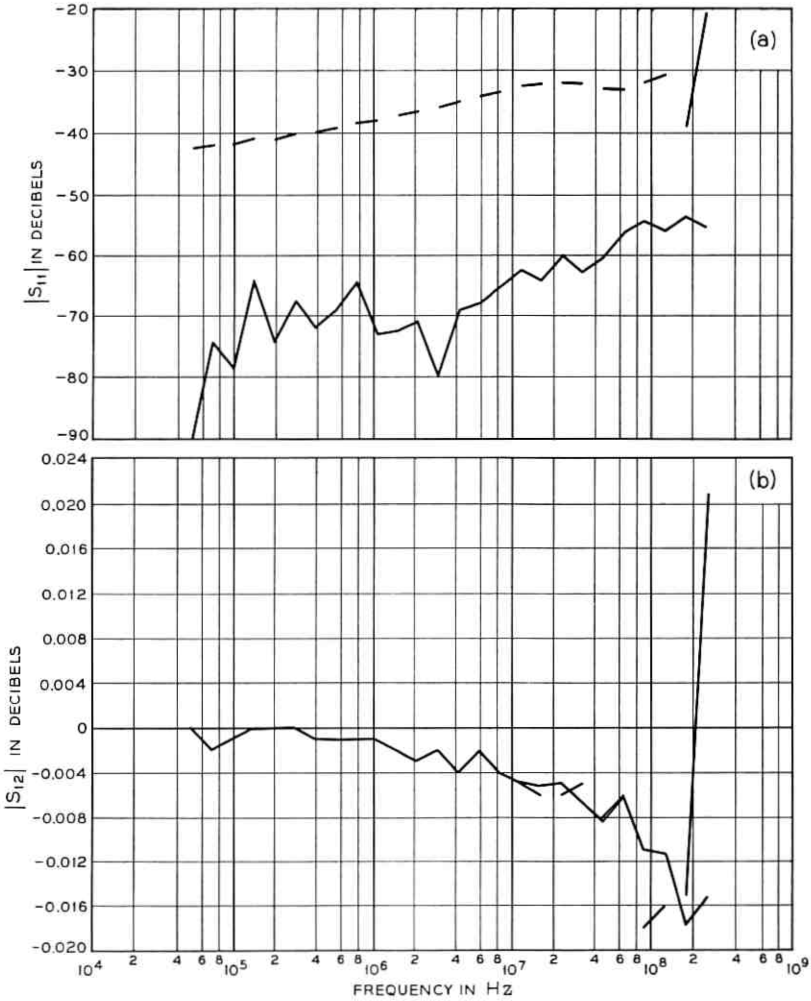


Fig. 24 — Normalized S parameters of precision 30 cm air line normalized to an ideal line of zero length (a) with and (b) without corrections for misterminations.

7.1 *Additional Measurement Capabilities*

One applique now under development will measure networks placed in an environmental chamber having temperature and humidity controls. After the environmental conditions have stabilized, computer control will transfer the necessary portions of the basic test set to the

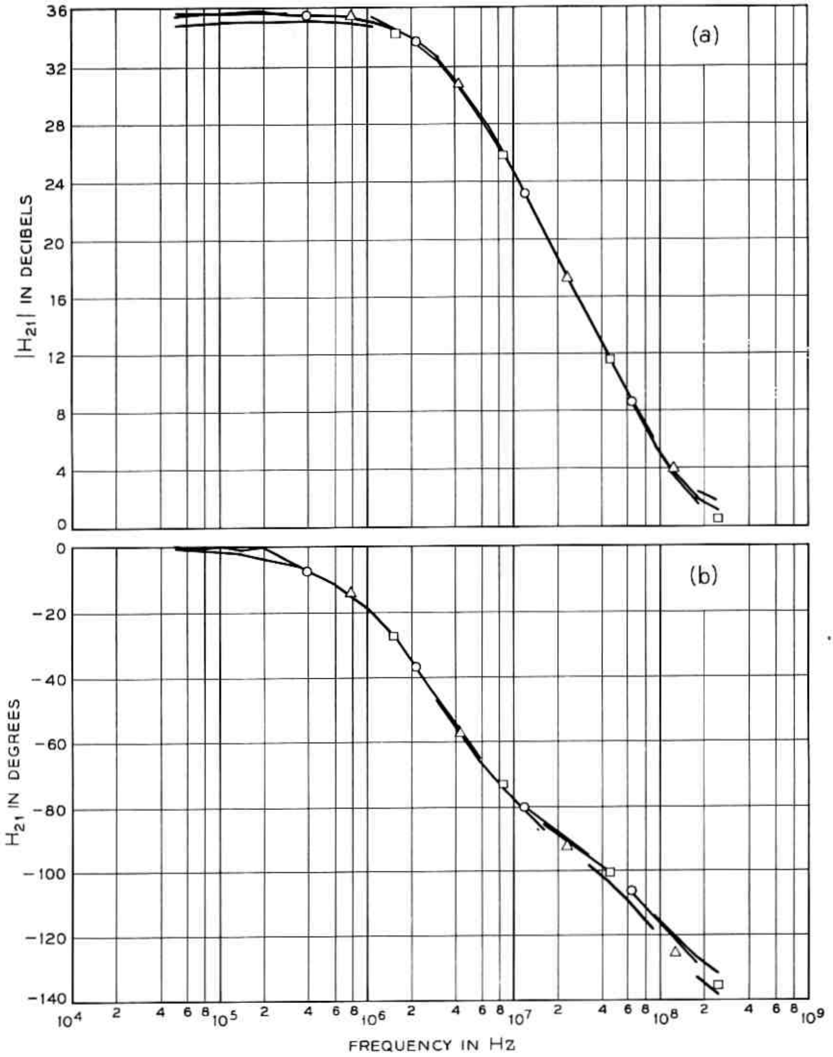


Fig. 25 — Transistor h parameters measured in the common emitter, common base, and common collector modes. ○ = 20C measured common emitter. △ = measured common base and computed common emitter. □ = measured common collector and computed common emitter.

applique for making measurements during idle periods on the basic set. The applique will have a 50 Hz to 250 MHz frequency range.

Other appliques are being developed to provide transmission measurements in the three microwave radio bands at 4, 6, and 11 GHz. These appliques will provide the microwave sources, control and switching, and down-converters. In the applique mode, about 80 percent of the basic set hardware is used.

Figure 26 is a block diagram of the 4 GHz applique which was recently completed.

A remote unit has been developed to permit transmission measurements hundreds of feet away from the basic set. Capabilities are not fully evaluated but 0.01 dB accuracy has been obtained over a 100 kHz to 100 MHz band at a distance of 275 feet from the basic set. Capabilities of the environmental applique and microwave appliques are estimated to be about 0.005 dB and 0.03°.

In the completed measurement center consisting of the basic set and appliques, it is planned to store data and control programs on a magnetic tape system. Then operation can be transferred from one applique to another in a few seconds.

7.2 *Maintenance Tests*

The question of test set accuracy is raised not only when a test set design is proven, but is a continuing question. The increased work load on the measurement center and the increased complexity of the total system inherent in the applique approach makes rapid and reliable fault detection and location vital. It is planned that the computer will enter an automatic test sequence whenever the measuring sets are idle. The test sequence would include tests for the computer as well as for the measuring sets.

7.3 *Frequency Extension*

Modular components are now available which will permit extension of the basic set operation up to 1000 MHz. Transmission circuits and frequency multipliers have been modified to operate to 1000 MHz, and a frequency converter to operate from 5 to 1000 MHz is being developed.

7.4 *Computer Interaction*

Coupling the measurement center to a larger computer via a data link is being considered. Two benefits of this connection are apparent.

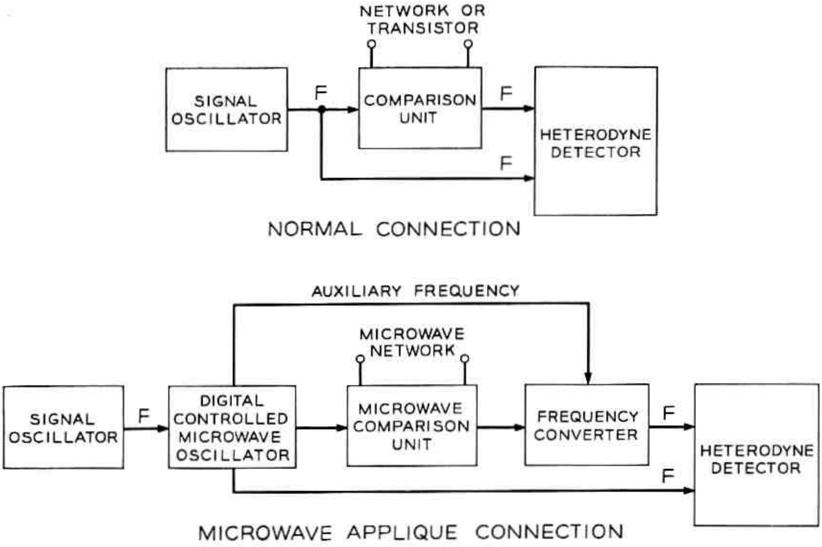


Fig. 26 — Extension of automated measurements to microwave frequencies.

The processing of two-port parameters now involves a considerable turn-around time. Direct connection to a larger computer could provide processed data in a few minutes. The problem of adjusting networks (particularly with interacting adjustments) might be substantially simplified if measurement data could be processed by a larger computer and adjustment information be returned to the test set.

VIII. SUMMARY

The computer-operated transmission measuring set provides high accuracy over a wide range of levels and frequency. The memory, control, and data processing capabilities of the computer provide the means to improve accuracy, operate at high speed, and provide versatility in the forms of output data.

The measurement capability of the set will make measurements an increasingly important part of the transmission system design process and, along with design aids such as computer analysis and modeling, improve the quality of systems being developed.

The flexibility inherent in the computer-operated set provides the opportunity for further development to increase its capability.

IX. ACKNOWLEDGMENTS

The computer-operated transmission measuring set hardware was developed by the Measuring Systems Design Department at Bell Laboratories. P. E. Rosenfeld suggested using the computer and developed the digital control and detection circuitry. J. G. Evans developed the transistor and two-port network characterization circuitry and technique (see Ref. 6). The computer program was developed by the Transmission Analysis Department at Bell Laboratories. We express particular appreciation to P. S. Meigs.

APPENDIX

Increased Resolution in Time Interval Method of Phase Measurements

A.1 *Basic Measurement*

A simplified block diagram of the time interval method of phase measurement is shown in Fig. 27. The measurement desired is $\theta_2 - \theta_1$. The pulse generators detect zero crossings; the first zero crossing occurs at $t_1 = \theta_1/\omega_1$, the second at $t_2 = \theta_2/\omega_1$. The exact time interval, Δt, is

$$\Delta t = t_2 - t_1 = \frac{\theta_2 - \theta_1}{\omega_1}$$

or in degrees

$$\Delta t = \frac{\varphi_2 - \varphi_1}{360 f_1}.$$

To measure the interval Δt with a counter using the pulse source F_s, the number of pulses, n, gated into the counter are: $2\pi(n \pm 1) = \omega_s \Delta t$.

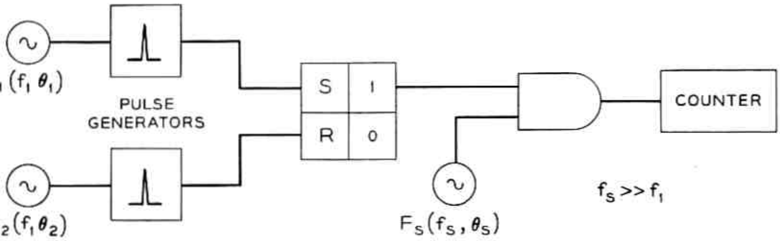


Fig. 27 — Time interval method of phase measurement.

Then:

$$\varphi_2 - \varphi_1 = 360 \frac{f_1}{f_s} (n \pm 1)^\circ .$$

For the case where $f_s = 100$ MHz and $f_1 = 27.777$ kHz,

$$\varphi_2 - \varphi_1 = 0.1n^\circ \pm 0.1^\circ .$$

This gives a resolution of 0.1°.

A.2 *Vernier Technique*

It will be shown that if f_s and f_1 are suitably related in frequency, multiple readings will provide increased resolution.

The circuit above can be modeled by integrating the product of two time functions. The gate opening is represented by $f(t)$, where

$$f(t) = \begin{cases} 1 & Kt_1 \leqq t < Kt_1 + \Delta t \\ 0 & Kt_1 + \Delta t \leqq t < (K+1)t_1 \end{cases}$$

and where

$$T = \text{the total measurement period}$$

$$K = 0, 1, 2, \cdots , \left(\frac{T}{t_1} - 1\right).$$

The pulse train from the source F_s is represented by $g(t)$, where:

$$g(t) = \sum_{m=o}^{\infty} \delta(t - \tau_o - m\tau)$$

and where: $\tau = 1/f_s$ and τ_o accounts for the relative phase between $f(t)$ and $g(t)$.

The total number of counts into the counter is represented by N, where

$$N = \int_0^T f(t)g(t)\, dt.$$

If we let the ratio of t_1/τ be I and I is an integer, the resolution (as in Section A.1) will be $360^\circ/I$. Integrating the product of $f(t)$ and $g(t)$ we obtain,

$$N = \sum_{K=0}^{(T/t_1 - 1)} \int_{Kt_1}^{Kt_1 + \Delta t} \sum_{m=o}^{\infty} \delta(t - \tau_o - m\tau)\, dt.$$

It is convenient to let $m = m' + Kt_1/\tau = m' + KI$, where m' indexes the delta function relative to the beginning of each gate opening.

$$N = \sum_{K=0}^{(T/I\tau-1)} \int_{KI\tau}^{KI\tau+\Delta t} \sum_{m=o}^{\infty} \delta(t - \tau_o - (m' + KI)\tau)\, dt.$$

In each integration, the delta functions will contribute when $KI\tau \leqq \tau_o + m'\tau + KI\tau < KI\tau + \Delta t$, or $-(\tau_o)/\tau \leqq m' < (\Delta t)/\tau$, which is independent of K. Thus the average of multiple readings will be the same as one reading. An example will illustrate this. If $t_1/\tau = I = 3600$ and $\Delta t = 1.5\tau$(that is, 0.15°), two cases are evident, as shown by Fig. 28.

In the first case there will be one count during each opening and in the second there will be two counts during each opening. Since the gate opening is periodic in t_1, the count will not vary and the reading will be $\Delta t \pm 0.5\tau$.

If the ratio of t_1/τ is varied so that the measurement is periodic in T, the precision is increased by T/t_1. In this example, during each counting period the delta function will move $(t_1/T)\tau$ seconds relative to $f(t)$ or a total of τ seconds during the measurement. This will provide a count of 1 during half the periods and a count of two in the other half. The average will be the correct number of 1.5.

To show this in the analysis, we define a slightly different frequency f_s' in the pulse source frequency so that $t_1/\tau' = I/(1 + s)$, where s is a term to give t_1/τ' a noninteger value. As before, T/t_1 has an integer value. Then

$$g'(t) = \sum_{m=o}^{\infty} \delta(t - \tau_o - m\tau') = \sum_{m=o}^{\infty} \delta(t - \tau_o - m\tau(1 + s))$$

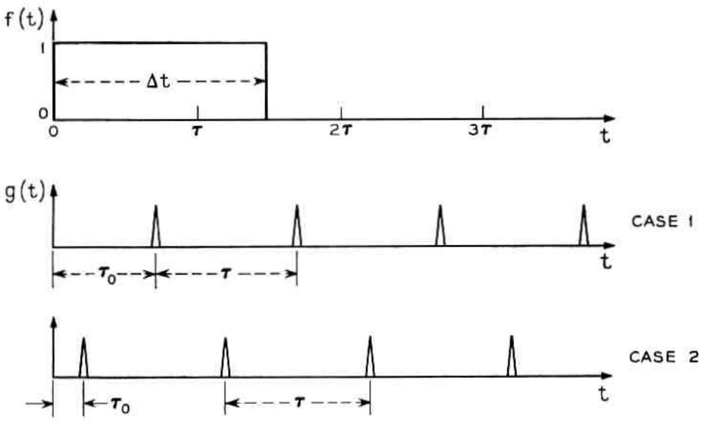


Fig. 28 — Example of average multiple reading being the same as one reading.

and

$$N' = \int_0^T f(t)g'(t)\,dt$$
$$= \sum_{K=o}^{(T/I\tau-1)} \int_{KI\tau}^{KI\tau+\Delta t} \sum_{m=o}^{\infty} \delta(t - \tau_o - m\tau(1 + s))\,dt.$$

By substituting $m = m' + KI$ as before and letting $s = \tau/T = t_1/(IT)$

$$N' = \sum_{K=o}^{(T/I\tau-1)} \int_{KI\tau}^{KI\tau+\Delta t} \sum_{m=o}^{\infty} \delta\left(t - \tau_o - (m' + KI)\left(1 + \frac{t_1}{IT}\right)\tau\right) dt. \tag{1}$$

The delta functions make contributions when

$$KI\tau \leqq \tau_o + (m' + KI)\left(1 + \frac{t_1}{IT}\right)\tau < KI\tau + \Delta t.$$

Solving for the integers m', two inequalities are obtained.

$$m' \geqq \frac{\dfrac{\tau_o}{-\tau} - K\dfrac{t_1}{T}}{1 + \dfrac{\tau}{T}} \tag{2}$$

and

$$m' < \frac{\dfrac{\Delta t}{\tau} - \dfrac{\tau_o}{\tau} - K\dfrac{t_1}{T}}{1 + \dfrac{\tau}{T}}. \tag{3}$$

For the ranges $0 \leqq (\tau_o)/\tau < 1$ and $0 \leqq K \leqq [(T/t_1) - 1]$, the values of m' provided by equation 2 are:

$$m' \geqq 0 \quad \text{for all } K$$
$$m' = -1 \quad \text{for } K \leqq \frac{T}{t_1}\left(1 + \frac{\tau}{T} - \frac{\tau_o}{T}\right).$$

Equation 3 depends on the measured quantity Δt and on the counting period, K. Using the previous example where $t_1/\tau = 3600$ and $\Delta t/\tau = 1.5$ assume that T/t_1 (the increase in resolution) $= 100$ and that $\tau_o/\tau = 0.5$.

Then from equation 2,

$$m' \geqq 0 \quad \text{for} \quad 0 \leqq K \leqq 99$$

$$m' \geqq -1 \quad \text{for} \quad 51 \leqq K \leqq 99.$$

Then the sum of the integrals yield

$$N' = \underbrace{100(1)}_{m'=0} + \underbrace{49(1)}_{m'=-1} = 149 \text{ for an average of 1.49 counts.}$$

As another example let $t_1/\tau = 3600$, $T/_{t1} = 100$, $(\tau_o)/\tau = 0.9$, and $(\Delta t)/\tau = 3599.9$ (that is, 359.99°).

Then from equation 2

$$m' \geqq 0 \qquad \text{for} \quad \text{all} \quad K$$

$$m' \geqq -1 \qquad \text{for} \quad 10 < K < 99.$$

And from equation 3

$$m' = 0, 1, \cdots, 3598 \quad \text{for all} \quad K.$$

Then the sum of the integrals yield

$$N = \underbrace{100(3599)}_{m'=0 \text{ to } 3598} + \underbrace{89(1)}_{\substack{m'=-1 \\ K>10}} = 100(3599.89)$$

for an average of 3599.89 counts.

In each case the resolution is increased from 0.1° to 0.001°.

REFERENCES

1. Leed, D. and Kummer, O., "A Loss and Phase Set for Measuring Transistor Parameters and Two-Port Networks Between 5 and 250 mc," B.S.T.J., *40,* No. 3 (May 1961), pp. 845–847.
2. Haynie, G. D. and Rosenfeld, P. E., "An Automated 20–20,000-cps Transmission Measuring Set for Laboratory Use," B.S.T.J., *42,* No. 6 (November 1963), pp. 2501–2531.
3. Elliott, J. S., "A High-Precision Direct-Reading Loss and Phase Measuring Set for Carrier Frequencies," B.S.T.J., *41,* No. 5 (September 1962), pp. 1493–1517.
4. Bode, H. W., *Network Analysis and Feedback Amplifier Design,* D. Van Nostrand Co., Princeton, N. J., 1945, p. 223.
5. Evans, J. G., "Linear Two-Port Characterization Independent of Measuring Set Impedance Imperfections," Proc. IEEE, *56,* No. 4 (April 1968), pp. 754–755.
6. Evans, J. G., "Measuring Frequency Characteristics of Linear Two-Port Networks Automatically," B.S.T.J., this issue, pp. 1313–1338.
7. Leed, D., "An Insertion Loss, Phase and Delay Measuring Set for Characterizing Transistors and Two-Port Networks Between 0.25 and 4.2 gc," B.S.T.J., *45,* No. 3 (March 1966), pp. 397–440.

Sun Tracker Measurements of Attenuation by Rain at 16 and 30 GHz

By ROBERT W. WILSON

(Manuscript received November 28, 1968)

This paper describes an instrument for measuring attenuation statistics on an earth-space path simultaneously at 16 and 30 GHz; the high attenuations result from heavy rain. The sun is used as a signal source during the time of day when the sun ordinarily is visible; a measuring range of more than 30 dB is achieved at both frequencies. The brightness temperature of the atmosphere also is measured both day and night. At night the antenna beam is stationary on the local meridian. Daytime brightness temperatures in conjunction with direct attenuation measurements are used to determine the equivalent absorber temperatures which are necessary for the reduction of night brightness temperatures to attenuation values. This paper discusses the measurements made during the first 12 months of operation and gives statistics of these measurements and an analysis of errors in the system.

I. INTRODUCTION

The advent of high performance booster rockets makes it possible to put very high-capacity microwave repeaters in synchronous orbit, possibly resulting in low cost per channel.[1] The large bandwidth required for such a system is in direct conflict with the crowded condition of the microwave spectrum below 10 GHz. We must therefore consider the possibility of operating such a system at frequencies above 10 GHz and must assess the magnitude of large attenuations which may be caused by heavy rain. Estimates based on attenuations for surface rainfall conditions[2] and models of the structure of rain storms indicate that such a system is feasible,[3] but direct measurement of the attenuation statistics is necessary.*

* The overall plan for a system calls for ground-station space diversity, but that is not discussed here.

The apparatus described in this paper has been set up at Crawford Hill, Holmdel, New Jersey, to measure the attenuation statistics of an earth-space path at 16 and 30 GHz using the sun as a signal source. At night the same equipment monitors the temperature of the antenna with the beam in the local meridian. Attenuations up to about 10 dB can be deduced from these temperatures. Attenuations of greater than 30 dB can be measured in the sun-tracking mode; the output time constant is two seconds so that even relatively fast fades can be followed.† Daily cycling of the equipment is automatic and sun coordinates are stored for a week's unattended operation. The sun tracker has been measuring at 30 GHz since October 1967 and at both 16 and 30 GHz in almost continuous operation since December 1967.

II. THEORY OF OPERATION

At 16 and 30 GHz the sun is transparent down to the lower chromosphere so the radiation temperature is fairly constant with time and fairly uniform across the disk of the sun. The value of the disk temperature is about 11,500°K at 16 GHz and 7,500°K at 30 GHz.[4]

If an antenna is pointed at the sun, the increase in antenna temperature because of the sun T_s is given by the product of the disk temperature of the sun and the fraction of the antenna's response which the sun subtends. At 30 GHz somewhat less than half of the response of the sun-tracker antenna falls on the disk of the sun, so $T_s \approx 3000$°K. If the remainder of the antenna's pattern is directed to cold space, the total antenna temperature will equal T_s. If we introduce a uniform lossy medium of transmission coefficient t and physical temperature T_c between the antenna and the sun, the antenna temperature (T_a) will be changed from T_s to

$$T_a = tT_s + (1 - t)T_c \tag{1}$$

where radiation from the attenuating medium takes the place of some of the radiation from the sun. In our case T_c is the temperature of some component of the earth's atmosphere (in particular, rain) and will be about 270°K; but we are not able to measure it directly. For attenuations greater than about 12 dB, the second term of equation (1) will dominate and a simple measurement of antenna temperature therefore would not provide a linear measurement of attenuation; for attenuations greater than about 20 dB the errors resulting from the

† If the signal going into a 2 second time constant is rapidly reduced to zero, the output drops at 2.2 dB per second, whereas if the signal is rapidly increased the output comes within 2 dB of the final value in 2 seconds.

unknown value of T_c would start to be significant. In the sun tracker these problems are solved by having the antenna's main beam scan on and off of the sun at a 1 Hz rate with an angular excursion of 2.6°. When the beam is 2.6° away from the sun, virtually the only radiation is that of the attenuating medium and

$$T_a = (1 - t)T_c\,. \tag{2}$$

The output of the receiver is sampled during the time the antenna is pointed at the sun and when it is pointed away from the sun; thus a difference voltage is generated. By subtracting equation (2) from (1), one sees that this voltage is proportional to tT_s. As long as T_s is constant any changes in the difference voltage can be interpreted as changes in t; thus, it is not necessary to know T_s in any absolute sense, just as a reference level at the radiometer output. In the sun tracker the difference voltage is passed through a logarithmic converter and presented on a chart recorder so the attenuation can be read directly in dB.

At night the sun is not available and only equation (2) can be used as a measure of attenuation. In this case a Dicke switch is used and the temperature of the antenna is subtracted from the temperature of a reference termination at about 290°K. The quantity plotted on the chart recorder is

$$\Delta T = 290°\text{K} - (1 - t)T_c\,.$$

Uncertainties in the value of T_c limit the range for which t can be recovered to about 10 dB in this mode of operation.

An additional output is obtained during the daytime by using the Dicke switch to connect the input of the receiver to the reference termination during the transition portion of the scanning cycle, that is, when the main beam is neither fully on nor off of the sun. A radiometer output similar to that in the nighttime operation is obtained by comparing the off-sun antenna temperature with the reference termination temperature. From the simultaneous measurements of T_a and t, T_c can be calculated.

III. EQUIPMENT

Figure 1 shows the physical layout of the equipment. A five-by-nine-foot plane reflector is mounted as a polar heliostat to reflect the sun's rays in the direction of the earth's north polar axis. A four-foot aperture conical horn-reflector antenna looks south along the same

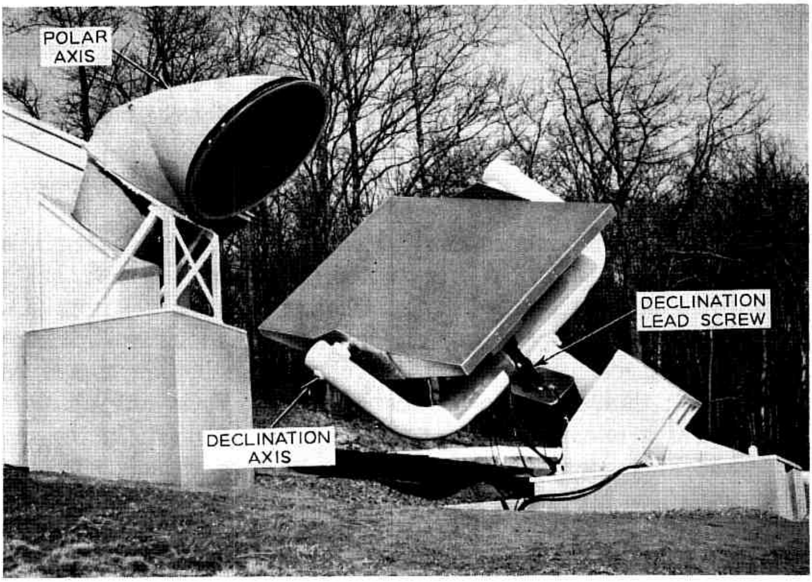


Fig. 1 — View of sun tracker from southwest showing equipment cab, horn reflector, and polar heliostat flat.

axis and collects most of these rays. The hour-angle motion for tracking the sun during any one day is provided by driving the reflector about its polar axis at a 24-hour per revolution rate. This tracking motion is automatically started each morning from an approximate starting position when the read-outs indicate a coincidence between the antenna beam and the ephermeris positions of the sun for that day. The seasonal motion of the sun (in declination) is corrected daily by motion of the reflector about its declination axis at half the angular rate. A motor-driven lead screw is automatically energized for a timed interval each morning to provide the required motion. The declination axis of the reflector is also used for the 1 Hz scanning motion mentioned in Section II. The upper end of the declination lead-screw connects to a crank shaft which is turned at about 1 Hz in the sun tracking mode of operation. A resolver, turned by the same shaft, generates timing signals for the radiometers.

The output of the horn-reflector antenna is in a circular waveguide. One linear polarization is split off by a polarization coupler for the 16 GHz receiver and the orthogonal polarization passes through a waveguide taper to the 30 GHz receiver. Figure 2 is a block diagram of the radiometer system.

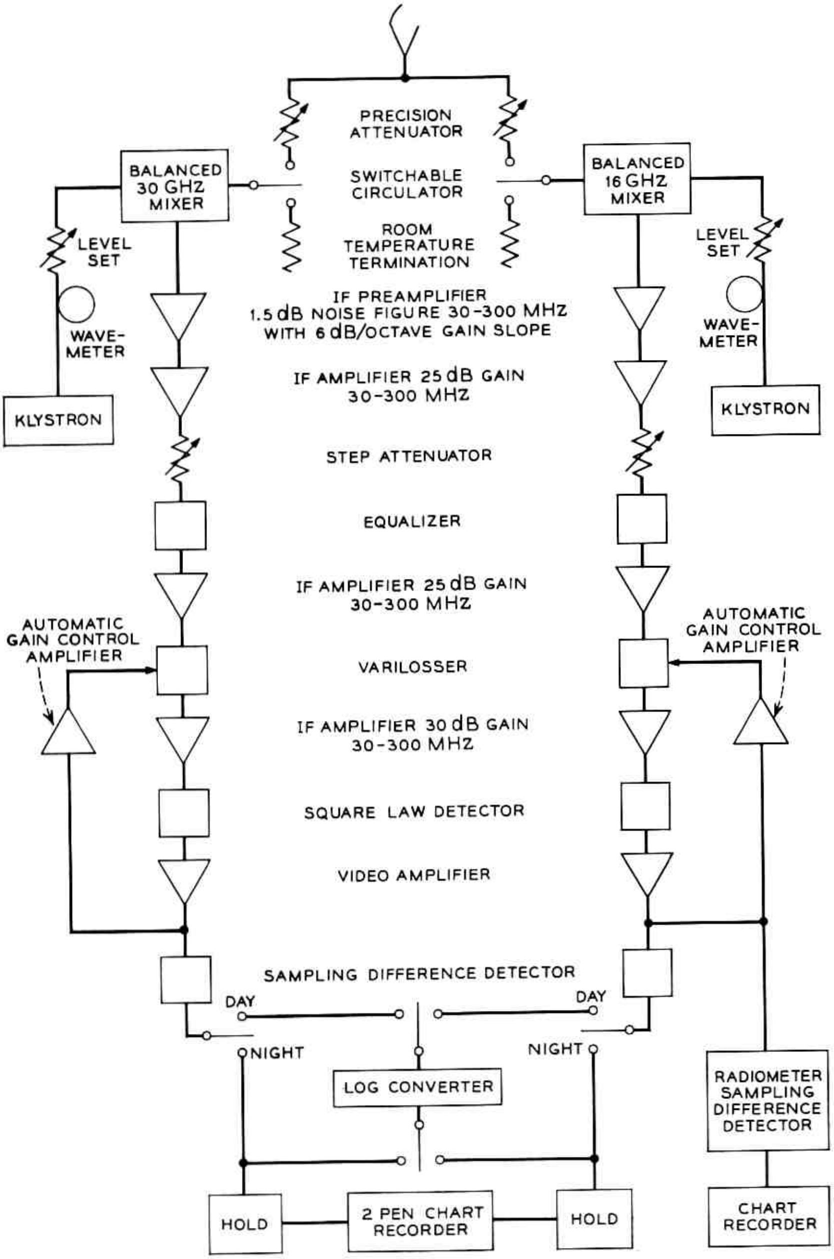

PRECISION ATTENUATOR
BALANCED 30 GHZ MIXER
SWITCHABLE CIRCULATOR
BALANCED 16 GHZ MIXER
LEVEL SET
ROOM TEMPERATURE TERMINATION
LEVEL SET
IF PREAMPLIFIER 1.5 dB NOISE FIGURE 30-300 MHz WITH 6 dB/OCTAVE GAIN SLOPE
WAVE-METER
WAVE-METER
IF AMPLIFIER 25 dB GAIN 30-300 MHZ
KLYSTRON
KLYSTRON
STEP ATTENUATOR
EQUALIZER
IF AMPLIFIER 25 dB GAIN 30-300 MHZ
AUTOMATIC GAIN CONTROL AMPLIFIER
AUTOMATIC GAIN CONTROL AMPLIFIER
VARILOSSER
IF AMPLIFIER 30 dB GAIN 30-300 MHZ
SQUARE LAW DETECTOR
VIDEO AMPLIFIER
SAMPLING DIFFERENCE DETECTOR
DAY
DAY
NIGHT
NIGHT
LOG CONVERTER
RADIOMETER SAMPLING DIFFERENCE DETECTOR
HOLD
2 PEN CHART RECORDER
HOLD
CHART RECORDER

Fig. 2 — Block diagram of sun tracker receivers.

The balanced Schottky-barrier-diode mixers have broadband input circuits allowing double sideband response and are directly connected to transistor IF preamplifiers. The resulting down converter has a low noise temperature (T_r) which varies somewhat over the IF bandpass of the system with a broad minimum around 70 MHz. The gain has a 6 dB per octave slope characteristic of high frequency transistors operated in the β cutoff region. The equalizer following the second IF amplifier changes the sloping frequency response of the system to a weighting function [gain $\alpha(T_o/T_r)^2$] which minimizes the fluctuation level at the output of the square law detector when referred back to the input temperature. The weighted average double sideband noise temperature of the receivers is 840°K at 30 GHz and 1300°K at 16 GHz. The noise bandwidth exceeds 200 MHz in both cases. The noise temperatures of the receivers are degraded in the radiometer system by the combined loss of about 1 dB in the calibrating attenuator and switchable circulator.

The operating cycle of the radiometer, when tracking the sun, is shown in Fig. 3. The top curve shows the declination of the antenna beam as a function of time; the second curve shows the resulting output of the square law detector of the 30 GHz receiver. The third curve shows the drive to the circulator switches which connect the receiver inputs to the room temperature reference terminations during the quarters of the cycle while the declination is changing rapidly. This switch causes the shoulder in the second detector output. The fourth curve shows the drive to the main sampling difference detectors. Positive sampling occurs during the quarter of the cycle when the beam is closest to the sun and negative sampling during the quarter cycle when the beam is farthest from the sun. The positive and negative samples are stored on separate capacitors with a charging time constant of 0.5 second.

The sampling duty cycle of ¼ gives a speed of response equivalent to a 2-second time constant. The fifth trace shows that the logarithmic converter (Fig. 2) operates on the output of the 30 and 16 GHz sampling difference detectors alternately during the remaining two quarters of the cycle. The sixth curve shows the drive to the radiometer sampling difference detector; it samples positively when the receiver is connected to the room temperature termination and negatively while the main beam is pointed away from the sun producing an output proportional to ΔT of equation (3).

The last trace shows the drive to the automatic gain control sampling circuit. The action of the automatic gain control is to adjust the

IF variolosser as necessary to keep the output level of the receiver at a fixed value during the portion of the cycle that its input is connected to the room temperature termination. An integrator in the automatic gain control amplifier prevents the gain from changing rapidly. Dead times in the cycle have been exaggerated for clarity in the figure.

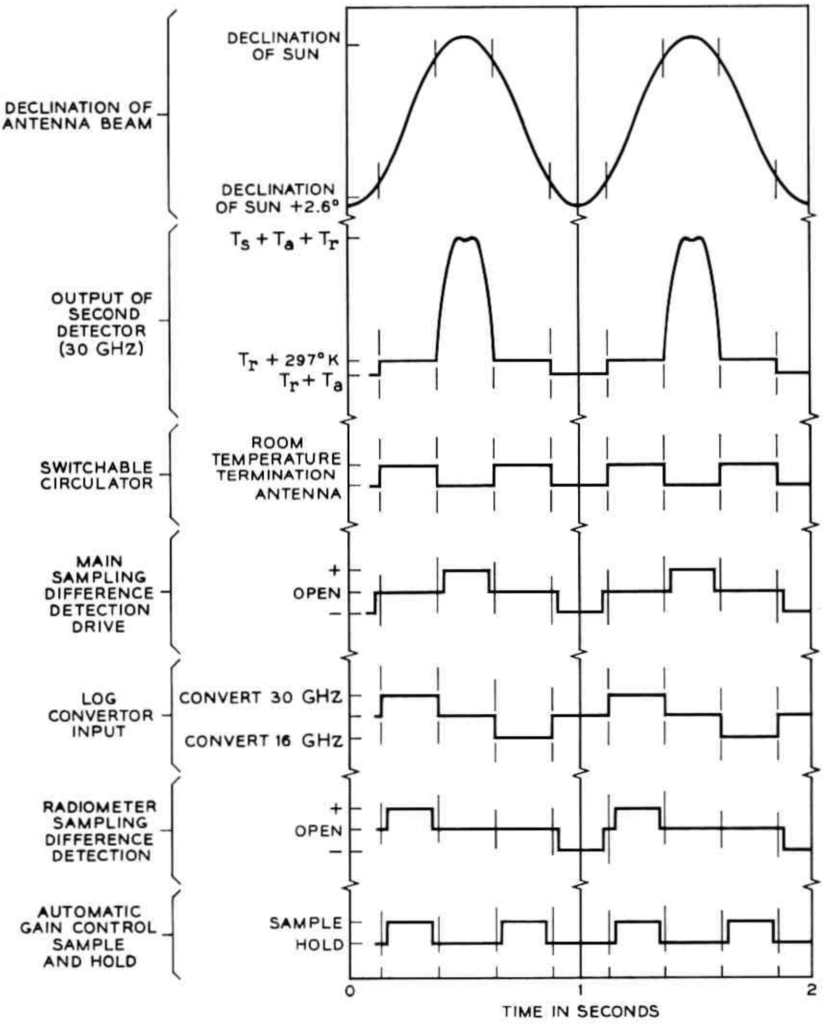

DECLINATION OF ANTENNA BEAM
DECLINATION OF SUN
DECLINATION OF SUN +2.6°
OUTPUT OF SECOND DETECTOR (30 GHZ)
$T_s + T_a + T_r$
T_r + 297°K
$T_r + T_a$
SWITCHABLE CIRCULATOR
ROOM TEMPERATURE TERMINATION
ANTENNA
MAIN SAMPLING DIFFERENCE DETECTION DRIVE
+
OPEN
−
LOG CONVERTOR INPUT
CONVERT 30 GHZ
CONVERT 16 GHZ
RADIOMETER SAMPLING DIFFERENCE DETECTION
+
OPEN
−
AUTOMATIC GAIN CONTROL SAMPLE AND HOLD
SAMPLE
HOLD
0
1
2
TIME IN SECONDS

Fig. 3 — Switching cycle of sun-tracker radiometers.

During nighttime operation the antenna beam is pointed to the meridian at the declination the sun had the previous day. The switchable circulators are operated as Dicke switches at about 2 Hz and the sampling difference detectors operate as phase sensitive detectors with blanking at the switching time. Their outputs go directly to the hold circuits which drive the recorder.

Various parameters of the system are summarized in Table I. The temperatures quoted refer to the antenna terminals. Room temperature and liquid nitrogen cooled absorbers are used as temperature standards at 30 GHz and a noise lamp and a room temperature absorber are used at 16 GHz.

IV. CLEAR WEATHER ATTENUATION

The antenna temperature (pointed away from the sun) has been measured as a function of elevation on a typical clear winter day (+3°C, 60 per cent relative humidity) and again on a hot summer day (37°C, 52 per cent relative humidity). Values of 8.3° and 17.1° K per atmosphere at 30 GHz and 4.1° and 7.8°K per atmosphere at 16 GHz respectively, were found. These correspond to 0.15 and 0.25 dB per atmosphere at 30 GHz and 0.06 and 0.12 dB per atmosphere at 16 GHz, assuming the absorption took place at 250°K (winter) and 284°K (summer). Attempts to measure these rather small atmospheric absorptions directly using the sun, under atmospheric conditions similar to the above, have been frustrated by variations in atmospheric absorption, solar brightness, or antenna gain during the course of the measurement. Consistent results have obtained only at low elevation angles where the thickness of the atmosphere changes rapidly with hour angle.

The normalization procedure which is normally used on sun tracker records cancels out clear weather attenuation. Thus attenuations quoted in other parts of this paper are increases above the clear weather value.

Table I — Parameters of the Sun Tracker

	16 GHz	30 GHz
Antenna beam width	0.92	0.54°
Antenna temperature of sun (T_s) (30° elevation)	1900°K	3000°K
Receiver double side band noise temperature	1700°K	1100°K
Measuring range on sun (1.5 dB peak error)	30 dB	35 dB

V. TYPICAL RECORDS

Figure 4 shows tracings of the output of the sun tracker during two 24-hour periods. The left two-thirds of both charts is night operation with antenna temperature presented on linear scales for both frequencies. The right portion is in the sun-tracking mode with a scale factor of 10 dB per major division. The sun is behind some trees during part of the sunrise.

The upper record was taken on a clear day. During the lower record, several showers with rainfall up to 75 mm per hour occurred near the site. The high temperature peaks on the night portion of the 30 GHz record show rounding, indicating that the peak antenna temperature of about 275°K is close to the temperature of the attenuating rain. The three peaks of attenuation in the daytime portion of the record occurred at solar elevation of 8°, 15°, and 18°.

Figure 5 contains tracings of the sun-tracker output on three other days. Figure 5a is from a 24-hour period when the sky was heavily overcast and occasional drizzle occurred. The attenuation did not exceed 2 dB during this period even for low elevations and was less than 1 dB most of the time.

The lower records were obtained before the 16 GHz receiver was installed so that only 30 GHz levels are plotted. During the night that the lower level record was taken (b) passage of a cold front produced snow and ground level temperatures fell below 0°C. The next morning about ¼ inch of rough ice was frozen on the reflector of the sun tracker. When the ice was removed from the reflector (at the right end of b), the signal level from the sun returned to normal.

The following sequence of events is postulated to explain the record. As snow fell on the warm reflector it melted to slush. The liquid water content of the slush has a very high absorption coefficient when its thickness amounts to an appreciable fraction of a wavelength above the aluminum reflector; as the slush collected, the antenna temperature approached ambient temperature. After the snowfall stopped, the antenna temperature remained constant until the air temperature lowered sufficiently to slowly freeze the slush into ice which is a dielectric with a low absorption coefficient; thus the antenna temperature dropped to a value typical of the overcast night (Fig. 5b). When the sun rose and the sun tracker started tracking it, however, the signal level was about 7 dB below normal because of phase perturbations (and a consequent reduction in gain) caused by the rough dielectric on the reflector. Removal of the ice returned operation to normal (final short segment in b).

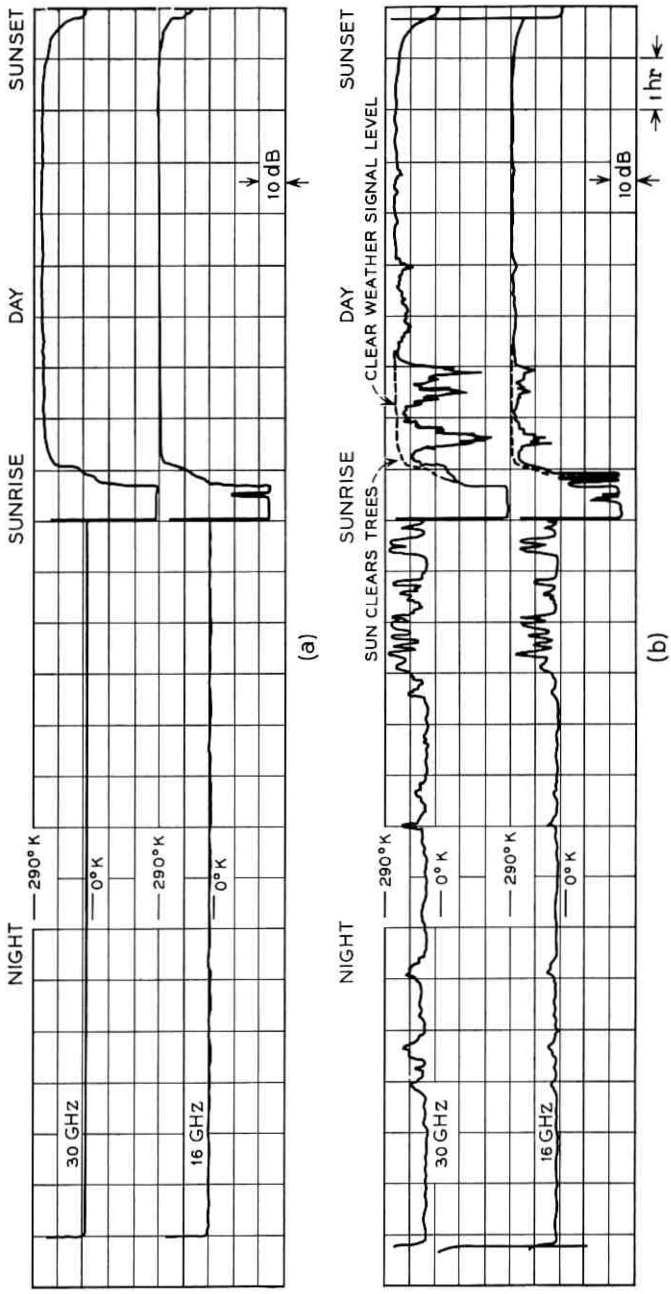


Fig. 4 — (a) Output of sun racker during a clear 24 hour period; the left portion of the record is nighttime operation with antenna temperature scales as indicated. The right portion shows sun tracking operation in which the scale factor is 10 dB per major division. (b) Output of sun tracker during occasional moderate showers (December 12, 1967).

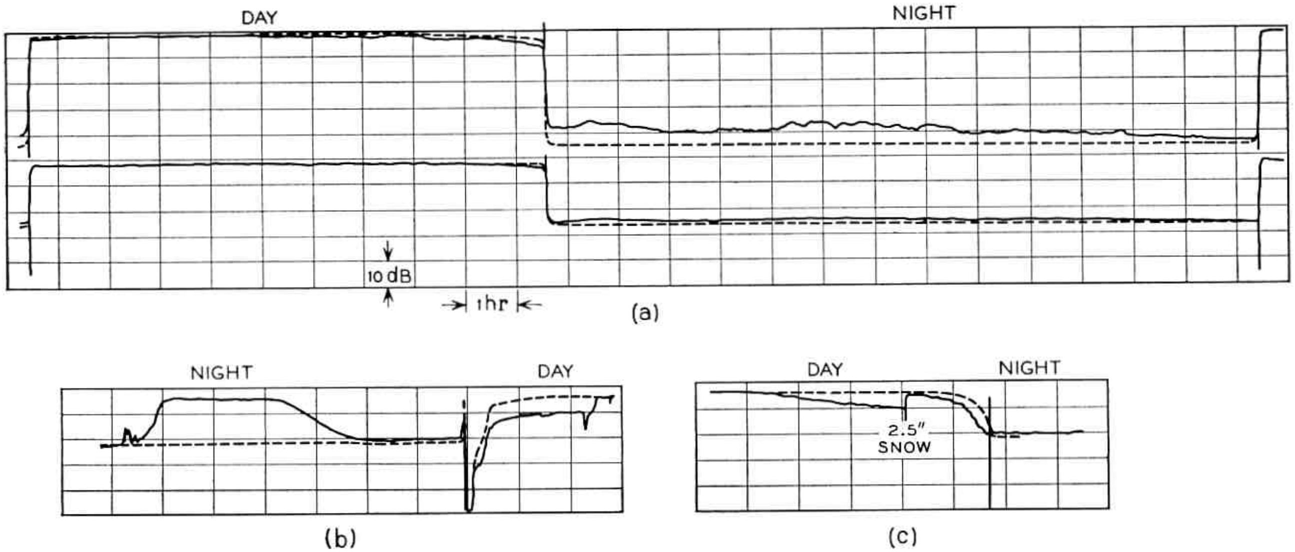


Fig. 5 — (a) Output of sun tracker during a 24 hour period of overcast skys and occasional drizzle. The dotted line indicates clear weather output. Notice that the gain has been increased in this record compared with Fig. 4, but things are otherwise the same (March 10, 1968). (b and c) Records at 30 GHz for times of wet and dry snowfall; see text for explanation (November 15, 1967 and November 30, 1967).

The lower right record (c) was taken on a cold afternoon when snow fell uniformly for several hours. In 2½ hours about 2½ inches of snow collected on the reflector. When the snow was removed from the reflector to measure the effect of the falling snow, the signal level returned to about 1 dB below the normal level. More snow collected on the reflector, but the antenna temperature was hardly affected in the night-time mode.

VI. ATTENUATION STATISTICS

At this writing, the sun tracker has been in full operation for more than a year. On two occasions the attenuation at 30 GHz continuously exceeded the measuring range of the system (35 dB) for more than 30 minutes. During one of these periods the attenuation at 16 GHz exceeded 30 dB four separate times for a total of 15 minutes. It is clear from these results that if a reliable communication satellite system is to be constructed using these frequencies it will be necessary to have some form of ground-station space diversity for operation during such periods.

A summary of the percentage of time that the attenuation exceeded various levels is shown in Tables II and III for 30 and 16 GHz. In both cases day and night statistics are shown separately since the measurement technique is different. Figures 6 and 7 are histograms showing the number of fades exceeding 9 dB at 30 and 16 GHz as a function of duration of the fade. No attempt has yet been made to divide this statistical data according to the elevation of the sun. It is expected that some differences will occur as a function of elevation; however, one of the two long-term high-attenuation periods, mentioned before, occurred when the elevation angle was about 60° and the other immediately before sunset.

VII. RATIO OF ATTENUATION AT 30 GHZ TO ATTENUATION AT 16 GHZ

If one knew the drop size distribution in an attenuating rain, the ratio of attenuation at 30 GHz to that at 16 GHz could be calculated. Using surface drop size distributions, Hogg has calculated that the ratio will lie between about 3.8 for small drops characteristic of 0.1 mm per hour rain and 2.2 for large drops characteristic of a 100 mm per hour rain.[5] Values over this range have been observed at various times.

Figure 8 is a scatter plot of attenuation at 16 GHz against attenuation at 30 GHz for various sample times during the first two hours of

TABLE II — CUMULATIVE DISTRIBUTION OF ATTENUATION at 30 GHz FOR DAY AND NIGHT OBSERVATIONS

Attenuation at 30 GHz (in dB)	Percent of total observing time* (3851 daylight hours)
> 3	1.97
> 6	1.00
> 9	0.55
> 15	0.309
> 21	0.174
> 27	0.105
> 33	0.069
	(4826 nighttime hours)
> 3	1.148
> 6	0.300
> 9	0.113
> 12	0.052

* December 8, 1967 through December 8, 1968. The daytime observations include all elevations from a maximum of 74° down to as low as 2° or 3°. Nighttime observations are made at a constant elevation varying from 73° on June 21 to 26° on December 21. Elevation effects may contribute to the differences between day and night distributions as well as rainfall differences.

the daytime portion of the record shown in Fig. 4(b). Except for two of the points, the dashed line which represents a constant ratio of 3.4 to 1 is a good fit to the data. Figure 9 shows points from a thunder storm in which the ratio taken from the second order-fitted curve varies from 3 at high attenuations to more than 4 at low attenuations. Figure 10 shows points from another thunderstorm during which two separate observers remarked on the unusually large size of the raindrops. The ratio in this case was about 2.2 to 1. The ratio for other rains has fallen within the range indicated above.

VIII. SUN VERSUS SKY BRIGHTNESS MEASUREMENTS

As explained in Section II a sky brightness measurement is made at one frequency simultaneously with the measurements of attenuation using the sun. With both attenuation and brightness the equations of Section II can be solved in either of two ways. In Fig. 11 a scatter plot has been made of attenuation derived from the sky-brightness measurement using equation (2) against simultaneous attenuation measured in the direction of the sun. It can be seen from this and other data that if the correct value for T_c is used (272°K in this

TABLE III— CUMULATIVE DISTRIBUTION OF ATTENUATION AT T6 GHz FOR DAY AND NIGHT OBSERVATIONS

Attenuation at 16 GHz (in dB)	Percent of total observing time* (3839 daylight hours)
> 1	1.59
> 2	0.84
> 3	0.45
> 5	0.259
> 7	0.158
> 9	0.112
> 11	0.085
> 13	0.065
> 15	0.049
> 20	0.034
> 25	0.023
> 30	0.013
> 33	0.009
	(4812 nighttime hours)
> 1	0.46
> 2	0.13
> 3	0.05
> 6	0.022
> 9	0.016
> 12	0.012

* Same dates as Table II. The daytime observations include all elevations from a maximum of 74° down to as low as 2° or 3°. Nighttime observations are made at a constant elevation varying from 73° on June 21 to 26° on December 21. Elevation effects may contribute to the differences between day and night distributions as well as rainfall differences.

case), measured sky brightness values can be interpreted as attenuations with reasonably small scatter up to and perhaps beyond 10 dB. (Some of the scatter in Fig. 11 is undoubtedly caused by real differences in attenuation in the two directions.)

A more interesting way of looking at this same data is to invert the equations and compute the apparent medium temperature T_c. In Fig. 12 the derived value of T_c has been plotted (dots) against measured attenuation for the same data as used in Fig. 11. At low values of attenuation the average value of T_c seems to be below the ice point even though the air temperature near the earth's surface was about 295°K during this rain. Super cooling might play some role in causing this low apparent temperature, but it is more likely that scattering as discussed in Section IX causes the main effect.

At high values of attenuation the measured value of T_c goes up to as high as 290°K; this is a very definite effect since the measured

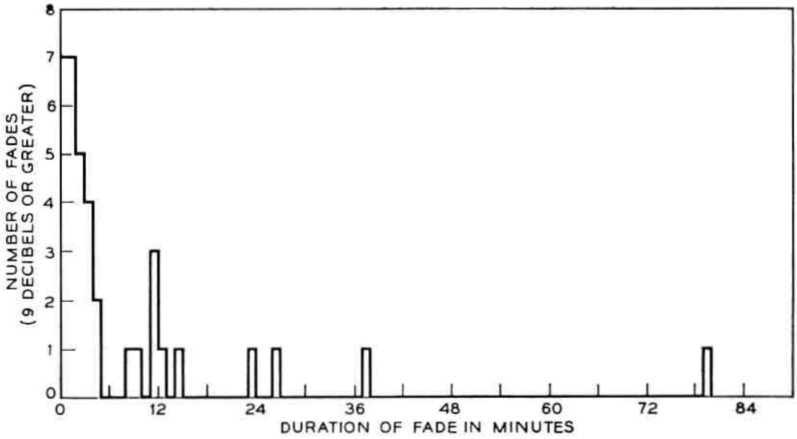


Fig. 6 — Number versus duration for fades of 9 dB or greater at 30 GHz for the period December 1967 to August 1968.

brightness temperature rises to 290°K. However, if a plot like Fig. 11 is made using such a high value of T_c, there is a definite curve at attenuations above 5 dB and the fit is unacceptable by 10 dB. There are two contributions to the higher values of T_c at higher attenuations. First, scattering ceases to be very effective in lowering T_c; instead of scattering radiation from the cold sky into the antenna beam, the lower drops scatter radiation from upper drops into the beam. Second, at high values of attenuation only the lower and hotter portion of the rain contributes effectively to the brightness since the lower drops absorb the radiation from the upper drops and replace it with their own.

IX. DEVIATIONS FROM SIMPLE THEORY

The output of the sun tracker could depart from the true attenuation in the path to the sun for several reasons:

(*i*) Nonlinearities and instabilities in the radiometers, are small enough to be negligible.

(*ii*) Mispointing the antenna beam as a result of atmospheric refraction, use of noon solar positions during an entire day, and mechanical misalignment cause the signal from the sun to decrease more at low elevations than one expects from atmospheric absorption. These effects fortunately are quite repeatable from one day to the next so that clear weather days provide a reference level below which excess attenuation is measured.

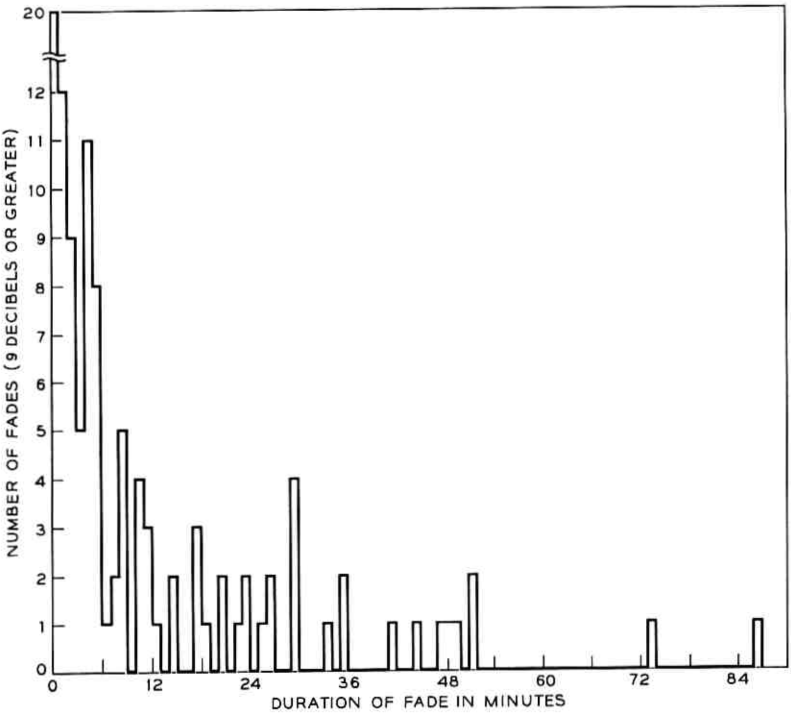


Fig. 7 — Number versus duration for fades of 9 dB or greater at 16 GHz from December 1967 to August 1968.

(*iii*) The brightness of the sun may not be constant. Clear weather records to date show no noticeable variations from day to day, except for a large increase for 40 minutes during the solar event of July 8, 1968 and increases of 1 dB or less lasting only a few minutes on several other occasions.

(*iv*) Part of the received signal might result from forward scattering by the precipitation. The scattered energy might therefore be collected by the relatively broad beam of the antenna and be indistinguishable from the direct signal. However, since rain drops are not large compared with a wavelength, the forward scattering lobe will be relatively weak and large in angular diameter. Moreover, approximately equal scattered power will be picked up in the direction of the sun and in the reference direction 2.6° away. Forward scattered energy should therefore cancel out, resulting in a proper measurement of attenuation. In measuring sky brightness at low attenuations, how-

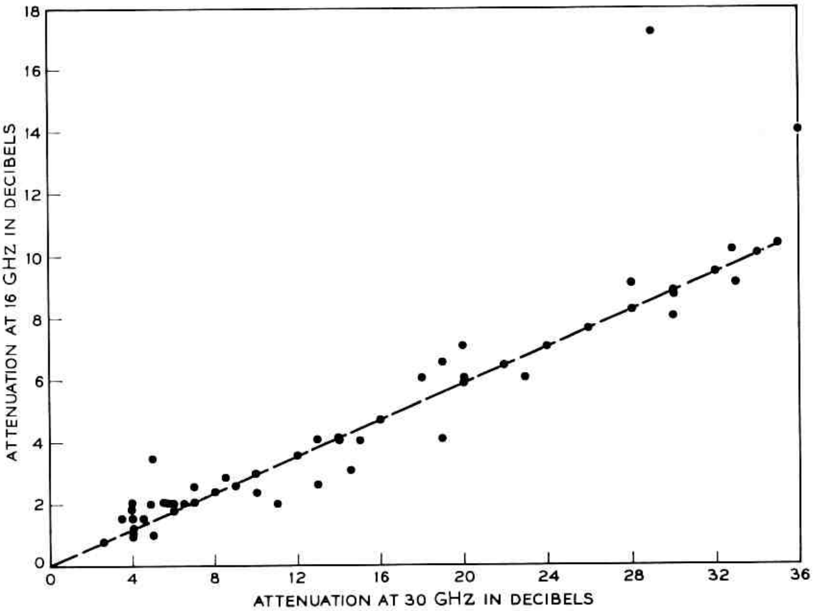


Fig. 8 — Scatter plot of attenuation at 16 GHz against simultaneous attenuation at 30 GHz (December 12, 1967).

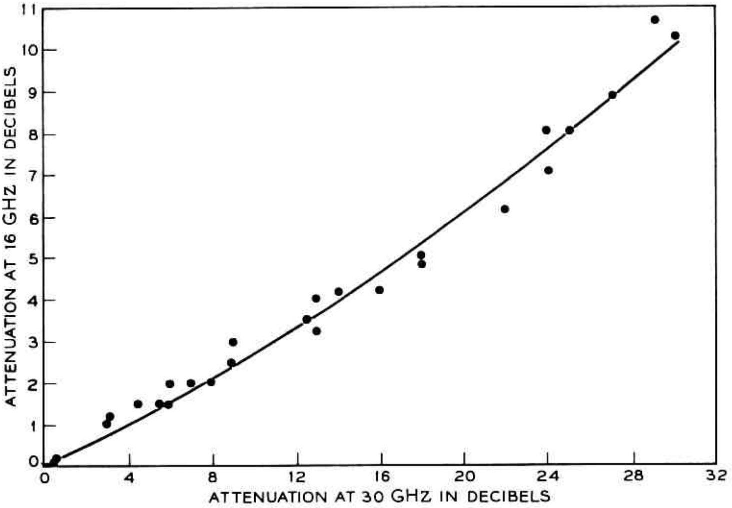


Fig. 9 — Scatter plot of attenuation at 16 GHz against simultaneous attenuation at 30 GHz showing change of ratio with rain rate (August 7, 1968).

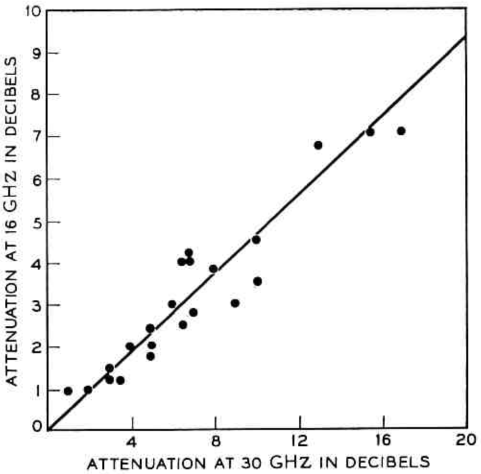


Fig. 10 — Scatter plot of attenuation at 16 GHz against simultaneous attenuation at 30 GHz for rain with noticably large drops (June 3, 1968).

ever, forward scattering will scatter energy from the cold sky into the antenna beam; in other words, it will not contribute to sky brightness in the same way that absorption does. The apparent value of T_c in equations (1) to (3) will be lower than the actual temperature of the absorbing water, consistent with the low values shown in Fig. 12 for small attenuations. The presence of the hot sun in the cold sky does not alter this conclusion since the sun's contribution averaged over the upper hemisphere will be much less than 1°K.

(*v*) In clear weather the transmissivity of the standard atmosphere will, in general, be different in the directions of the sun and of the reference region because of the difference in elevation angle of the two regions. Thus, even when not tracking the sun, the output of the sun tracker is not zero. At 30 GHz it can be as high as 25 dB below the sun at the low elevation cutoff of our observations. This effect does not limit the measuring range of the sun tracker because the false signal is attenuated as the sun is attenuated. If the high attenuation region caused by rain were concentrated near the sun tracker, the false signal from the atmosphere would be attenuated by the same amount as the sun. Also if the high attenuation region had the same temperature as the rest of the atmosphere, its position in the atmosphere would not matter; by the same argument the false signal would be attenuated by the same amount as the sun. In the unlikely case that the high attenuation region is beyond the atmosphere, the effect of the false signal would still be reduced with attenuation, but the maximum reduction would be the ratio of the average atmospheric temperature to the difference between the average atmospheric temperature and the temperature of the high attenuation region. In the actual case, precipitation will at worst be distributed through the atmosphere and have nearly the same temperature as the atmosphere leading to the conclusion that the false signal is not a practical limit to the measuring range of the sun tracker.

(*vi*) Precipitation collecting on the surfaces of the antenna can cause attenuation especially if the radio waves pass through a wetted surface such as a weather cover. This attenuation should not be attributed to the atmosphere. For tests on the effect of water on the surfaces, a fire truck with a fine spray nozzle was used. At a water fall rate of 15 inches per hour, 3 dB attenuation was observed at 30

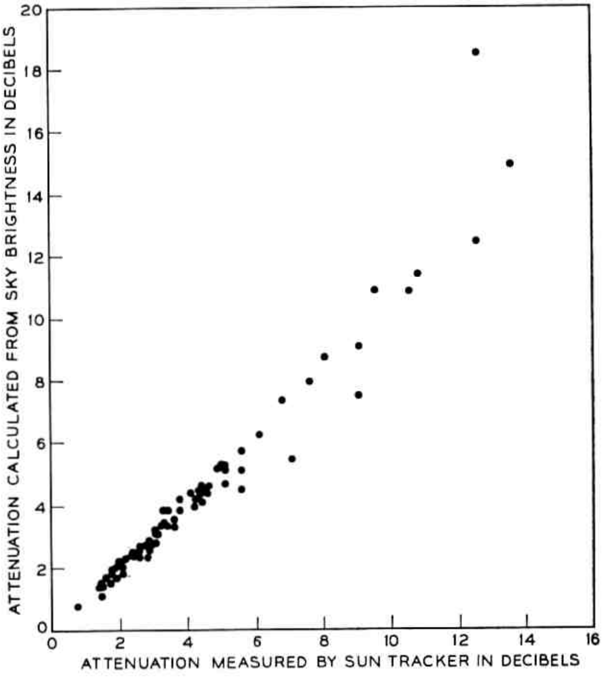


Fig. 11 — Scatter plot of attenuation calculated from sky brightness against the value measured simultaneously using the sun (16 GHz; $T_e = 272°\text{K}$; June 12, 1968).

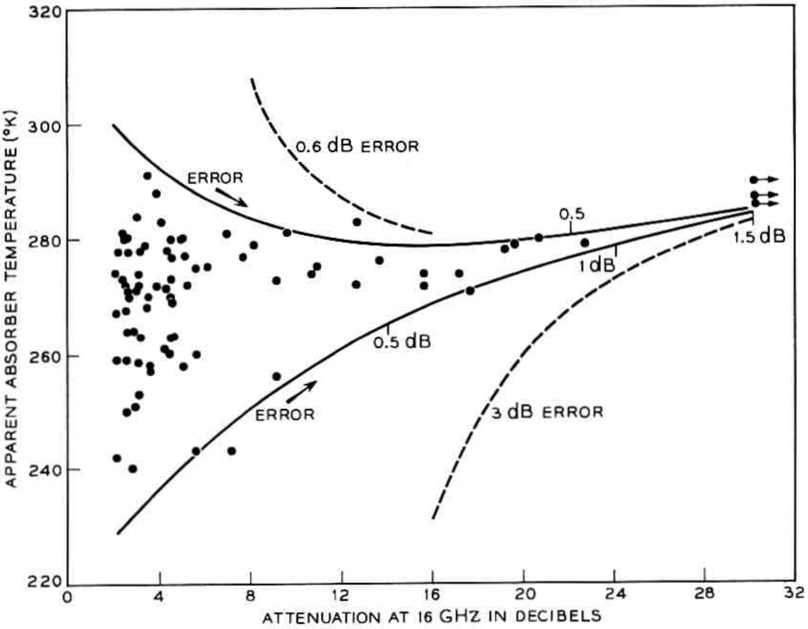


Fig. 12 — Apparent absorber temperature derived from sky brightness measurements plotted against attenuation. The dots are experimental points. The dashed lines show the effect in this plot of 10 dB differences between the attenuation in the direction of the sun and the reference region. The solid lines show the effect of 20 percent differences in attenuation. The resulting error in attenuation measured by the sun tracker is labeled on the curves. A linear change in absorber temperature with attenuation has been assumed in calculating the lines (June 12, 1968).

GHz and 1½ dB at 16 GHz. Snow is also an offender and measurements during snow have been discarded except immediately following removal of the accumulation.*

(*vii*) The high loss region (for example, a raincell) may not be uniform over the 2.6° lobing angle thereby resulting in a difference in brightness between the medium in the direction of the sun and the medium in the direction of the reference region. The main salvation in this case is that the sun produces an antenna temperature much higher than the physical temperature of the atmosphere. If the transmissivity of the medium in front of the sun is t_1 and in front of the reference region t_2, then on subtracting equation (2) from (1) we will have

* However, as discussed in Section V, an inch or so of dry snow on the antenna has only a small effect on antenna temperature.

$$\Delta T_a = t_1 T_s[1 + (t_2/t_1 - 1)(T_c/T_s)].$$

Since T_c/T_s is about 0.1 at both frequencies the maximum over-estimate of t_1 will be about 0.5 db as T_2 becomes much smaller than t_1. In the cast $t_2 > t_1$ the second term in the brackets would begin to dominate if $t_2/t_1 > 10$. Thus t_2 would be measured instead of t_1 for large ratios. Fluctuations of more than 10 dB in the transmissivity of the atmosphere over the 2.6° lobing angle would cause a significant average under-estimate of attenuation in the path to the sun.

The 16 GHz radiometer data plotted in Fig. 12 can be used to estimate the errors in attenuation measurement resulting from actual differences in attenuation. As shown above, a 10 dB excess of attenuation in the direction of the sun would cause about a 3 dB under-estimate of that attenuation. The expected apparent values of T_c in this condition are shown as the lower dashed line in Fig. 12. The upper dashed line indicates the values of T_c expected in the equally probable opposite case where the attenuation in the reference beam is 10 dB greater than in the direction of the sun. In this case the total error is only 0.6 dB. It is seen from Fig. 12 that the data excludes differences which are this great. In both of these cases and in the one to follow, a linear increase of T_c with attenuation has been assumed, namely from 265° at 0 dB to 285° at 30 dB.

A more realistic model of the fluctuations in attenuation is that they are some fraction of the total attenuation. The solid lines in Fig. 12 show the apparent values of T_c with plus and minus 20 percent differences in attenuation. These lines come remarkably close to being envelopes of the scattered points. The error in measured attenuation implied by this model is shown along the lines at 0.5 dB intervals. The maximum error of 1.5 dB out of 30 dB is acceptably small for the type of measurements intended with the sun tracker. The same type of plot has been made for other rains and with 30 GHz data with similar results.

If the temperature of the attenuating medium is T_1 in front of the sun and T_2 in the reference region, but the transmissivity is a constant value t, on subtracting equation (2) from (1) one obtains

$$\Delta T_a = tT_s\left[1 + \frac{1-t}{t}\frac{T_1 - T_2}{T_s}\right].$$

In this case the temperature difference appears linearly in the correction term so that positive and negative errors are equally likely. Thus if there were significant temperature differences over the 2.6° lobing angle, one would see periods of zero or negative receiver output during times of high attenuation. To date, when large attenuations have occurred, this behavior has not been observed and the output has had the same appearance as receiver noise in the absence of signal.

X. ACKNOWLEDGEMENTS

Many people at the Crawford Hill Laboratory have contributed to the sun tracker. However, I wish in particular to acknowledge the contributions of H. W. Anderson who made the major mechanical design; C. A. Burrus and W. M. Sharpless who provided the 30 GHz and 16 GHz down converters, respectively; J. T. Ruscio who did much of the wiring, conducted numerous tests, handled most of the data reduction, and has taken many of the operation and maintenance responsibilities; and D. C. Hogg whose advice and help has been greatly appreciated during many phases of the project.

REFERENCES

1. Tillotson, L. C., "A Model of a Domestic Satellite Communication System," B.S.T.J., *47,* No. 10 (December 1968), pp. 2111–2137.
2. Medhurst, R. G., "Rainfall Attenuation of Centimeter Waves: Comparison of Theory and Measurement," IEEE, Trans. on Antennas and Propagation, *AP-13,* No. 4 (July 1965), pp. 550–564.
3. Hogg, D. C., "Millimeter Wave Communication Through the Atmosphere," Science, *159,* No. 3810 (January 5, 1968), pp. 39–46.
4. Kruger, A. and St. Michael, H., Nature, *206,* No. 4984, (May 8, 1965), pp. 601–602.
5. Hogg, D. C., unpublished work.

High-Power Single-Frequency Lasers Using Thin Metal Film Mode-Selection Filters

By PETER W. SMITH, M. V. SCHNEIDER, and HANS G. DANIELMEYER

(Manuscript received December 12, 1968)

In this paper we present the theory of mode selection by use of a thin metal film in the laser cavity and we derive formulae both by a rigorous method and by using a lumped-circuit approach. Experiments performed with a 500-*mW argon ion laser showed that* 350 *mW or* 70 *percent of the multimode power could be obtained in single-frequency operation using this technique. Somewhat lower efficiencies were obtained with a neodymium-doped yttrium aluminum garnet laser. We compare this with other mode-selection techniques.*

I. INTRODUCTION

Troitskii and Goldina recently showed that a thin metal film can be used inside a He-Ne laser cavity to produce single-frequency output.[1] A thin lossy film will favor oscillation on a mode which has a standing-wave minimum at the film position.

The simplicity of this technique is very attractive. We have, therefore, investigated both theoretically and experimentally its efficiency and its application to high power continuous wave (CW) lasers.

In Section II we develop formulas, both rigorously and using a lumped-circuit approach, which relate the complex refractive index of the metal film to its characteristics as a mode filter. We also show how the complex refractive index of a given metal film can be determined from measurements of the reflectivity and transmissivity of the film. Section III describes experiments using this thin-film technique to obtain single-frequency operation of a continuous wave argon ion laser; Section IV describes the results obtained with a neodymium-doped yttrium aluminum garnet laser. In Section V we discuss these results and the applications of this technique.

II. THIN METAL FILMS FOR MODE-SELECTION FILTERS

2.1 *Optical Properties of Thin Metal Films*

The optical properties of a thin metal film can be characterized by a complex index of refraction and by an effective optical thickness. The parameters which are easily measured are transmittance, reflectance, and average physical film thickness. From these parameters one can deduce the index of refraction and the optical thickness. This procedure does not necessarily lead to meaningful optical constants since thin films often consist of separate islands or of material which is considerably different from the bulk metal because of special problems in the deposition process.

Thin metal films can have high losses if the free space transmittance and reflectance are about equal. This property can be used in optical mode selection filters. The film is placed in the null of the E-field of one particular desired mode which experiences little loss because of the film. Undesired modes with nulls in a different plane are attenuated and hopefully eliminated. Best results are obtained for the thinnest film with the highest complex index of refraction. We require that the film be continuous, that is, that is does not consist of a large number of separate aggregates. (But see Ref. 2.) Chromium and titanium are particularly useful materials because they do not tend to form islands on quartz substrates. Continuous thin films can also be obtained with evaporated nickel-chromium alloys (Nichrome) since the high vapor pressure of chromium leads to fractional distillation during evaporation and consequently gives a base layer of chromium directly on the substrate. A further advantage of Nichrome is its high stability with respect to atmospheric contaminants; Nichrome can also be fully evaporated from a tungsten coil.

2.2 *Computation of Reflectance and Transmittance of Thin Metal Films*

The notation used in the following computation is shown in Fig. 1a. The complex index of refraction of the metal film is $N_1 = N - jK$, and the propagation constant in the film, ρ, is given by

$$\rho = \frac{2\pi N_1}{\lambda} \tag{1}$$

where λ is the wavelength in vacuum.

The amplitude reflection and transmission coefficients r and δ for a film with thickness D are[3]

$$r = \frac{(N_2-N_1)(N_1+N_0)\ \exp\ (j\rho D)+(N_2+N_1)(N_1-N_0)\ \exp\ (-j\rho D)}{(N_2+N_1)(N_1+N_0)\ \exp\ (j\rho D)+(N_2-N_1)(N_1-N_0)\ \exp\ (-j\rho D)} \quad (2)$$

$$\delta = \frac{4N_2N_1}{(N_2+N_1)(N_1+N_0)\ \exp\ (j\rho D)+(N_2-N_1)(N_1-N_0)\ \exp\ (-j\rho D)}. \quad (3)$$

Power reflectance R, transmittance T, and loss A are given by

$$R = rr^* \quad (4)$$

$$T = \frac{N_0}{N_2}\,\delta\delta^* \quad (5)$$

$$A = 1 - R - T. \quad (6)$$

For thin films with $D \ll \lambda$ one can simplify equations (1) and (2) with $\exp\ (\pm j\rho D) = 1 \pm j\rho D$. In addition we let $N_2 = N_0 = 1$ and obtain

$$r = +\frac{j\rho D(1 - N_1^2)}{2N_1 + j\rho D(1 + N_1^2)} \quad (7)$$

$$\delta = +\frac{2N_1}{2N_1 + j\rho D(1 + N_1^2)}. \quad (8)$$

For $|\ N_1^2\ | \gg 1$ one obtains finally

$$r = -\frac{1}{1 + \dfrac{2}{j\rho DN_1}} \quad (9)$$

$$\delta = +\frac{1}{1 + \dfrac{j\rho DN_1}{2}}. \quad (10)$$

This means that the thin film with $|\ N_1^2\ | \gg 1$ can be characterized by one single physical parameter

$$j\rho DN_1 = j\,\frac{2\pi D}{\lambda}\,N_1^2. \quad (11)$$

These approximations are appropriate when considering infrared wavelengths. For other cases one has to use the rigorous expressions of equations (2) and (3).

It is often useful to derive an equivalent lumped-film admittance Y based on equations (2) and (3) or equations (7) and (8). The lumped-film admittance Y is shown in Fig. 1b in a transmission line with admittance Y_2 and Y_0. The reflection and transmission coefficients are

$$r = \frac{Y_2 - Y - Y_0}{Y_2 + Y + Y_0} \tag{12}$$

$$\delta = \frac{2Y_2}{Y_2 + Y + Y_0}. \tag{13}$$

For $Y_2 = Y_0 = 1$ one obtains

$$r = -\frac{1}{1 + \frac{2}{Y}} \tag{14}$$

$$\delta = +\frac{1}{1 + \frac{Y}{2}}. \tag{15}$$

An expression for Y can be obtained from equations (2) or (3). A good approximation valid for thin absorbing films can be derived by using equation (7)

$$r = +\frac{j\rho D(1 - N_1^2)}{2N_1 + j\rho D(1 + N_1^2)} = -\frac{1}{1 + \frac{2}{Y}}. \tag{16}$$

The result for Y is

$$Y = \frac{j\rho D(N_1^2 - 1)}{N_1 + j\rho D}. \tag{17}$$

The rigorous as well as lumped approach have been used for computing transmittance, reflectance, and loss of the film as listed in Table I. The film was one used in the experiments reported in Sections III and IV. The transmittance and reflectance were measured with a traveling-wave beam external to the laser cavity. By successive trials, the value of N_1 was found which gave the best fit to the experimental measurements. One can conclude that this film is thin and lossy enough for using the lumped model.

2.3 *Thin Metal Film in Front of a Mirror*

A thin film in front of a mirror is shown in Fig. 1c. The high reflectance mirror can be considered as a short or open circuit which is spaced by a length L from the back end of the thin metal film. The reflectance and loss can be computed from a rigorous expression derived from equations (2) and (3) or from the approximate model

Table I—Measured and Computed Thin Film Properties

Film material: Nichrome (80% nickel, 20% chromium)
Film thickness: 50 Å
Wavelength: 10645 Å

	Transmittance T	Reflectance R	Loss A
$N_2 = 1.5$ quartz; $N_0 = 1.0$ air			
Measured	0.78	0.01	0.21
Computed, eqs. (2) and (3), $N - jK = 1.66 - j\cdot 2.83$	0.773	0.0116	0.214
Computed, eqs. (2) and (3) $\exp(\pm j\rho D) = 1 \pm j\rho D$	0.776	0.0115	0.212
Computed, lumped admittance, eq. (17), $Y = 0.272 - j\cdot 0.193$	0.777	0.0115	0.211
$N_2 = 1.0$ air; $N_0 = 1.5$ quartz			
Measured	0.78	0.08	0.14
Computed, eqs. (2) and (3), $N - jK = 1.66 - j\cdot 2.83$	0.773	0.0832	0.143
Computed, eqs. (2) and (3) $\exp(\pm j\rho D) = 1 \pm j\rho D$	0.776	0.0833	0.140
Computed, lumped admittance, eq. (17), $Y = 0.272 - j\cdot 0.193$	0.777	0.0819	0.140

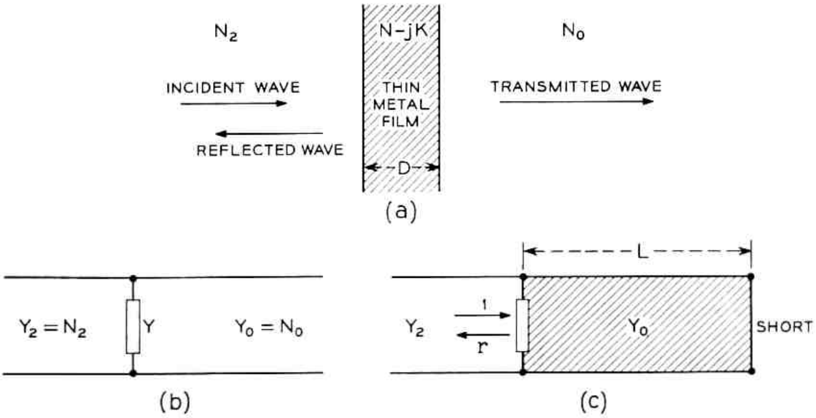


Fig. 1 — Notation used for computing optical film properties: (a) thin metal film with index $N_1 = N - jK$ and thickness D; (b) lumped admittance, $Y = j\rho D(N_1^2 - 1)/(N_1 + j\rho D)$, and (c) admittance Y in front of mirror.

with a lumped admittance Y. The rigorous result for r is

$$r = \frac{E + F}{G + H} \tag{18}$$

where

$$E = (N_2 - N_1)[(N_1N_0 + N_0^2)e^{i\beta} + (N_1N_0 - N_0^2)e^{-i\beta}]e^{i\rho D} \tag{19}$$

$$F = (N_2 + N_1)[(N_1N_0 - N_0^2)e^{i\beta} + (N_1N_0 + N_0^2)e^{-i\beta}]e^{-i\rho D} \tag{20}$$

$$G = (N_2 + N_1)[(N_1N_0 + N_0^2)e^{i\beta} + (N_1N_0 - N_0^2)e^{-i\beta}]e^{i\rho D} \tag{21}$$

$$H = (N_2 - N_1)[(N_1N_0 - N_0^2)e^{i\beta} + (N_1N_0 + N_0^2)e^{-i\beta}]e^{-i\rho D} \tag{22}$$

and

$$\beta = \frac{2\pi N_0 L}{\lambda}. \tag{23}$$

Reflectance obtained from equations (18) to (23) for a 50-Å Nichrome film at $\lambda = 10645$ Å, and a 150-Å film at $\lambda = 5145$ Å is plotted in Fig. 2. The reflectance is plotted as a function of $\Delta\beta/2\pi$ where $\beta = \pi/2 + 2\pi n + \Delta\beta$ and n is an integer. As β can be varied either by changing the film-to-mirror spacing or by changing the frequency of incident radiation, we have written $\Delta\beta = N_0\Delta L/\lambda$ or $N_0L\Delta\nu/c$. The values of N_0 and N_2 are chosen to correspond to the experimental situations in which the films are used. Of particular importance are the minimum and the maximum absorption listed in Table II. Note that if the film has no loss and $N_0 = N_1 = N_2$, the reflectivity is 1 regardless of the value of L or λ.

The data are based on the assumption that the quartz substrate is lossless and that the surface roughness of the substrate is much less than the listed film thickness.

Reflectance and loss can also be computed from the thin-film equivalent circuit of Fig. 1c. The metal film is characterized by the lumped admittance Y, the mirror by a short, and the distance between mirror and film by the effective optical length N_0L. The short is transformed into a susceptance Y_S in parallel with Y given by

$$\frac{Y_S}{Y_0} = -j \cot \frac{2\pi N_0 L}{\lambda}. \tag{24}$$

The reflection coefficient r is

$$r = \frac{Y_2 - Y - Y_S}{Y_2 + Y + Y_S}. \tag{25}$$

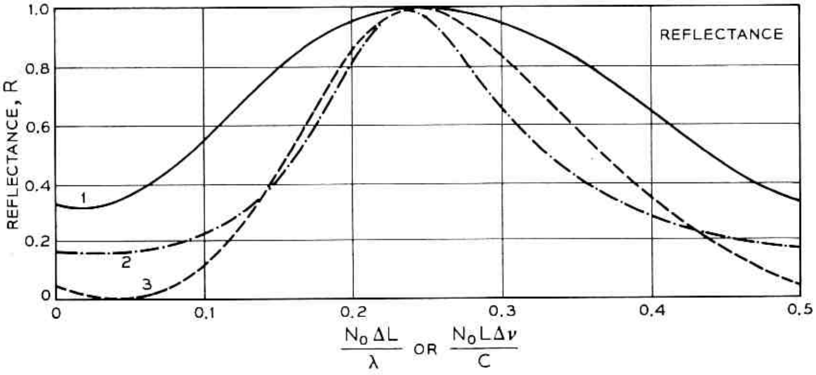


Fig. 2 — Reflectance for mode selectors using 50-Å or 150-Å Nichrome films on quartz substrates as a function of their spacing from the high reflectivity mirror. Notice that the values of N_0, N_2, and λ are chosen to correspond to the experimental circumstances under which each film was used.

Curve	Ni − Cr (Å)	λ (Å)	N_1	N_0	N_2
1	50	10645	$1.66 - j\cdot 2.83$	1.5	1.0
2	150	5145	$1.33 - j\cdot 1.30$	1.0	1.5
3	100	10645	$2.4 - j\cdot 3.5$	1.5	1.0

For the special case $Y_2 = 1$ (air) and $Y_0 = 1.5$ (quartz) one obtains

$$r = \frac{1 - Y + 1.5j \cot(\pi L/\lambda)}{1 + Y - 1.5j \cot(\pi L/\lambda)}. \tag{26}$$

The lumped admittance model always gives a minimum loss of zero because it is based on a limit process in which the film thickness approaches zero while the product $Y \approx j\rho D N_1$ remains constant.

It is clear from the form of equation (25) that the maximum reflectivity comes for $2\pi N_0 L/\lambda = n\pi$ where n is an integer. The fre-

Table II—Minimum and Maximum Loss for Nichrome Films on Quartz

Film Thickness	Wavelength	Index of refraction			Minimum loss	Maximum loss
		$N_1 = N - jk$	N_0	N_2		
50 Å	10645 Å	$1.66 - j\cdot 2.83$	1.5	1.0	8.5×10^{-5}	0.680
150 Å	5145 Å	$1.33 - j\cdot 1.30$	1.0	1.5	1.3×10^{-2}	0.841

quency spacing between reflectance peaks is just $c/(2N_0L)$. Thus for single-frequency operation the film should be situated sufficiently close to the laser cavity end mirror that $c/(2N_0L)$ is greater than the oscillation width of the laser medium. It is advantageous to make N_0L as large as possible without exceeding this requirement, however, since the selectivity of the mode filter decreases as N_0L is decreased.

The minimum reflectance and the shape of the filter curve in the vicinity of the maximum reflectance are important parameters governing the mode selection properties of the filter. We separate the admittance Y into a real and imaginary part

$$Y = G + jB \tag{27}$$

and obtain from equations (4) and (26) for the reflectance R

$$R = 1 - \frac{4G}{(G+1)^2 + \left(B - 1.5 \cot \dfrac{\pi L}{\lambda}\right)^2}. \tag{28}$$

The minimum reflectance occurs in the vicinity of the nulls of the cotangent function. Close to maximum reflectance, $\cot (\pi L/\lambda) \gg 1$ and we obtain from equation (28)

$$R = 1 - 1.78G \tan^2 \frac{\pi L}{\lambda}. \tag{29}$$

For rigorous computations one has to use equations (18) to (23). Filter curves based on rigorous equations with $N_1 = N - jK$ as a parameter are shown in Fig. 3. The film thickness for all curves is $D = 150$ Å and the wavelength $\lambda = 5145$ Å.

Notice that we have assumed plane waves in all of these calculations. In practice, if the flat metal film is situated close to a plane laser end mirror, this condition will be well satisfied. If a plane metal film must be situated some distance from the laser end mirrors, the laser cavity must be designed so that there will be a beam waist at the metal film.

The problems encountered in practical filter design are often that films with suitable index of refraction are not stable or *vice versa*. Additional protective coatings have to be deposited which may change the filter characteristics, or a compromise has to be found with one single stable film or two stable films spaced at an appropriate distance inside the laser cavity.

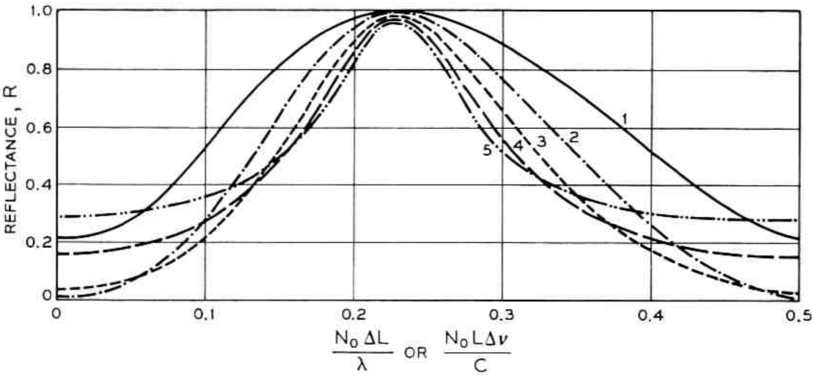


Fig. 3 — Reflectance of a mode selector using a 150-Å metal film for various values of $N_1 = N - jK$: (1) $1 - j$, (2) $1.5 - j \cdot 1.5$, (3) $2 - j \cdot 2$, (4) $2.5 - j \cdot 2.5$, and (5) $3 - j \cdot 3$. The curves are plotted for $\lambda_0 = 5145$ Å, $N_0 = 1.5$, and $N_3 = 1.0$.

2.4 *Film Fabrication*

Films for use as mode selection filters are evaporated from a tungsten coil in a vacuum of $4 \cdot 10^{-7}$ torr. The coils are made from 4-strand tungsten wire with a diameter of 0.015-inch per strand. The source to substrate distance for a 150-Å film is 3.0 inches and the Nichrome charge is a 0.010-inch diameter wire with a length of 0.854 inch. Total evaporation time is less than 10 seconds. The substrates are cleaned in isopropyl alcohol, immersed in methanol, and blow dried with dry nitrogen. It is concluded from separate experiments with a Tolansky interferometer that the Nichrome material is completely evaporated from the tungsten coil.

III. ARGON ION LASER EXPERIMENTS

Experiments were performed with a dc-excited discharge tube with an active plasma length of 60 cm and a 3-mm diameter bore. A Brewster-angle prism was used inside the cavity to select the desired laser transition; the cavity consisted of a 5-m mirror and a flat mirror separated by 150 cm. The metal film was situated 2 cm from the flat mirror. With this configuration, no adjustable aperture was required to obtain fundamental transverse mode operation.

Films of pure nickel or Nichrome were deposited on one side of a fused quartz plate. These plates were of optical quality suitable for their use as Brewster-angle windows for laser tubes. In order to eliminate the effects of Fabry-Perot interferences between the front and back surfaces of the plates, an additional plate with an antireflection coating on one side was contacted with optical matching oil to the bare surface of the metal-coated plate.

The measurements reported here were made using this composite plate with an antireflection coating on one side and the metal film on the other. Virtually the same laser output power was observed when the simple plate with a metal film on one side, and no coating on the other, was used in the laser cavity.

Several films of different thickness of nickel and Nichrome were used for these experiments. The best results were obtained with a 150-Å Nichrome film. Because details of the deposition technique may affect the properties of the film obtained it is perhaps more informative to list the characteristics of the film measured with an external (traveling-wave) beam. These were

$$T = 0.60 \pm 0.01, \qquad R = 0.13 \pm 0.005, \qquad A = 0.27 \pm 0.01$$

for the beam incident on the metal film and

$$T = 0.61 \pm 0.01, \qquad R = 0.013 \pm 0.002, \qquad A = 0.38 \pm 0.01$$

for the beam incident on the antireflection coating. These measurements were made at 5145 Å. Virtually the same results were obtained at 4880 Å. These results were used to find the complex index of refraction used for the calculations in Section II. Figure 4a shows the laser output at 4880 Å as a function of the distance between the metal film and the end mirror of the laser cavity. This distance was varied by a ramp voltage applied to a piezoelectric ceramic transducer element on which the laser mirror was mounted. As the relative film position is varied, different longitudinal modes of the laser find themselves with a standing-wave minimum at the metal film and thus are able to oscillate. The overall outline of the pattern indicates the profile of the gain curve. The side humps are caused by the axial magnetic field applied to the laser tube.

In order to verify that we had indeed achieved single-frequency operation, a scanning interferometer was set up to observe the frequency spectrum of the laser output. Figure 4b shows the output versus frequency for the laser operating without a metal film in the cavity. This picture corresponds to the maximum available output of 500 mW at 4880 Å. With a 150-Å Nichrome film in the laser cavity,

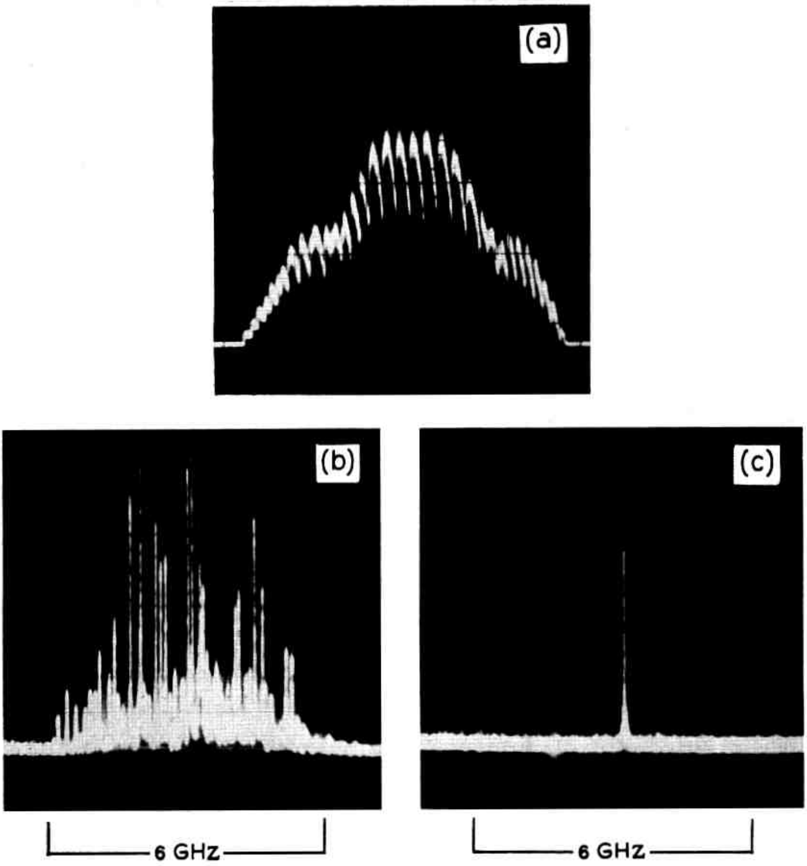


Fig. 4 — Experimental results using metal-film mode selector in an argon ion laser oscillating at 4880Å: (a) Single-frequency laser output as a function of separation between metal film and laser end reflector; (b) Multimode output of laser without mode selector as a function of frequency: total output power 500 mW. The total oscillation bandwidths is ≈6 GHz; (c) Single mode laser output obtained using metal film as a function of frequency: output power 350 mW.

single-frequency output was obtained at 4880 Å as shown in Fig. 4c. Over 350 mW or 70 percent of the multimode output could be obtained in a single frequency. At 5145 Å, 50 percent of the multimode power could be obtained in a single frequency, using the same film.

These figures can be compared with those for an interferometric mode selector of the type described in Ref. 4. A mode selector of that type was constructed for use with the argon ion laser. It was found that with the same type of laser 50 percent of the multimode output power at 5145 Å could be obtained in single-frequency output; 70 percent of the multimode output power at 4880 Å could be obtained in single-frequency output. Thus, for this laser, the two schemes appear identical in power output.

From curve 2 of Fig. 2 we find that the loss produced by the metal film is less than the laser gain ($\approx$ 25 percent) for a frequency range of roughly 1.3 GHz. The fact that single-frequency operation was obtained with this film indicates that mode-competition effects must have extended over this range of frequencies. This is not surprising as the natural linewidth for the argon laser is about 500 MHz and radiation broadening will increase this homogeneous linewidth in a laser well above threshold.[5] Mode competition effects are expected between adjacent modes spaced by less than the homogeneous linewidth. Thus we see that, for the argon laser, single-frequency operation can be obtained with a much lower selectivity mode selector than would be required if mode competition were not present.

IV. *Nd:YAG* LASER EXPERIMENTS

4.1 *Description of the Laser*

The laser system consisted of a 30- by 2.5-mm neodymium: yttrium aluminum garnet (Nd:YAG) rod, pumped with a 1-kW tungsten lamp in an elliptic cylinder, a high reflectivity plane mirror, and an output mirror with 10-m curvature and 1.6 percent transmission. The mirror separation was $M = 20$ cm which resulted in a longitudinal mode spacing of 670 MHz. Without insertion of any mode selector, this cavity configuration gave fundamental transverse mode operation up to 850-W pump power. Figure 5a shows the output spectrum at that pump level observed with a scanning Fabry-Perot interferometer. The total output power was 200 mW with a maximum linearly polarized component of 130 mW. This component increased to 220 mW at 960-W pump power, but this power was not all in the fundamental mode, and the amplitudes of individual modes were very unstable.

To obtain single-frequency operation, the plane mirror was replaced by a fused silica flat 2.5 mm thick (free spectral range 40 GHz) which was high-reflectivity coated on one side and metal coated on the other. This arrangement produced stable single-frequency output, as evidenced by Fig. 5b, which was photographed from a screen averaging over 10 scans in one second (the persistance time of the screen). The output stability was achieved by keeping one particular node of one

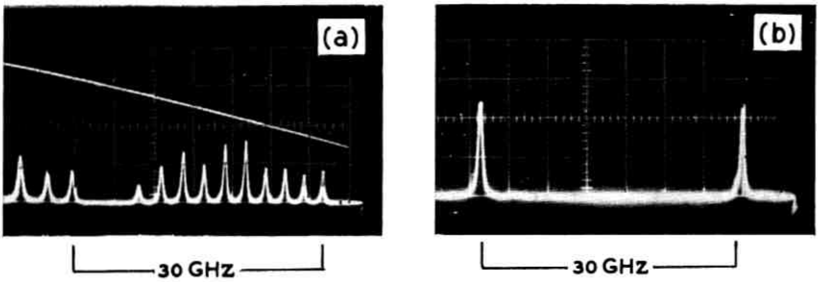


Fig. 5 — Experimental results using metal-film mode selector in neodymium-doped YAG laser oscillating at 1.06μ. (a) Multimode output of laser without mode selector as a function of frequency. Linearly polarized output power 130 mW. The frequency spacing between modes corresponds approximately to $c/2L$ for the YAG rod. The upper trace shows the ramp voltage used to scan the Fabry-Perot interferometer. The output spectrum display repeats itself with the spacing of the interferometer free spectral range (30 GHz). The total oscillation bandwidth of the laser shown here is ≈22 GHz. (b) Single mode laser output obtained using metal film as a function of frequency: linearly polarized output power 60 mW.

longitudinal mode close to the film center. Their relative positions should be within about ±20 A since the nodes corresponding to adjacent longitudinal modes are spaced at $N_0L\lambda/2M = 84$ Å in the vicinity of the film. This implies temperature stabilization of the quartz flat to within ±0.1°C [$\partial(N_0L)/L\partial T = 8 \times 10^{-6}$/°C], and cavity length stabilization to within ±1,000 Å. In addition, the film becomes inefficient if its tilt exceeds one adjacent longitudinal node spacing across the beam diameter. Thus the quartz flat must be parallel to about 2 seconds of arc and its flatness should be better than $\lambda/20$.

4.2 *Results*

Best results at lowest threshold were obtained with a 50-Å nickel-chromium film. Its transmission and reflection, as measured with a YAG laser beam, are shown in Table I. A power of 60 mW (maximum linearly polarized component) could be obtained in a single frequency which was 27 percent of the maximum multimode power output. However, the absorption of the film was not sufficient to obtain single-frequency output up to the pump limit. At 960 watts pump power, the total (multimode) output power was 150 mW with a frequency range of 4 GHz. Curve 1 of Fig. 2 predicts a 4 GHz range for a net gain of 3 percent ($c/N_0L = 80$ GHz). In addition it was observed that the output was much more stable than that of the free-running laser: the power in an individual mode was constant to within 20 percent.

Therefore, inserting a metal film into the cavity is a simple technique for obtaining a stable YAG laser output in a narrow frequency range.

Several other films have been tried. A 100-Å nickel-chromium film, for instance, had close to zero minimum reflectance according to curve 3 of Fig. 2. This curve was calculated from the transmission (48 percent) and reflections (18 percent and 32 percent) measured with a YAG laser beam. It was verified that the minimum loss was larger than for the 50-Å film since the threshold for oscillation had increased to 800 W pump power (compared with 600 W for the 50-Å film). Although curve 3 in Fig 2 indicates a higher selectivity for this film, the maximum single-frequency output was again 60 mW. If the pump power was increased beyond this point, multimode operation was obtained.

Obviously, the metal film technique works less efficiently for the YAG laser than for the argon ion laser. The reasons for this are (*i*) the apparent lack of mode competition which makes it necessary for the YAG laser to completely suppress all but one mode (not just to provide a little more loss for the other modes), and (*ii*) its low gain which makes the YAG laser output very sensitive to small additional losses. Therefore, it is generally much more difficult to obtain a high-power single-frequency output from a YAG laser than from an argon ion laser or helium-neon laser.

V. DISCUSSION AND CONCLUSION

The theoretical and experimental results show that it is possible, under suitable circumstances, to obtain high-power single-frequency operation of a laser using the metal film technique. In practice, however, it is not always possible to find a material that has the required loss in a sufficiently thin film. This is in contrast with interferometric mode selectors whose selectivities are determined simply by the reflectivities of the elements. For the argon ion laser an interferometric mode selector has some advantages over the metal film;[4] however, it is difficult to apply to the YAG laser because of its much greater oscillation width (about 100 GHz). The metal film method described here does have the advantage of simplicity, however, and the system is relatively easy to make mechanically stable. The metal film technique should be of particular interest to people working in the fields of Brillouin scattering or holography where a narrow bandwidth source is required. It is relatively easy with a metal film to restrict the laser oscillation to a few neighboring modes. Thus a drastic reduction in bandwidth can be made, often at little expense in total output power.

VI. ACKNOWLEDGMENTS

We wish to thank Mr. P. J. Maloney and Mr. J. M. Barro for technical assistance, Mr. S. Shah for evaporating the metal films, and Mr. D. W. Doughty for making the dielectric coatings.

REFERENCES

1. Troitskii, Yu. V. and Goldina, N. D., "Separation of One Mode of a Laser," J. Experimental and Theoretical Phys. Letters, *7,* (January 30, 1968) pp. 36–38.
2. Troitskii, Yu V. and Goldina, N. D., "Thin Scattering Film in the Field of a Standing Wave of Optical Frequencies and its Use in Selecting Modes of an Optical Resonator," Opt. Spectroscopy, *25,* No. 3 (September 1968), pp. 255–256.
3. Wolter, H., "Optik Dünner Schichten," *Handbuch der Physik,* Springer-Verlag, Berlin, 1956, *24,* pp. 461–473.
4. Smith, P. W., "Stabilized, Single-Frequency Output From a Long Laser Cavity," IEEE J. Quantum Elec., *QE1,* No. 11 (November 1965) pp. 343–348; and "On the Stabilization of a High-Power Single-Frequency Laser," IEEE J. Quantum Elec., *QE2,* No. 9 (September 1966) pp. 666–668.
5. Webb, C. E., Miller, R. C., and Tang, C. L., "New Radiative Lifetime Values for the 4s Levels of AII," IEEE J. Quantum Elec. *QE4,* (May 1968), p. 357.

Microstrip Lines for Microwave Integrated Circuits

By M. V. SCHNEIDER

(Manuscript received December 2, 1968)

Microstrips, transmission lines of metallic layers deposited on a dielectric substrate, are very useful for the microwave and millimeter wave hybrid integrated circuits required for solid-state radio systems because of their simplicity and planar structure. To design hybrid integrated circuits with microstrips requires computation or measurement of the impedance, the attenuation, the guide wavelength, and the unloaded Q *of the line. These parameters can be obtained from the effective dielectric constant and the characteristic impedance of the corresponding air line. This paper gives the exact design data for all line parameters for the most important cases.*

We report the impedance and attenuation measurements performed on microstrips. Satisfactory agreement is obtained with theoretical results based on conformal mapping with logarithmic derivatives of theta functions and expressions involving the partial derivatives of the impedance with respect to independent line parameters.

I. INTRODUCTION

Transmission lines and passive lumped or distributed circuit elements, which are manufactured and assembled from planar metal conductors or conducting stripes on insulating substrates, are essential basic elements in microwave and millimeter wave hybrid integrated circuits. The metal strips or microstrips are deposited by thin-film or thick-film technology on dielectric substrates; the processing steps are substantially different compared to conventional coaxial and waveguide circuit technology. Circuits built with microstrip transmission lines or microstrip components have three important advantages:

(*i*) The complete conductor pattern can be deposited and processed on a single dielectric substrate which is supported by a single metal ground plane. Such a circuit can be fabricated at a substantially lower cost than waveguide or coaxial circuit configurations.

(*ii*) Beam-leaded active and passive devices can be bonded directly to metal stripes on the dielectric substrate.

(*iii*) Devices and components incorporated into hybrid integrated circuits are accessible for probing and circuit measurements (with some limitations imposed by external shielding requirements).

The purpose of this paper is to derive formulas for the electric parameters which are the impedance, attenuation, propagation constant, and unloaded Q of the microstrip transmission line. In addition to the electrical design data, attenuation measurements at 30 GHz are presented because:

(*i*) The attenuation is the most important electrical parameter of a microstrip because it determines the circuit losses of microwave and millimeter wave hybrid integrated circuits.

(*ii*) There are many solid-state radio systems for which hybrid integration looks attractive, such as the radio pole line, high-capacity domestic-satellite systems, *Picturephone*® visual telephone distribution, and mobile telephone systems.[1,2] Hybrid integration of circuits is essential for many other applications in order to achieve small overall size, minimum weight, and low production cost.

II. DEFINITION AND CLASSIFICATION

A strip line or microstrip line is a parallel two-conductor line made of at least one flat strip of small thickness. For mechanical stability the strip is deposited on a dielectric substrate which is usually supported by a metal ground plane. This basic configuration is shown in Fig. 1a.

A parallel two-conductor line of this type may need modification because:

(*i*) A radio frequency shield may be required to eliminate radiation losses. The shield dimensions or the sheet conductivity of the shielding material have to be chosen in such a way that excitation of transverse electric modes, transverse magnetic modes, and box resonances is suppressed.

(*ii*) Proximity of the air-dielectric interface with the strip conductor can lead to excitation of plane-trapped surface waves. This problem can be solved by using a substrate with a low dielectric constant or by choosing a sufficiently small frequency-thickness product

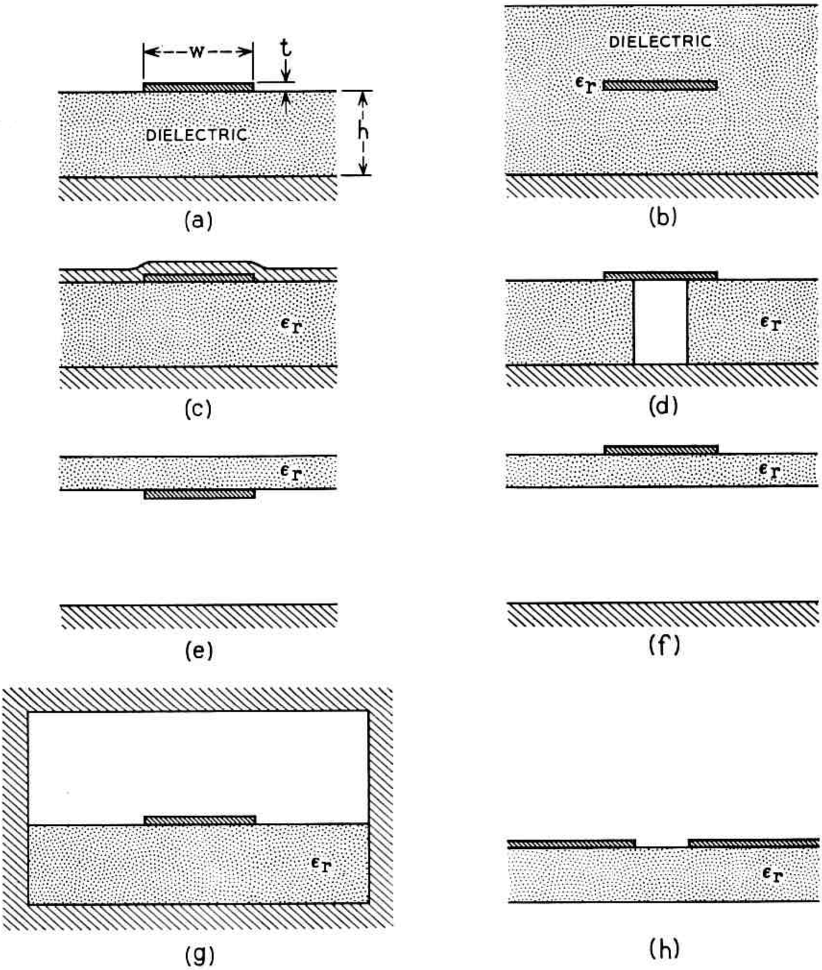


Fig. 1 — Basic types of microstrip transmission lines with one strip conductor supported by a dielectric substrate: (a) standard microstrip, (b) embedded microstrip, (c) microstrip with overlay, (d) microstrip with hole, (e) standard inverted microstrip, (f) suspended microstrip, (g) shielded microstrip, (h) slot transmission line.

for the microstrip. It can also be solved by removing the air-dielectric interface into the far field region as shown in Fig. 1b.

(*iii*) If the substrate is a semiconductor, surface passivation may be necessary to protect against atmospheric contaminants. This can be achieved by a thin dielectric film as shown in Fig. 1c.

(*iv*) Solid-state devices with substantial heat dissipation such as IMPATT, GUNN, and LSA diodes as well as high-power varactor diodes have to be shunt mounted in the microstrip in order to achieve a small thermal spreading resistance in the ground plane. A hole in the dielectric is required in Fig. 1d for mounting a solid state device between the two microstrip conductors.

IMPATT diodes, bulk sources, and high-power varactors are typical examples of solid-state devices which are usually shunt mounted in transmission line circuits. Other solid-state devices or materials which require shunt mounting are ferrites for circulators and isolators and high-Q dielectric resonators for microwave band-pass filters. Shunt mounting is facilitated in inverted microstrips and suspended microstrips shown in Figs. 1e and 1f. Solid-state devices which require a dc bias or a dc return have to be mounted by means of a pressure contact or bonded contacts between the ground plane and the strip conductor shown in Fig. 1e. Complete shielding of such a line is essential because fringe field effects are enhanced by increased electric field intensities in the dielectric support material. An attractive solution is to suspend the substrate symmetrically between the ground plane and the top shield. Such lines have been discussed by Brenner and have been used for balanced transistor amplifiers and ferrite circulators.[3–6] A major advantage of all microstrip configurations with an air gap is that the effective dielectric constant is small. This means that the effective dielectric loss tangent is substantially reduced; also, all circuit dimensions can be increased, which leads to less stringent mechanical tolerances, better circuit reproducibility, and therefore lower production cost.

Figure 1g shows a completely shielded standard microstrip and Fig. 1h is a schematic diagram of a slot line which consists of two conductors deposited on the same side of a high permittivity substrate.[7] The slot line can be tightly coupled to the lines of Figs. 1a through g by depositing the slot line metallization on one side of the substrate and the microstrip conductor on the opposite side of the same substrate. Standard microstrips supporting transverse electromagnetic modes and structures supporting slot modes can thus be combined on one single substrate for obtaining the widest possible choice of circuits to be built with existing hybrid integrated circuit technology.

III. IMPEDANCE, ATTENUATION, AND UNLOADED Q

The electrical parameters of the microstrips of Figs. 1a through g which are required for circuit design are impedance, attenuation, unloaded Q, wavelength, and propagation constant. These parameters are interrelated for all microstrips of Figs. 1a through g assuming that

(*i*) The propagating mode is a transverse electromagnetic mode, or it can be approximated by a transverse electromagnetic mode.

(*ii*) Conductor losses in the metal strips are predominant, which means dielectric losses can be neglected.

(*iii*) The relative magnetic permeability of the substrate material is $\mu_r = 1$.

The basic reason for the subsequently explained relationship of the line parameters is that the inductance per unit length depends only upon the conductor geometry and is absolutely independent of the geometry and the dielectric properties of the supporting structure. The relationship between line parameters is shown in Fig. 2.

Let us assume in Fig. 2a that the conductor geometry is defined by a stripe width w_o, a ground plane spacing h_o, and a small stripe thickness t_o. Let us also assume that this is an air line with a characteristic impedance Z_o, a wavelength λ_o, an attenuation per unit length α_o, and an unloaded Q_o. If the conductor dimensions remain the same, and if the microstrip is fully embedded in a dielectric medium with a relative dielectric constant ϵ_r, one obtains the new line parameters given in Fig. 2b. If the line is only partially filled with dielectric support material with a relative dielectric constant ϵ_r, one obtains for the line parameters of Fig. 2c

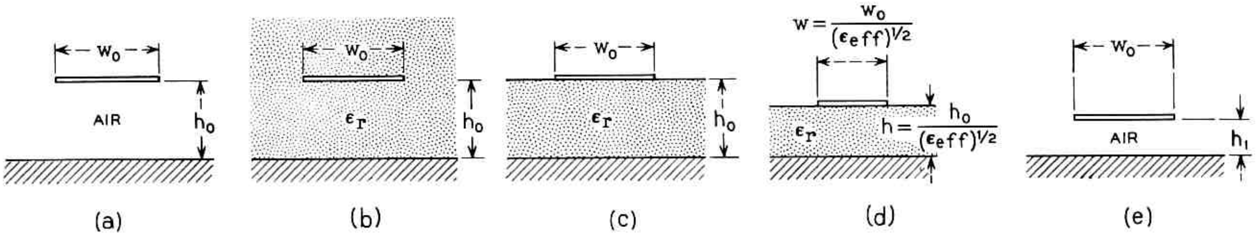
$$Z = \frac{Z_o}{(\epsilon_{eff})^{\frac{1}{2}}} \qquad \text{impedance} \tag{1}$$

$$\lambda = \frac{\lambda_o}{(\epsilon_{eff})^{\frac{1}{2}}} \qquad \text{wavelength} \tag{2}$$

$$\alpha = (\epsilon_{eff})^{\frac{1}{2}}\alpha_o \quad \text{attenuation} \tag{3}$$

$$Q = Q_o = \frac{20\pi}{\ln 10}\frac{1}{\alpha_o\lambda_o} \qquad \alpha_o\lambda_o \text{ in dB.} \tag{4}$$

The effective dielectric constant ϵ_{eff} has to be computed or measured



Z_0 IMPEDANCE	$\frac{Z_0}{(\epsilon_r)^{1/2}}$	$\frac{Z_0}{(\epsilon_{eff})^{1/2}}$	$\frac{Z_0}{(\epsilon_{eff})^{1/2}}$	$\frac{Z_0}{(\epsilon_{eff})^{1/2}}$
λ_0 WAVELENGTH	$\lambda = \frac{\lambda_0}{(\epsilon_r)^{1/2}}$	$\lambda = \frac{\lambda_0}{(\epsilon_{eff})^{1/2}}$	$\lambda = \frac{\lambda_0}{(\epsilon_{eff})^{1/2}}$	$\lambda = \lambda_0$
α_0 ATTENUATION	$\alpha = (\epsilon_r)^{1/2} \alpha_0$	$\alpha = (\epsilon_{eff})^{1/2} \alpha_0$	$\alpha = \epsilon_{eff} \cdot \alpha_0$	$\alpha = (\epsilon_{eff})^{1/2} \alpha_0$
$Q_0 = \frac{20\pi}{\ln 10} \frac{1}{\alpha_0 \lambda_0}$	$Q = Q_0$	$Q = Q_0$	$Q = \frac{Q_0}{(\epsilon_{eff})^{1/2}}$	$Q = \frac{Q_0}{(\epsilon_{eff})^{1/2}}$

Fig. 2 — Impedance, wavelength, attenuation, and unloaded Q of microstrip transmission lines.

as discussed in Section 4.3. The following inequalities are valid for the standard microstrip in Fig. 1a and the inverted microstrip of Fig. 1e

$$\frac{1+\epsilon_r}{2} \leqq \epsilon_{eff} \leqq \epsilon_r \qquad \text{standard microstrip} \tag{5}$$

$$1 \leqq \epsilon_{eff} \leqq \frac{1+\epsilon_r}{2} \qquad \text{inverted microstrip.} \tag{6}$$

If one has to compare the attenuation or the unloaded Q of different microstrips one has to consider lines which have the same impedance level. It is also necessary that the electrical length of both or at least one critical conductor dimension w or h of Fig. 2 is the same. By critical conductor dimension we mean the dimension which is more critical with respect to excitation of transverse electric modes or transverse magnetic modes. Plane-trapped surface waves or hybrid modes are not considered in this comparison.

Figure 2d gives the line parameters for partial dielectric filling with reduced dimensions $w = w_o/(\epsilon_{eff})^{1/2}$ and $h = h_o/(\epsilon_{eff})^{1/2}$. This insures that the electrical dimension of the two basic line parameters is the same as the electrical dimension of the air line of Fig. 2a. In order to obtain the same impedance for the partially filled microstrip of Fig. 2d and the air line one reduces the ground plane spacing h_o to h_1 as shown in Fig. 2e such that the characteristic impedance of the air line is reduced to $Z_o/(\epsilon_{eff})^{1/2}$.

We can now state that:

(*i*) The microstrip with dielectric material of Fig. 2d and the microstrip without dielectric material of Fig. 2e have the same impedance.

(*ii*) If we assume that the current distribution is uniform for the air line over the width w_o on the ground plane and the adjacent bottom face of the strip we obtain the same unloaded Q for both lines of Fig. 2d and Fig. 2e. The attenuation of the air line is lower by a factor $(\epsilon_{eff})^{1/2}$ as given in Fig. 2e.

IV. COMPUTATION OF LINE PARAMETERS

4.1 *Exact Analytic Solution for Impedance by Conformal Mapping*

The charactreistic impedance of the microstrip of Fig. 2a with thickness $t = 0$ can be obtained by Schwarz-Christoffel integrals which transform the upper half of a complex z_1 plane into a rectangle in the complex z plane.[8–10] More specifically, one has to find an analytic function which maps the two strip boundaries in the z_1-plane on two opposite sides of the rectangle as shown in Fig. 3. The Schwarz-Christoffel integral for this specific case can be expressed in terms of the theta function ϑ_1 and ϑ_4. Theta functions are well behaved analytic functions of a complex variable, their properties are well known, and rapidly converging series have been published.[11,12] These functions and their logarithmic derivatives are essential mathematical tools for solving the following engineering problems:

(*i*) characteristic impedance of conductors with strip geometries,

(*ii*) junction capacitance in semiconductor diodes with strip junctions,

(*iii*) heat flow and thermal resistance from a line source into a solid, and

(*iv*) series resistance of bulk devices with stripe contacts.

The conformal transformation $z_1 = z_1(z)$ expressed in terms of the logarithmic derivative of the theta function ϑ_1 and its parameter $\kappa =$

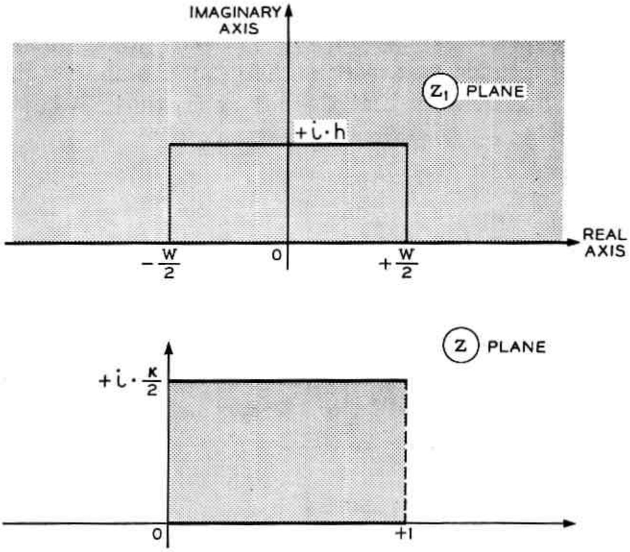


Fig. 3 — Conformal mapping of a microstrip by the logarithmic derivative of the theta function $\vartheta_1(z, \kappa)$.

$$z_1 = -\frac{2hK}{\pi}\frac{\partial \ln \vartheta_1(z,\kappa)}{\partial z}, \; Z_0 = \left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}} \cdot \frac{\kappa}{2}$$

K'/K is

$$z_1 = -\frac{2hK}{\pi}\frac{\partial}{\partial z}\ln\vartheta_1(z,\kappa) \tag{7}$$

where $K = K(m)$ and $K' = K'(m)$ are complete elliptic integrals of the first kind with modulus m.

The characteristic impedance Z_o of the microstrip with width w, height h, and thickness $t = 0$ is obtained by solving the following equations

$$\frac{w}{h} = \frac{2}{\pi}\frac{\partial}{\partial\zeta}\ln\vartheta_4(\zeta,\kappa) \tag{8}$$

$$\mathrm{dn}^2(2K\zeta) = \frac{E}{K} \tag{9}$$

$$Z_o = \frac{1}{2}\left(\frac{\mu_o}{\epsilon_o}\right)^{\frac{1}{2}}\frac{K'}{K}. \tag{10}$$

$E = E(m)$ is the complete elliptic integral of the second kind, dn the Jacobian elliptic function, μ_o and ϵ_0 the magnetic and dielectric permeabilities of free space. With $(\mu_0/\epsilon_0)^{1/2} = 120\pi$ ohm and $\kappa = K'/K$ one obtains

$$Z_o = 60\pi\kappa \text{ ohm}. \tag{11}$$

For a very narrow strip $w \ll h$ and a very wide strip $w \gg h$ one obtains the simple expressions

$$Z_o = 60 \ln\frac{8h}{w} \text{ ohm} \qquad w \ll h \tag{12}$$

$$Z_o = \frac{120\pi h}{w} \text{ ohm} \qquad w \gg h. \tag{13}$$

The exact computation for one important intermediate case by means of a series expansion for the logarithmic derivative of the theta function ϑ_4 is treated in the appendix.

4.2 *Impedance Design Formulas*

The rigorous solution for computing Z_o from equations (8), (9), and (10) is not recommended for most engineering applications. Useful expressions in terms of rational functions or series expansions can be obtained by generalization of equations (12) and (13) as follows

$$Z_o = 60 \ln \sum_{n=1}^{-\infty} a_n\left(\frac{h}{w}\right)^n \text{ ohm} \qquad w \leqq h \tag{14}$$

$$Z_o = \frac{120\pi}{\sum_{n=1}^{-\infty} b_n \left(\frac{w}{h}\right)^n} \text{ ohm} \qquad w \geqq h. \tag{15}$$

The number of terms after which the series is terminated determines the accuracy of the approximations. The following formulas obtained by rational function approximation give an accuracy of ±0.25 per cent for $0 \leqq w/h \leqq 10$ which is the range of importance for most engineering applications

$$Z_o = 60 \ln \left(\frac{8h}{w} + \frac{w}{4h}\right) \text{ ohm} \qquad \frac{w}{h} \leqq 1 \tag{16}$$

$$Z_o = \frac{120\pi \text{ ohm}}{\frac{w}{h} + 2.42 - 0.44 \frac{h}{w} + \left(1 - \frac{h}{w}\right)^6} \qquad \frac{w}{h} \geqq 1. \tag{17}$$

The accuracy obtained for strips with $w/h > 10$ from equation (17) is ±1 per cent.

Table I compares the impedance obtained with theta functions, the impedance calculated from the rational function approximations, and the measured value, with a time domain reflectometer for $w/h = 1$. The physical dimensions of the line used for the time domain reflectometer measurement are listed in Table II.

The estimated maximum error for Z_o is ±0.7 percent. Measurements for different ratios w/h by the same procedure have also given excellent agreement with data obtained by means of the logarithmic derivative of the theta function $\vartheta_4\ (\zeta, \kappa)$.

Figure 4 is a plot of Z_o as a function of w/h. The impedance for the important case of the standard microstrip of Fig. 1a is also plotted for two materials which look attractive for hybrid integrated circuits in the microwave and the millimeter wave range. These materials are

Table I — Characteristic Impedance for $w/h = 1$

Method	Z_o Ohm
Rigorous solution with theta functions eqs. (8), (9), (10)	126.553
Measured impedance with time domain reflectometer (Table II)	126.60
Approximation with narrow strip rational function equation (16)	126.613
Approximation with wide strip rational function equation (17)	126.507

TABLE II — IMPEDANCE MEASUREMENT DATA WITH TIME DOMAIN REFLECTOMETER

Impedance standard, General Radio coaxial precision air line	GR 900-L 50 Ω
Time domain reflectometer, Hewlett Packard	hp 1415A
Microstrip ground plane spacing h,	0.750 inch
width w,	0.750 inch
thickness t	0.001 inch
Dielectric constant of polyfoam support and polyfoam cover	$\epsilon_r = 1.032$
Measured impedance for thickness $t = 0.001$ inches, dielectric constant $\epsilon_r = 1.032$	124.42 Ω
Extrapolated impedance Z for thickness $t = 0$ from measurements for $t = 0.001$ inch, 0.0115 inch, 0.0265 inch, 0.0525 inch and 0.0625 inch	124.62 Ω
Microstrip air line impedance $Z_0 = (\epsilon_r)^{\frac{1}{2}}Z$	126.60 Ω

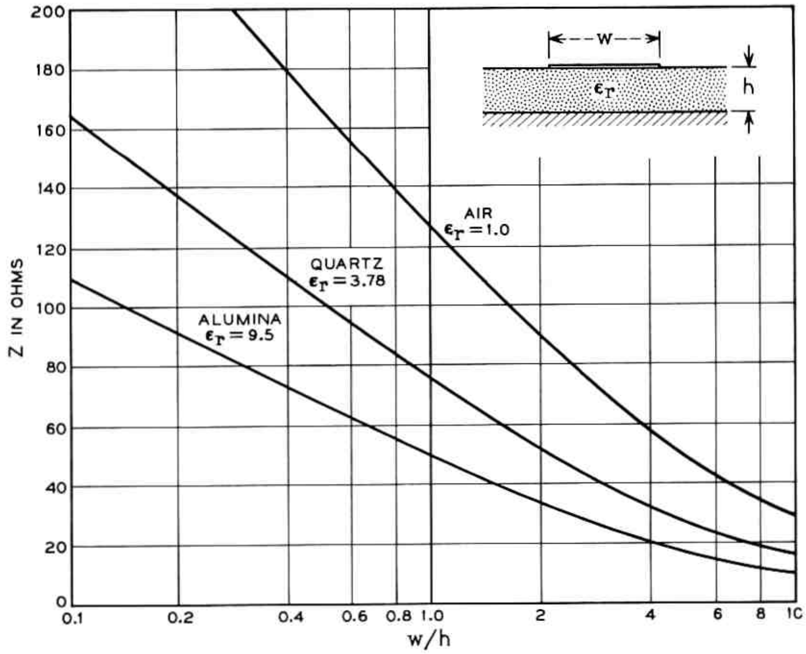


Fig. 4 — Characteristic impedance of the standard microstrip for $\epsilon_r = 1$ and impedance of the standard microstrip for $\epsilon_r = 3.78$ (quartz) and $\epsilon_r = 9.5$ (alumina) as a function of w/h.

fused quartz (SiO_2) with $\epsilon_r = 3.78$ and 99.5 percent alumina (Al_2O_3) with $\epsilon_r = 9.5$ whose impedance curves are based on computed effective dielectric constants treated in Section 4.3.

4.3 *Computation and Measurement of Effective Dielectric Constant*

The electrical parameters of any microstrip can be computed if the characteristic impedance Z_o of the corresponding air line and the dielectric constant $(\epsilon_{eff})^{1/2}$ are known. The basic equations required for this computation are listed in Fig. 2.

The effective dielectric constant ϵ_{eff} is a function of the ratio w/h, the relative dielectric constant ϵ_r, and the geometrical shape of the boundary between air and the dielectric support material. The effective dielectric constant can be obtained by starting from the transformation given by equation (7), by mapping the boundaries between air and dielectric into the rectangle in the z-plane of Fig. 3, and by treating the new geometrical configuration obtained inside the rectangle of the z-plane as a parallel plate capacitor which is partially filled with dielectric.

Notice that the fringe field problem is eliminated in the z-plane because the complete upper half of the plane is transformed into one rectangle. The procedure is rigorous since conformal mapping preserves the angle of refraction of electric field lines at the boundary between dielectric and air. If the capacitance of the parallel plate configuration in the z-plane of Fig. 3 is C_o without dielectric and C with partial dielectric filling one obtains

$$\epsilon_{eff} = \frac{C}{C_o}. \tag{18}$$

The method which is outlined above has been used by Wheeler for the standard microstrip of Fig. 1a by starting from an approximate conformal mapping transformation and by using an approximation for the transformed parallel plate capacitance.[13] The square root of the effective dielectric constant $(\epsilon_{eff})^{1/2}$ obtained by this method is shown in Fig. 5 as a function of w/h and ϵ_r.

In order to find a function which approximates the set of curves of Fig. 5 over the total range $0 \leqq w/h < \infty$ and $1 \leqq \epsilon_r < \infty$ we define a function $F(\epsilon_r, w/h)$ by

$$\epsilon_{eff} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2} F\left(\epsilon_r, \frac{w}{h}\right). \tag{19}$$

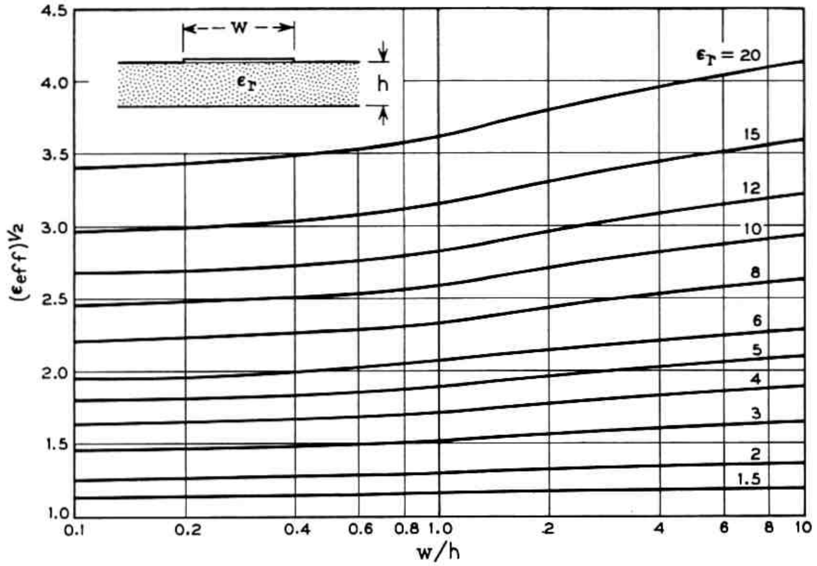


Fig. 5 — Square root of the effective dielectric constant for the standard microstrip. $(\epsilon_{eff})^{1/2}$ plotted as a function of w/h with ϵ_r as parameter.

From equation (5) we find for the standard microstrip of Fig. 1(a)

$$0 \leqq F\left(\epsilon_r, \frac{w}{h}\right) \leqq 1. \tag{20}$$

One class of functions which fulfills this requirement is the class of irrational functions

$$F\left(\epsilon_r, \frac{w}{h}\right) = \left[1 + \sum_{n=1}^{N} c_n \left(\frac{h}{w}\right)^n\right]^m \tag{21}$$

with c_n being functions of ϵ_r and $m \leqq 0$. The set of curves of Fig. 5 can be approximated with $m = -0.5$ and one single term of the series by

$$F\left(\epsilon_r, \frac{w}{h}\right) = \left(1 + \frac{10h}{w}\right)^{-\frac{1}{2}}. \tag{22}$$

The final result with an accuracy of ±2 per cent for ϵ_{eff} and an accuracy of ±1 per cent for $(\epsilon_{eff})^{\frac{1}{2}}$ is

$$\epsilon_{eff} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2}\left(1 + \frac{10h}{w}\right)^{-\frac{1}{2}}. \tag{23}$$

Effective dielectric constants can also be obtained by static capacitance measurements or time domain reflectometer measurements. If the static capacitance per unit length is C with partial dielectric filling and C_o with the dielectric removed, one obtains $\epsilon_{eff} = C/C_o$ and from $Z = (L/C)^{1/2}$ with $L = Z_o/v_o$

$$Z = \frac{Z_o}{(\epsilon_{eff})^{\frac{1}{2}}} = \frac{1}{v_o(CC_o)^{\frac{1}{2}}} \tag{24}$$

where v_o is the velocity of light in vacuum, $v_o = 3.10^{10}$ cm per second.

Accurate measurements of ϵ_{eff} with a time domain reflectometer require a precision coaxial connector standard and a good transition from coaxial transmission line into the microstrip. Baseband transitions up to a few GHz can be made by building an oversize model of the partially filled microstrip as shown in Fig. 6. The inverted microstrip of Fig. 1(e) used for this measurement is supported by clear fused and polished quartz plates with a dielectric constant $\epsilon_r = 3.78$. The effective dielectric constant ϵ_{eff} plotted as a function of w/h is much lower than ϵ_r because only a small fraction of field lines passes through the quartz. Similar results are obtained if the line shown in Fig. 6 is completely shielded provided that the major part of the radio frequency energy remains concentrated in the

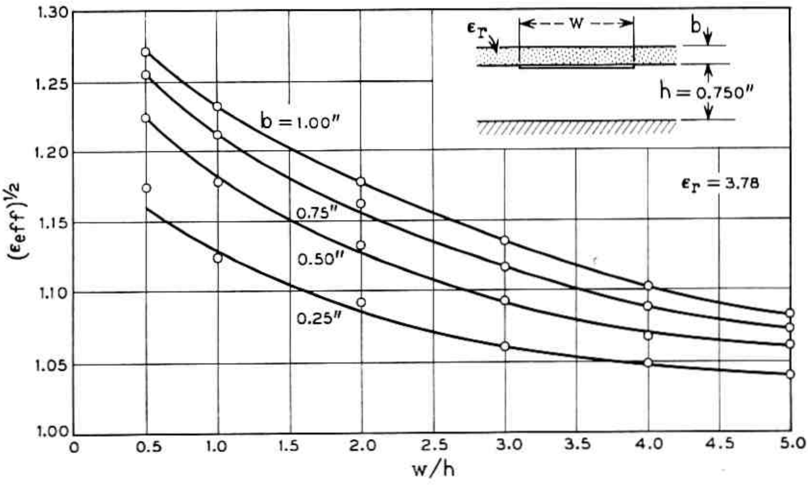


Fig. 6 — Square root of the effective dielectric constant for an inverted microstrip with quartz substrate. Oversize measurement with strip conductor thickness $t = 0.010$ inch.

air gap between the ground plane and the strip conductor. One concludes from this measurement that all of the electrical parameters of inverted microstrips are close to the electrical parameters of the air line. One also concludes that dielectric losses are substantially reduced because all the dielectric support material is removed into the low field region of the microstrip.

4.4 *Computation of Conductor Attenuation*

The attenuation of any lumped or distributed circuit element is known if its inductance as a function of the geometrical conductor parameters can be calculated. Inductance and conductor attenuation are related because the inductance is the normalized magnetic field energy of the circuit element and attenuation is proportional to the magnetic field energy stored in the metal conductor.[14] In order to calculate the attenuation one has to recede the metal surface by one skin depth or more generally by a small length δn normal to the conductor surface. If the corresponding increase in inductance is δL and if the skin resistance of the metal is R_s, then the radio frequency resistance R of the circuit or line element is

$$R = \frac{R_s}{\mu_o} \frac{\delta L}{\delta n} \tag{25}$$

with the skin resistance R_s given by

$$R_s = (\pi \mu_o f \rho)^{\frac{1}{2}} \text{ ohm} \tag{26}$$

where f is the frequency in Hz, ρ the conductor resistivity in ohm·cm, and $\mu_o = 4\pi \cdot 10^{-9}$ henry per cm. The skin resistance in ohms as a function of frequency is plotted in Fig. 7 with ρ in ohm·cm as a parameter.

The inductance L and the attenuation α_o in neper per unit length of a microstrip which supports a transverse electromagnetic mode are given by

$$L = (\epsilon_o \mu_o)^{\frac{1}{2}} Z_o \tag{27}$$

$$\alpha_o = \frac{R}{2Z_o}. \tag{28}$$

From equations (25), (27), and (28) one obtains

$$\alpha_o = \left(\frac{\epsilon_o}{\mu_o}\right)^{\frac{1}{2}} \frac{R_s}{2Z_o} \frac{\delta Z_o}{\delta n}. \tag{29}$$

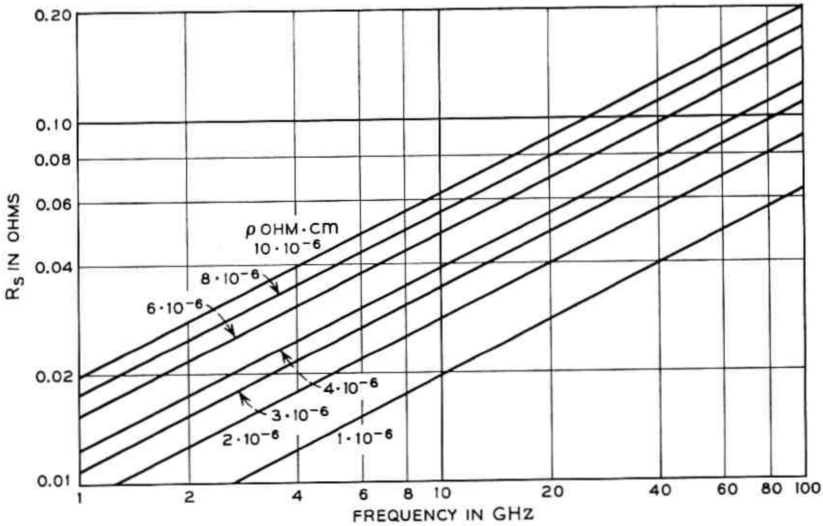


Fig. 7 — Skin resistance R_s of metals as a function of frequency. Bulk resistivity at dc and 20°C for suitable conductors is 1.7 μohm·cm for copper, 1.6 μohm·cm for silver, 2.3 μohm·cm for gold, and 2.8 μohm·cm for aluminum.

The geometrical conductor parameters of the microstrip are width w, height h and thickness t. Let us assume first that the skin resistance of the ground plane is different from the skin resistance of the strip, for example, the two conductor materials are not the same. The attenuation α_1 owing to the ground plane with a skin resistance R_{s1} is obtained by receding the metal surface by $\delta n = \delta h$

$$\alpha_1 = \left(\frac{\epsilon_o}{\mu_o}\right)^{\frac{1}{2}} \frac{R_{s1}}{2Z_o} \frac{\partial Z_o}{\partial h}. \tag{30}$$

The strip attenuation α_2 with a skin resistance R_{s2} is obtained by reducing the strip dimensions by $2\delta w$ and $2\delta t$ as well as increasing the ground plane spacing by δh

$$\alpha_2 = \left(\frac{\epsilon_o}{\mu_o}\right)^{\frac{1}{2}} \frac{R_{s2}}{2Z_o} \left[\frac{\partial Z_o}{\partial h} - 2\frac{\partial Z_o}{\partial w} - 2\frac{\partial Z_o}{\partial t}\right]. \tag{31}$$

The total attenuation α_o is

$$\alpha_o = \alpha_1 + \alpha_2 . \tag{32}$$

If the conductor materials for the ground plane and the strip are the same we obtain with $R_{s1} = R_{s2} = R_s$

$$\alpha_o = \left(\frac{\epsilon_o}{\mu_o}\right)^{\frac{1}{2}} \frac{R_s}{Z_o} \left(\frac{\partial Z_o}{\partial h} - \frac{\partial Z_o}{\partial w} - \frac{\partial Z_o}{\partial t}\right). \tag{33}$$

It is useful to write the partial derivatives:

$$\frac{\partial Z_o}{\partial w} = +\frac{1}{h} \frac{\partial Z_o}{\partial\left(\frac{w}{h}\right)} \tag{34}$$

$$\frac{\partial Z_o}{\partial h} = -\frac{w}{h^2} \frac{\partial Z_o}{\partial\left(\frac{w}{h}\right)} \tag{35}$$

$$\frac{\partial Z_o}{\partial t} = +\frac{\partial Z_o}{\partial w} \frac{\partial w}{\partial t} \tag{36}$$

with $\partial w/\partial t$ being the derivative of w with respect to t for constant Z_o. The attenuation α_o in dB per unit length is finally

$$\alpha_o = -\frac{R_s}{6\pi \ln 10} \frac{\partial Z_o}{\partial\left(\frac{w}{h}\right)} \frac{1 + \frac{w}{h} + \frac{\partial w}{\partial t}}{hZ_o}. \tag{37}$$

The partial derivative $\partial w/\partial t$ can be derived from approximate expressions published by Wheeler, Caulton, Hughes, and Sobol.[13,15] They define an effective width $w_{eff} = w + \Delta w$ by considering two different microstrips with the same characteristic impedance Z_o and different dimensions given by w, h, $t \neq 0$ and w_{eff}, h, $t = 0$. The approximations are

$$\Delta w = w_{eff} - w = \frac{t}{\pi}\left(1 + \ln \frac{4\pi w}{t}\right) \qquad \frac{w}{h} \leqq \frac{1}{2\pi} \tag{38}$$

$$\Delta w = w_{eff} - w = \frac{t}{\pi}\left(1 + \ln \frac{2h}{t}\right) \qquad \frac{w}{h} \geqq \frac{1}{2\pi}. \tag{39}$$

Additional restrictions for applying equations (38) and (39) are $t \ll h$, $t < w/2$, and $t/\Delta w < 0.75$. Notice also that the ratio $\Delta w/t$ obtained from equations (38) or (39) is divergent for $t \rightarrow 0$. This does not present a problem since equations (29) to (37) are only applicable if the conductor thickness exceeds several skin depths.

Being aware of these limitations, we obtain the partial derivatives $\partial w/\partial t$ by computing $\partial w_{eff}/\partial t$ from equations (38) and (39)

$$\frac{\partial w}{\partial t} = \frac{1}{\pi} \ln \frac{4\pi w}{t} \qquad \frac{w}{h} \leqq \frac{1}{2\pi} \tag{40}$$

$$\frac{\partial w}{\partial t} = \frac{1}{\pi} \ln \frac{2h}{t} \qquad \frac{w}{h} \geqq \frac{1}{2\pi}. \tag{41}$$

It is convenient for design purposes to define the normalized attenuation A in dB per ohm as follows

$$A = \frac{h\alpha_o}{R_s} = -\frac{1}{6\pi \ln 10} \frac{\partial Z_o}{\partial\left(\frac{w}{h}\right)} \frac{1 + \frac{w}{h} + \frac{\partial w}{\partial t}}{Z_o}. \tag{42}$$

A is plotted in Fig. 8 as a function of w/h with $\partial w/\partial t$ as a parameter. The normalized attenuation A based on the assumption of uniform current distribution over the width w of the bottom conductor and the adjacent bottom side of the strip conductor is also shown in Fig. 8 for comparison. The formula valid for uniform current distribution is

$$A = \frac{20}{\ln 10} \frac{h}{wZ_o} \frac{\text{dB}}{\text{ohm}}. \tag{43}$$

One can show that equations (42) and (43) give the same result for $w/h \gg 1$ since $Z_o = 120\pi h/w$ and $\partial Z_o/\partial(w/h) = -120\pi h^2/w^2$. This is expected because fringe fields can be neglected for wide strips. One obtains a lower attenuation from equation (42) for narrow strips because currents are flowing on the top and bottom side of the strip and also because of the beneficial effect of wider current distribution in the ground plane because of fringe fields. For narrow strips the result is with $Z_o = 60 \ln (8h/w + w/4h)$ ohm

$$A = \frac{10}{\pi \ln 10} \frac{\left(\frac{8h}{w} - \frac{w}{4h}\right)\left(1 + \frac{h}{w} + \frac{h}{w}\frac{\partial w}{\partial t}\right)}{Z_o \exp\left(\frac{Z_o}{60}\right)} \qquad \frac{w}{h} \leqq 1. \tag{44}$$

For wide strips one obtains from equations (17) and (42)

$$A = \frac{Z_o}{720\pi^2 \ln 10}\left[1 + \frac{0.44h^2}{w^2} + \frac{6h^2}{w^2}\left(1 - \frac{h}{w}\right)^5\right]\left(1 + \frac{w}{h} + \frac{\partial w}{\partial t}\right)$$
$$\frac{w}{h} \geqq 1. \tag{45}$$

For design purposes it is recommended to read R_s and A from Figs. 7 and 8 and to obtain α_o in dB per unit length from

$$\alpha_o = \frac{R_s A}{h}. \tag{46}$$

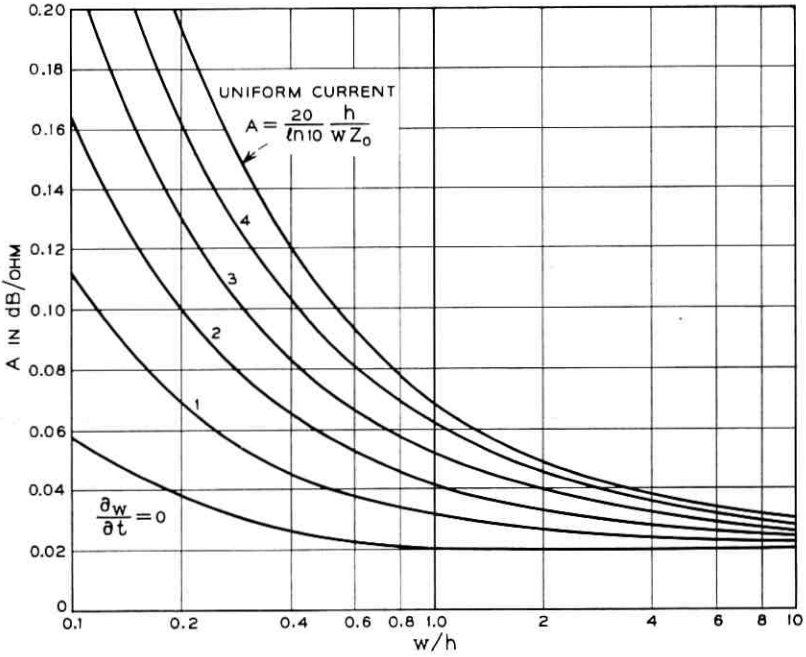


Fig. 8 — Normalized conductor attenuation $A = \alpha_0 h/R_s$ in dB per ohm for a standard microstrip with $\epsilon_r = 1$. The partial derivative $\partial w/\partial t$ is a function of the conductor thickness t and given by equations (40) and (41). The conductor attenuation for partial dielectric filling is $\alpha = (\epsilon_{eff})^{1/2}\ \alpha_o$ as given by equation (3).

The conductor attenuation for partial dielectric filling is obtained from equation (3).

4.5 *Measurement of Microstrip Attenuation*

Measurements of the microstrip attenuation in the 1 to 6 GHz frequency range have been performed by Caulton, Hughes, Sobol, Pucel, Massé, and Hartwig.[15,16] Good agreement between theory and experiment has been obtained in Ref. 15 based on the assumption of uniform current distribution. Good agreement is also obtained in Ref. 16 based on the assumption of the correct nonuniform current distribution. This can be explained in part because the skin resistance R_s used for the calculations in Ref. 16 is based on the dc resistivity of the copper conductor plus a sizable correction in order to account for surface roughness. This correction increases R_s by 13 percent at 1 GHz and 33 percent at 6 GHz. From recent work by L. U. Kibler one concludes that this correction may be too large even if one takes into account the fact that the data obtained by Kibler in Ref. 17 are based on electroformed oxygen free copper without any additional treatment for improving the surface finish.

The measurement of attenuation at 30 GHz requires a low loss transition from waveguide into microstrip. Such a transition has been developed by W. F. Bodtmann.[18] Clear fused and polished quartz substrates are used for the substrate material in order to obtain a low effective dielectric constant. Evaporated and photoetched nichrome-gold layers with a thickness of 2 μm are used for the conductor materials on both substrate surfaces. Table III summarizes the properties of the microstrip.

Table IV gives the attenuation measured for a 3-inch long microstrip line by means of a transmission measurement with two-waveguide to microstrip transition at both ends of the microstrip. The theoretical loss based on the assumption of uniform current distribution and the theoretically computed current distribution is given in Table V.

The agreement between measured and calculated data does not necessarily support the uniform current theory. It indicates as expected that the radio frequency film resistivity ρ at 30 GHz is higher than the dc resistance of Table V. The dc resistivity is calculated from a measurement of the composite nichrome-gold resistance and a thickness measurement with a Tolansky interferometer.

The attenuation α' per guide wavelength is 0.0609 dB. A value from 0.060 to 0.068 dB has been measured in the 26.5 to 30.5 GHz frequency range. The unloaded Q is given by

$$Q = \frac{20\,\pi}{\ln 10}\frac{1}{\alpha_o\lambda_o} = \frac{27.3}{\alpha'} = 450. \tag{47}$$

TABLE III — MICROSTRIP DATA

Type of microstrip	standard of Fig. 1a
Substrate material*	clear fused quartz
Substrate thickness h	0.030 inch
Conductor width w	0.030 inch
Conductor thickness t	2 μm Nichrone-gold
Metal deposition	evaporated
Thickness of Nichrome base layer	100 to 150 A
Line fabrication	photoetching
Conductor resistivity	$\rho = 3.0 \cdot 10^{-6}$ Ohm·cm

* 99.8 percent SiO_2, Amersil Inc., Hillside, New Jersey.

TABLE IV—MEASURED MICROSTRIP LOSS AT 30 GHz

Measured total loss of waveguide to microstrip transitions and 3-inch long microstrip at 30 GHz	0.88 dB
Measured insertion loss for both transitions (two transitions back to back)	0.10 dB
Attenuation for line length l = 3 inches	0.78 dB

This is believed to be the highest Q obtained for a microstrip in this frequency range.

V. MODE PROPAGATION IN MICROSTRIPS

Microstrip transmission lines which are fully shielded and completely filled with dielectric material can propagate transverse electromagnetic, transverse electric, and transverse magnetic modes. Partially filled and fully shielded lines cannot support these modes because the boundary conditions at the interface between air and dielectric cannot be rigorously fulfilled. Zysman and Varon have shown that a hybrid mode can be found which satisfies all boundary conditions and which can be decomposed into sums of transverse electric and transverse magnetic space harmonics.[19] From their results one concludes that the hybrid mode propagates at all frequencies and that it approaches the transverse electromagnetic mode at low frequencies or for sufficiently small line dimensions.

The problem of hybrid mode propagation has also been treated by Pregla, Schlosser, Hartwig, Massé, and Pucel.[20,21] One concludes from the results that the frequency dependent behavior or the dispersion of the propagation constant and the effective dielectric constant is

TABLE V—THEORETICAL MICROSTRIP LOSS AT 30 GHz

Square root of effective dielectric constant for $\epsilon_r = 3.78$ and $w/h = 1$, equation (23)	$(\epsilon_{eff})^{\frac{1}{2}} = 1.68$
Conductor skin resistance for $f = 30$ GHz and $\rho = 3.0 \times 10^{-6}$ ohm·cm, Fig. 7	$R_s = 0.060$ ohm
Normalized attenuation, uniform current distribution, $w/h = 1$, Fig. 8	$A = 0.0685$ dB per ohm
Normalized attenuation, nonuniform current distribution, $\partial w/\partial t = 2.1$, Fig. 8	$A = 0.0420$ dB per ohm
Attenuation, line length $l = 3''$, $(\epsilon_{eff})^{\frac{1}{2}} R_s A l/h$, uniform current	0.690 dB
Attenuation, line length $l = 3''$ Nonuniform current distribution	0.423 dB

particularly pronounced for lines with substrates which have a high dielectric constant, such as alumina with $\epsilon_r = 9.6$ and rutile with $\epsilon_r = 104$. It is also shown that the frequency of operation has to be lower than the cutoff frequency f_c of the lowest order transverse electric surface wave which is given by

$$f_c = \frac{75}{h(\epsilon_r - 1)^{\frac{1}{2}}} \text{ GHz} \tag{48}$$

where h is the substrate thickness in millimeter.[21] The cutoff frequency obtained for the line of Table III with $h = 0.75$ mm and $\epsilon_r = 3.78$ is $f_c = 60$ GHz. For high density alumina with $\epsilon_r = 9.6$ the cutoff is considerably lower with $f_c = 34$ GHz.

VI. CONCLUSIONS

The electrical properties of microstrips can be derived from the characteristic impedance of the air line and the effective dielectric constant if the propagating mode can be approximated by a transverse electromagnetic mode. Substrates with a low dielectric constant are useful for circuit applications because dispersion of the line parameters is less pronounced. Structures with an air gap are recommended if circuit losses have to be minimized. Complete shielding is essential for most applications in order to reduce radiation loss and to reduce the coupling between different circuits.

VII. ACKNOWLEDGMENT

The author expresses his thanks to S. Michael and W. W. Snell for performing exact measurements and to S. Shah for providing high quality films.

APPENDIX

Computing Microstrip Impedance with Theta Functions

The following example gives the numerical procedure for computing the characteristic impedance of a microstrip. It is convenient to calculate w/h and Z_o as a function of the modulus m of the complete elliptic integrals K, K', and E. We assume $m = 0.86$ for this example and use equations (8), (9), and (10)

(*i*) $m = 0.86$ modulus of complete elliptic integrals

$$(ii)\quad \left.\begin{aligned} K &= 2.42093\\ K' &= 1.63058\\ E &= 1.13600\\ \kappa &= K'/K = 0.673532.\end{aligned}\right\}\ \text{From tables, Ref. 22.}$$

(*iii*) Characteristic impedance from equation (10) with $(\mu_o/\epsilon_o)^{\frac{1}{2}} = 120\pi$ ohm

$$Z_o = 60\pi \frac{K'}{K} = 126.958 \text{ ohm}. \tag{49}$$

(*iv*) From equation (9) and tables of the Jacobian elliptic functions we obtain[23]

$$\mathrm{dn}^2(2K\zeta) = \frac{E}{K} = 0.469240 \tag{50}$$

$$2K\zeta = \mathrm{arc\ dn}\left(\frac{E}{K}\right)^{\frac{1}{2}} = 1.02806 \tag{51}$$

$$\zeta = 0.212328. \tag{52}$$

(*v*) We use the rapidly converging series expansion[12]

$$\frac{\partial}{\partial\zeta} \ln \vartheta_4(\zeta, \kappa) = 4\pi \sum_{n=1}^{\infty} \frac{\exp(-n\pi\kappa)}{1 - \exp(-2n\pi\kappa)} \sin(2n\pi\zeta) \tag{53}$$

and obtain for the sum S of the first 10 terms $S = 0.124095$.

(*vi*) From equation (10) we obtain

$$\frac{w}{h} = \frac{2}{\pi}\frac{\partial}{\partial\zeta} \ln \vartheta_4(\zeta, \kappa) = 0.992762. \tag{54}$$

The result listed in Table I is based on quadratic interpolation from a table made with closely spaced moduli m.

REFERENCES

1. Ruthroff, C. L., Osborne, T. L., and Bodtmann, W. F., "Short Hop Radio System Experiment," scheduled for B.S.T.J., *48*, No. 6 (July–August 1969).
2. Tillotson, L. C., "A Model of a Domestic Satellite Communication System," B.S.T.J., *47*, No. 10 (December 1968), pp. 2111–2137.
3. Brenner, H. E., "Use a Computer to Design Suspended-Substrate Integrated Circuits," Microwaves, *7*, No. 9 (September 1968), pp. 38–45.
4. Engelbrecht, R. S., and Kurokawa, K., "A Wideband Low Noise L-Band Balanced Transistor Amplifier," Proc. IEEE, *53*, No. 3 (March 1965), pp. 328–333.
5. Saunders, T. E., and Stark, P. D., "An Integrated 4-GHz Balanced Transistor Amplifier," IEEE J. Solid State Elec., *SC-2*, No. 1 (March 1967), pp. 4–10.
6. Bonfeld, M. D., Bonomi, M. J., and Jaasma, E. G., "An Integrated Micro-

wave FM Discriminator," 1968 G-MTT Int. Microwave Symp. Digest, Detroit, Michigan, May 20–22, 1968, pp. 139–146.

7. Cohn, S. B., "Slot Line—An Alternative Transmission Medium for Integrated Circuits," 1968 G-MTT Int. Microwave Symp. Digest, Detroit, Michigan, May 20–22, 1968, pp. 104–109.
8. Binns, K. J., and Lawrenson, P. J., *Electric and Magnetic Field Problems,* New York: MacMillan, 1963, pp. 157–224.
9. Moon, P., and Spencer, D. E., *Field Theory for Engineers,* New York: D. Van Nostrand, 1961, pp. 339–357.
10. Smythe, W. R., *Static and Dynamic Electricity,* New York: McGraw Hill, 1950, pp. 82–101.
11. Bellman, R., *A Brief Introduction to Theta Functions,* New York: Holt Rinehart and Winston, 1961, pp. 1-72.
12. Tölke, F., *Praktische Funktionenlehre, Zweiter Band, Theta-Funktionen und Spezielle Weierstrass' sche Funktionen,* Berlin: Springer Verlag, 1966, pp. 1–83.
13. Wheeler, H. A., "Transmission-Line Properties of Parallel Strips Separated by a Dielectric Sheet," IEEE Trans. Microwave Theory and Techniques, *MTT-13,* No. 2 (March 1965), pp. 172–185.
14. Wheeler, H. A., "Formulas for the Skin Effect," Proc. IRE, *30,* No. 9 (September 1942), pp. 412–424.
15. Caulton, M., Hughes, J. J., and Sobol, H., "Measurement of the Properties of Microstrip Transmission Lines for Microwave Integrated Circuits," RCA Review, *27,* No. 3 (September 1966), pp. 377–391.
16. Pucel, R. A., Massé, D. J., and Hartwig, C. P., "Losses in Microstrip," IEEE Trans. Microwave Theory and Techniques, *MTT-16,* No. 6 (June 1968), pp. 342–350.
17. Kibler, L. U., "The Properties and Uses of the Cutoff Frequency Region of a Lossy Rectangular Waveguide," Ph.D. Thesis, Polytechnic Institute of Brooklyn, June 1968.
18. Schneider, M. V., Glance, B., and Bodtmann, W. F., "Microwave and Millimeter Wave Hybrid Integrated Circuits for Radio Systems," scheduled for B.S.T.J., *48,* No. 6 (July–August 1969).
19. Zysman, G. I., and Varon, D., "Wave Propagation in Microstrip Transmission Lines," Int. Microwave Symp., IEEE Group on Microwave Theory and Techniques, Dallas, Texas, May 5–7, 1969.
20. Pregla, R., and Schlosser, W., "Waveguide Modes in Dielectrically Supported Strip Lines," Archiv der Elektrischen Uebertragung, *22,* No. 8 (August 1968), pp. 379–386.
21. Hartwig, C. P., Massé, D., and Pucel, A. P., "Frequency Dependent Behavior of Microstrip," 1968 G-MTT Int. Microwave Symp. Digest, Detroit, Michigan, May 20–22, 1968, pp. 110–116.
22. Abramowitz, M., and Stegun, I. A., *Handbook of Mathematical Functions,* National Bureau of Standards, Applied Mathematics Series 55, 2nd printing, November 1964, pp. 608–609.
23. Fettis, H. E., and Caslin, J. C., *Ten Place Tables of the Jacobian Elliptic Functions,* Aerospace Research Laboratories, Wright Paterson Air Force Base, Ohio, AD 631–869, September 1965.

Experimental Simulation of a Multiple Beam Optical Waveguide

By D. GLOGE and W. H. STEIER*

(Manuscript received November 7, 1968)

Two mirrors, 15 centimeters in diameter and 25 meters apart, form an optical delay line which can store two gaussian beams for 342 round trips or 60 microseconds. This paper reports experiments which studied the intensity profiles, the phase fronts, and the cross scattering between these beams after their retrieval from the delay line. In certain respects, the delay line simulates a multiple beam guide made of 684 mirror periscopes. The experimental results permit an estimate of the beam capacity, the crosstalk, and the transmission length of such a guide.

I. INTRODUCTION

The possibility of sending a multitude of gaussian light beams down a single lens waveguide has recently been suggested as an inexpensive means of multiplying the capacity of the waveguide.[1,2] Though the beams would overlap along the guide, appropriate optics could separate them in the receiver.

The density of resolvable beams in the system is determined by beam distortion and scattering rather than the spread of the ideal beams. Smooth imperfections of the optical surfaces cause the beam to deviate from the exact position and distort its profile and cross section.[3] This limits the density of the beams and determines the receiver size required to secure reception. Surface irregularities that are small compared with the beam size result in scattering that is collected by receivers of adjacent channels.[4] This crosstalk increases with the receiver size, the density of beams, and the number of scattering elements. The purpose of this experiment was to check the amount of distortion, to determine the receiver size required, and then to measure the scattering and find out what beam density and transmission dis-

* Formerly with Bell Telephone Laboratories at the Crawford Hill Laboratory. Now with the University of Southern California.

tance could be achieved with tolerable crosstalk. In a multiple beam guide, front surface mirrors probably will be preferred to lenses because, for the large apertures needed, lenses are apt to have imperfections within the material. A first simulation of such a mirror guide was tried here by folding two beams into a two-mirror cavity with a size comparable to one guide section. The setup was similar to optical delay lines built previously,[5] except that this line was optimized to exploit its full capacity.[1]

In a delay line, the folded beam wanders about the mirror surfaces, being submitted to always new and statistically independent mirror imperfections, similar to the waveguide situation. The distortion is therefore equivalent to the distortion in a guide. Two beams launched simultaneously follow adjacent paths comparable to two adjacent beams in a multiple beam waveguide. Their cross-scattering is equivalent to the cross-scattering of two neighboring beams in a waveguide.

II. THE FOLDED-BEAM CAVITY

Figure 1 shows the experimental setup with the two cavity mirrors in the background. Disregard the beam splitter for the moment and assume that only one gaussian beam, beam 1, is injected at an angle through the center hole in the front mirror. By introducing astigmatism to this mirror, as indicated by the arrows, the beam can be kept in the cavity for many round trips, writing a Lissajous pattern on each mirror.[5] Careful adjustment of this pattern permits recovery of the beam through the same hole at a slightly different angle. Figure 1 shows the two-lens telescope used to inject the laser beam and a little mirror at the focus of the telescope which deflects the output beam, beam 4, into a photomultiplier.

The delay line was designed so that a maximum number of round trips could be accommodated in an available 6-inch conduit, 25 meters long, with the beam axis never approaching the wall and the center hole closer than 2.5 beam radii. This clearance ratio is identical to the density factor k defined in Ref. 1.

Also from Ref. 1 one obtains the possible number of round trips

$$N_{\text{cavity}} = \frac{A^4}{4d^2\lambda^2k^4} \tag{1}$$

in a delay line of radius A and length d, using an optical wavelength λ. To allow for a slight misalignment of the conduit sections, we assumed an unobstructed cross section 12 cm in diameter. For $A = 6$ cm, $d =$

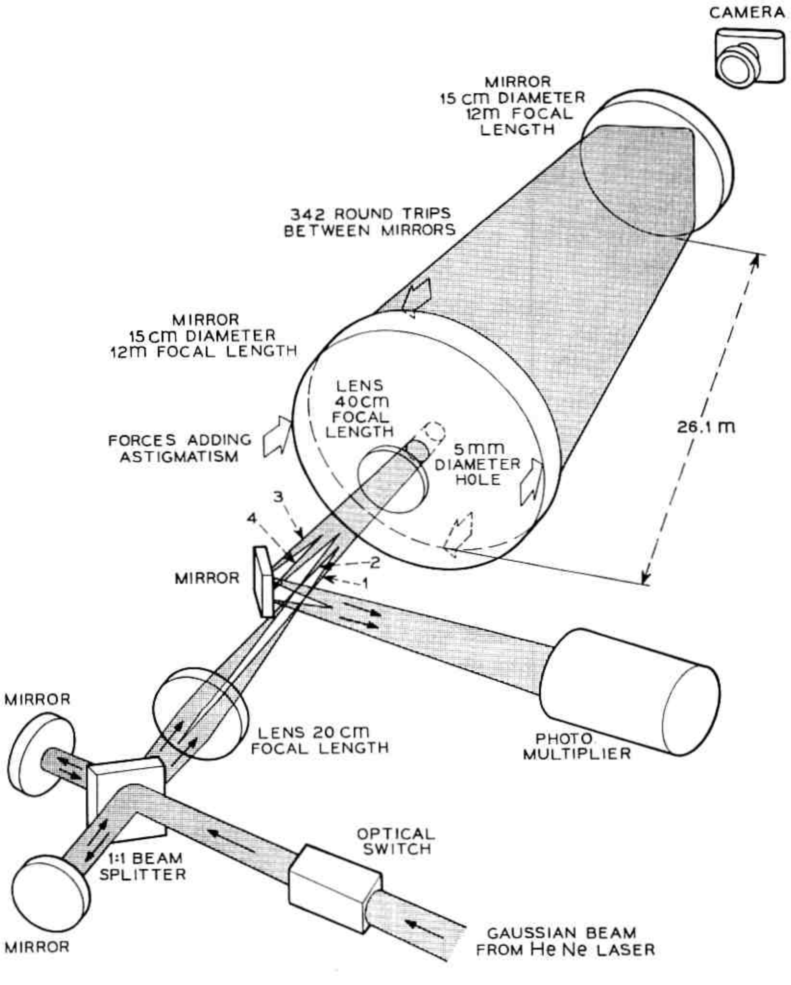


Fig. 1 — Injection and recovery of the two beams after 342 round trips in the delay line.

25 m, and $k = 2.5$, one obtains $N = 335$. We chose $342 = 18 \times 19$ round trips because, for optimum conditions, N must be a multiple of two consecutive integers.[1,5]

For this optimum design, Ref. 1 demands a focal length

$$f = \frac{d}{2 + \pi/N^{\frac{1}{2}}} \tag{2}$$

for the undistorted mirror and focal lengths

$$f_{h,v} = \frac{d}{2 + \frac{\pi}{N^{\frac{1}{2}}} \pm \frac{\pi}{N}} \tag{3}$$

for the astigmatic mirror in the horizontal and vertical planes, respectively. We chose $f = 12$ m and adjusted the mirror spacing to 26.1 m. This spacing was critical to within 1 mm. The astigmatic mirror had focal lengths $f_{h,v} = 12$ m ± 5 cm corresponding to a surface deflection of ±1 micron at the mirror edge when forces were applied as shown in Fig. 1. Both mirrors were 2.5 cm thick, 15 cm in diameter, polished spherical within $\lambda/10$, and coated for high reflectivity at 6328 Å, the wavelength of the He-Ne laser used.

The optimum design requires a beam radius

$$v = \frac{(d\lambda)^{\frac{1}{2}}}{N^{\frac{1}{4}}} \tag{4}$$

at the input.[1] For the chosen parameters $v = 1$ mm. We provided a center hole with a radius $kv = 2.5$ mm in the front mirror. The radius v is also the minimum radius the beam ever has in the cavity. Figure 2 is a photograph taken at the back of the rear mirror. It shows that

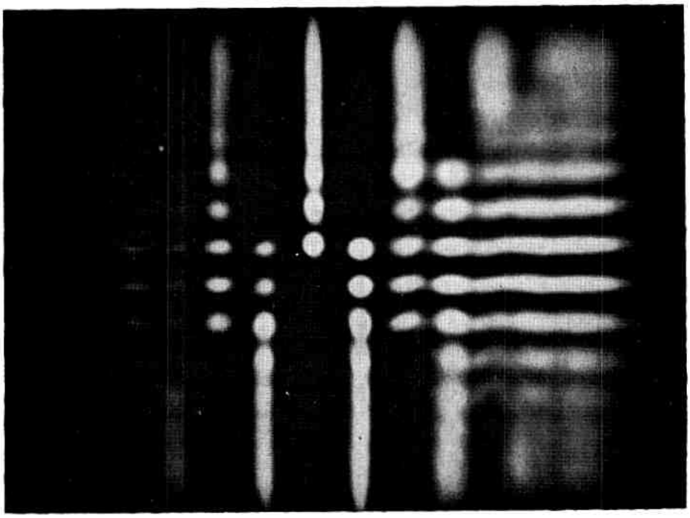


Fig. 2 — Lissajous pattern of one beam photographed at the back of the rear mirror.

the beam size is smallest in the center of the pattern. The beam widens horizontally when it is displaced horizontally and widens vertically when it deviates vertically. Consequently, the beams have elliptical cross sections everywhere except along the pattern diagonals. The ratio of the maximum to the minimum beam radius is

$$\frac{u}{v} = \frac{N^{\frac{1}{2}}}{\pi} \tag{5}$$

for the optimum design; consequently $u = 5.9$ mm.

To recover the beam after 342 round trips without interference from other paths, the Lissajous patterns on the mirrors must form 18 lines and 19 rows of spots as shown in Fig. 2. The spacing of these lines and rows decreases toward the pattern edges; the spots overlap as their sizes increase. In the middle of the pattern, the spots are spaced center-to-center 6 mm horizontally and 5 mm vertically. The same spacing holds for the spots around the center hole of the front mirror. A better output beam was obtained from the rectangular arrangement shown in Fig. 2 than from one with equal horizontal and vertical spacing. A possible cause of this is discussed in the Section III.

Figure 2 shows an increase in the pattern brightness from right to left caused by the nonuniform mirror transmission, which does not reflect a variation in beam intensity. The total loss for 342 round trips was 4.0 dB or 0.135 per cent per reflection. This loss is about three times that of the best mirrors reported.[6] Unfortunately, the reflection maximum of the rear mirror was not exactly centered on the 6328 Å laser line, and the coating was not completely uniform across the surface.

The conduit was mounted along the laboratory wall between two concrete tables which supported the ends. The mirrors were inside the airtight conduit. Their position and the astigmatism were adjusted from outside. Without evacuation, convection inside the pipe caused the beam to drift off the exit hole within minutes. A 1-inch fiberglass insulation around the pipe did not improve this situation. After the pipe had been evacuated to a pressure of 3 torr, the proper alignment could be kept for hours.

III. BEAM DISTORTION MEASUREMENTS

In a well-aligned perfect cavity, the input and output beams pass the center hole with the same size and phase front, but with a slight difference in propagation angle. This permits their separation at the focus of the launching telescope. Figure 3 shows a vertical and a horizontal scan of the output beam. The scans deviate little from the expected gaussian profiles. The width agrees well with that of the input beam. Obviously the high quality mirrors do not introduce appreciable distortion even after 684 reflections. This agrees with previous observations.[3]

The mirror imperfections might be large enough, however, to make the beam stray from its predicted path. The output beam did not show this deviation, as we could and did correct for it by adjusting the mirrors. But there was some evidence that this effect is not completely negligible. Theory predicts that, with perfect alignment, changing the direction of the input beam only changes the direction

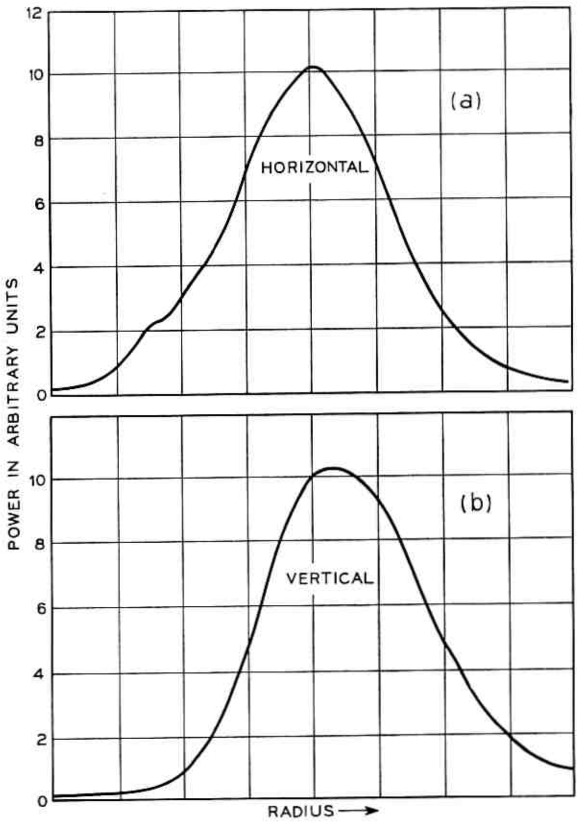


Fig. 3 — Horizontal and vertical scan of the beam after 684 reflections.

of the output beam, but both beams stay centered in the hole. Of course, this operation simultaneously changes the pattern size and either brings the outer paths close to the wall, or the inner paths close to the hole. Before we noticed any interference at the wall or the hole, the output beam would start moving off the hole center when we changed the input beam direction. Comparable experiments with and without astigmatism in a system with fewer round trips suggests that this imperfection is associated mainly with the way the astigmatism is introduced.

Straying from the designed beam path will cause crosstalk in a multiple beam guide. To learn more about this effect, the input was split into a lower beam (1) and an upper beam (2), as shown in Fig. 1. Beam 2 writes a pattern which is the mirror image of Fig. 2 about the horizontal axis. Figure 4 shows the composite pattern written by both beams. The output beams 3 and 4 are separated one above the other at the focus of the telescope and can be recovered separately or together by moving the deflection mirror up or down. The profile of beam 3 is very similar to the one shown in Fig. 3. Figure 5 shows the interference pattern of the output beams displayed on a card in front of the receiver. The straight lines indicate that the phase fronts of the two beams are tilted with respect to each other but are not noticeably different otherwise.

To avoid too optimistic a conclusion from this result, one has to investigate the respective paths of the two beams. To every reflection made by one beam, one can find a reflection by the other which occurs not more than 6 mm away. The effects of small imperfections add up in a commutative way. Consequently, the sequence of reflections is immaterial, and the total distortion of one beam is closely related to the distortion of the other beam because of their neighborhood in the cavity. The nature of this neighborhood is the same as with two beams in a multiple beam transmission system when they are launched and received 6 mm apart.

IV. BEAM SCATTERING MEASUREMENTS

A better analysis of the light output from the cavity is possible when light pulses are injected. This was done by pulsing the laser output at a rate of 1 kHz for intervals of 100 ns using a polarization switch as shown in the foreground of Fig. 1.[7] The pulses were shorter than the cavity round trip time of 174 ns so that the output from successive round trips could be resolved. The total delay of the primary

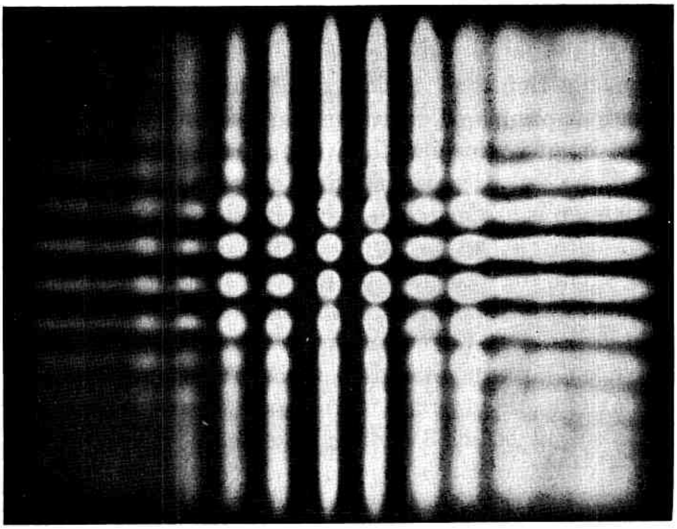


Fig. 4 — Composite pattern of both beams at the back of the rear mirror.

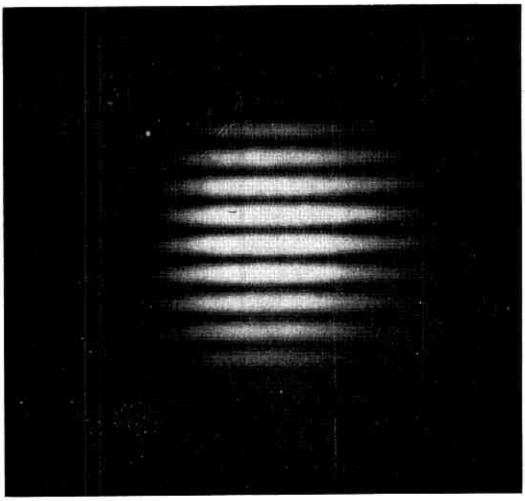


Fig. 5 — Interference pattern of the two beams after 684 reflections.

output pulse was 59.5 μs, confirming the projected number of 342 round trips.

Much weaker pulses were detected before and after the primary output pulse at periodic intervals corresponding to 18 or 19 round trips. By alternately blocking beams 1 and 2, we could attribute some of these pulses to beam 1 and some to beam 2. Blocking beam 2 avoids the strong primary pulse in beam 3 so that the weak pulses can be amplified without saturating the photomultiplier.

Figure 6 shows pulses generated by beam 1 which leave the cavity along path 2. They were detected by moving the deflection mirror into this path. The numbers indicate the round trips completed before detection. Pulse 342 was caused by the primary output pulse which leaves the cavity along beam 4. Although it is not intercepted by the deflection mirror, some scattering outside the cavity resulted in a weak light pulse in the receiver. The other pulses can be attributed to scattering inside the cavity. Investigation of the Lissajous pattern shows that the beam path tends to approach the center hole whenever 18 to 19 round trips are completed. The occurrence of scattered pulses with this periodicity suggests that the beams close to the center hole are responsible for the scattering.

Figure 7 is a sketch of the area around the center hole as viewed from the back of the front mirror. The numbers indicate the round trips completed before the respective reflection. The arrows show

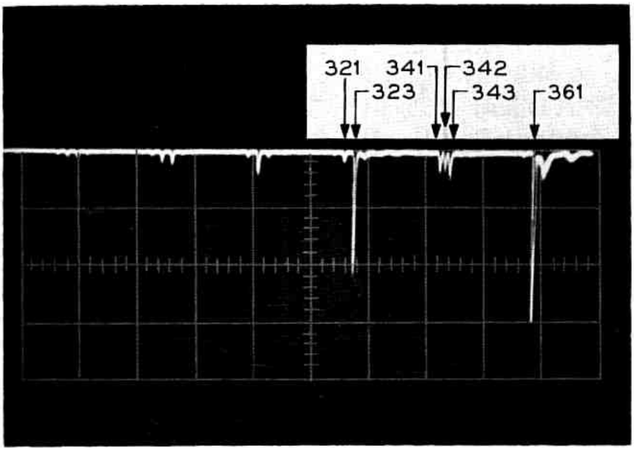


Fig. 6 — Pulse train of scattered light received in beam path 2 when only beam 1 is injected. The numbers indicate the round trips completed before reception.

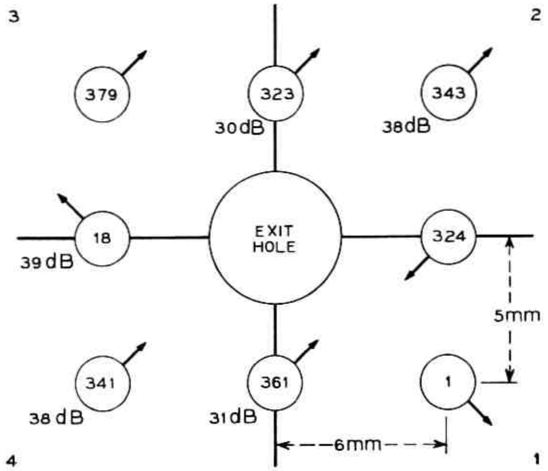


Fig. 7 — The reflections around the exit hole as seen at the back of the front mirror. The numbers indicate the round trips. The arrows point to the quadrant where the scattering is picked up. The dB values represent the ratio of total output to scattered light.

in which direction a particular beam scatters, that is, in what beam path it will be picked up. For example, all dots pointing toward quadrant 2 were received in beam path 2 and are present in the pulse train of Fig. 6. The additional pulses not labeled in Fig. 6 originated from reflections farther away from the hole. They were omitted in Fig. 7 to avoid confusion. The signal in Fig. 6 was calibrated by comparing it with the signal from the primary output pulse reduced by a 40-dB standard attenuator. The dB values in Fig. 7 represent the signal-to-crosstalk ratio obtained by calibrated reception in beam paths 2 and 3.

These observations seem to support the theory that every reflection scatters a small amount of light into a narrow cone about the primary beam.[4] Refocused by the mirrors, this light stays close to the beam; contributions from successive reflections add in power. This is why, after 323 round trips, 9 dB more scattering was measured than after 18 trips, though at that time both beams are the same distance from the hole. Notice that the dB levels indicated in Fig. 7 are related to the power of the output beam. If we consider the attenuation and relate the scattering levels to the beam powers at the respective reflections, the scattering is 42.8 dB below that power after 18 round trips and 39.2 dB after 323 round trips. The difference is 12.6 dB, that is, the scattering has increased 18 times from the 18th to the 323rd round trip, or about proportionally to the number of round trips. If the scattering could be measured after one reflection, the scattered power should be 30 dB + 10 log 646 = 57 dB smaller than the total power of the primary beam.

This, of course, holds only for the specific arrangement shown in Fig. 7: an output hole 5 mm in diameter and a beam being displaced by about 5 mm from this hole. If the reflection occurs 1.56 times farther away from the hole (for example, reflection 341 or 343), the scattered power intercepted by the hole is about 8 dB smaller. From this it is concluded that the scattered power density decreases with about the fourth power of the distance from the hole.

This result is subject to the specific measuring arrangement used, in particular the directional properties of the receiver. In our case, because of the deflecting mirror, the receiver collected one-quarter of the cone of light falling through the exit hole. The observation of a rapid fall-off of the scattered light around the primary beam agrees with measurements reported in Ref. 4. The fact that scattered signals occur even after the primary beam has left the cavity means that the cone of scattered light, though intercepted partly by the exit hole, keeps travelling around in the cavity.

V. AN EQUIVALENT MULTIPLE BEAM GUIDE

Envisage the two cavity mirrors to be replaced by a sequence of thin lenses with the same focal length and spacing. Consider the beam to be unfolded along this path. Periscopic mirror arrangements could be used as well as lenses.[8] Each periscope consists of two mirrors, thus there are two reflections at every focuser, twice as many as in the delay line. Consequently, after traveling through 684 sections, 25 m in length, the beam suffers a loss of 8 dB or about 0.47 dB per km.

In contrast with the delay line, it is not necessary to introduce astigmatism in the multiple beam guide. If the guide is installed above ground, the pressure in the conduit will have to be reduced to a few torr, but evacuation seems unnecessary in an underground installation.[8]

The experiment demonstrated that two adjacent beams show negligible distortion and are fully separated after 684 sections, or 17 km. A receiver area of 5 mm diameter, the size of the exit hole, is sufficient to collect practically all of the beam energy. It might be advantageous to reduce the detector diameter to 3 mm. Such a detector would still collect 90 to 95 per cent of the signal light but less of the light scattered from adjacent beams.

A double mirror periscope will cause twice the scattering of one delay line mirror. A detector with the size of the exit hole at the end of a 648-section guide will consequently receive twice the scattering measured in the experiment, that is, a level of $30 - 3 = 27$ dB for two beams 5 mm apart and $38 - 3 = 35$ dB for two beams about 8 mm apart. The contributions from beams farther away decrease fairly rapidly. If one assumes a decrease with the fourth power of the spacing, one obtains a crosstalk level of about 27 dB for a beam surrounded by equal beams with a mutual spacing of 8 mm or 4 beam widths. Reducing the detector diameter to 3 mm should improve this level to about 31.5 dB.

The level that can be tolerated depends on various aspects of the complete transmission system, but for a comparative figure, regard the composite scattering from all beams as gaussian background noise. Then, with binary envelope detection and no other noise present, a crosstalk level of 20 dB would guarantee an error rate of 10^{-9}.

With this figure in mind, one might consider increasing the transmission distance to 4,000 sections, or about 100 km, allowing a total attenuation of 48 dB. This increases the crosstalk by about 8 dB resulting in a signal to crosstalk ratio of $31.5 - 8 = 23.5$ dB for a mutual beam spacing of 4 beam widths and a detector diameter of 1.5 beam widths. Reference 2 calculates a diffraction crosstalk of about 60 dB for this beam spacing which is completely negligible compared with the scattering effect.

The number of beams that could be transmitted with a mutual spacing of $k = 4$ beam widths in a guide equivalent to the investigated cavity is[1]

$$N_{\text{guide}} = \frac{\pi^4}{32} \frac{A^4}{d^2 \lambda^2 k^4}. \tag{6}$$

For a section length $d = 25$ m, a useful cross section of $A = 6$ cm radius, and $\lambda = 6328$ Å, one obtains about 600 beams. Filling the guide with this capacity, however, requires that the receivers have a better directional selectivity than the one used in the experiment. On the other hand, better selectivity would reduce the scattering received from other beams below what was measured.

VI. CONCLUSIONS

A He-Ne laser beam was injected into an evacuated 25-m delay line and extracted with negligible distortion and only 4 dB loss after 342 round trips. This corresponds to 60 μs delay. The absolute deviation from the ideal path could not be measured, but two beams injected simultaneously were found to be well resolved after 342 round trips.

The light scattered at every reflection from the main beam traveled in a narrow cone about this beam. The power density of the scattered light seemed to decrease with about the fourth power of the distance from the beam. The crosstalk caused by scattering from earlier round trips was 30 dB below the signal level.

A multiple beam waveguide equivalent to this delay line would have mirror periscopes spaced 25 m apart. It could transmit 600 beams over a distance of 100 km with an attenuation of 48 dB and a crosstalk level of 23.5 dB.

VII. ACKNOWLEDGMENTS

We acknowledge the laboratory assistance of Messieurs G. T. Bowes and H. E. Earl and the help given by Mr. H. T. Schulte.

REFERENCES

1. Gloge, D. and Weiner, D., "The Capacity of Multiple Beam Waveguides and Optical Delay Lines," B.S.T.J., *47*, No. 10 (December 1968), pp. 2095–2109.
2. Goubau, G. and Schwering, F. K., "Diffractional Crosstalk in Beam Waveguides for Multibeam Transmission," Proc. IEEE *56,* No. 9 (September 1968), pp. 1632–1634.
3. Gloge, D. and Steier, W. H., "Pulse Shuttling in a Half-Mile Optical Lens Guide," B.S.T.J. *47,* No. 5 (May-June 1968), pp. 767–782.
4. Gloge, D., Chinnock, E. L., and Earl, H. E., "Scattering from Dielectric Mirrors," B.S.T.J., *48,* No. 3 (March 1969), pp. 511–526.
5. Herriott, D. H. and Schulte, H. T., "Folded Optical Delay Lines," Appl. Opt. *4,* (August 1965), pp. 883–889.
6. Gronros, W., and others, unpublished work.
7. Steier, W. H., "Coupling of High Peak Power Pulses from He-Ne Lasers," Proc. IEEE, *54,* No. 11 (November 1966), pp. 1604–1606.
8. Gloge, D., "Experiments with an Underground Lens Waveguide," B.S.T.J., *46,* No. 4 (April 1967), pp. 721–735.

A Companded One-Bit Coder for Television Transmission

By R. H. BOSWORTH and J. C. CANDY

(Manuscript received November 19, 1968)

We compand a one-bit coder by increasing its step size when a string of equal bits is detected in the transmitted code. To code and decode each string we use a weight sequence 1, 1, 2, 3, 5 $\cdots$ 5, the weight returns to unity when the string ends. Stability considerations restrict the choice of weights but those proposed give adequate stability as well as improve the signal-to-noise ratio about 5 dB. The weighted coder has a wide tolerance to changes of input, so that a ±3 dB change from the design value is hardly visible to most observers. Matching weights at the transmitter and receiver is uncritical because mismatches appear as small changes of contrast rather than as noise. The circuit is easily implemented because it is tolerant to changes of component values.

There is a description of an experimental coder and decoder, together with subjective and objective measures of performance. Signal-to-noise ratios of 50 dB are reported.

I. INTRODUCTION

Encoding an analog signal to digital form entails quantization of amplitude. This process introduces a noise into the analog signal that is recovered from the digits. The magnitude of the noise, relative to the signal, is determined by the bit rate in the digital representation and the spectrum of the signal. Successful coder designs make efficient use of the digits, avoiding worthless redundancies, and shape the noise to be subjectively least noticeable.

Delta modulation is one of the simplest and best known coding methods.[1] It changes its analog output positively or negatively by a fixed increment at regular instants, as illustrated by V in Fig. 3. Differential coding is a related method where, at regular instants, the output changes by any one of a set of prescribed values. Delta modulation is regarded as one-bit differential coding because at each sampling time it transmits either of two codes, a pulse or a space, representing a positive or a negative step, respectively. In general, an m-bit coder transmits one of 2^m codes at each sample time.

The advantages of one-bit coding are simplicity of circuitry and a high sampling rate. Thus, for a given bit rate in the digit channel, its sampling rate is m-times greater than that of a corresponding m-bit coder. Although the total noise power from a one-bit coder is greater than that from a multibit coder, much of the power occurs at higher frequencies where it is out of the signal band. The advantage of multibit coding is the ability to grade the step sizes to suit the signal values.[2] Thus, some large steps are provided to track large changes and some small steps are provided to accurately reproduce fine details. In this respect, ordinary one-bit coders are handicapped by having only a single step size.

Theoretical results by J. B. O'Neal show that multibit differential coders have larger signal-to-noise ratios than delta modulators.[3] Practical measurements confirm this and show that much of the advantage comes from companding the quantization levels.* We describe a method for varying the step size of a one-bit coder which has the advantages of both companding and a high sampling rate.

Several authors have described a method for companding delta modulators by changing the step size according to the average pulse rate in the digit channel.[4–7] The steps are smallest when there is an equal number of pulses and spaces; they increase when there is a higher proportion of either pulses or spaces for a significant time. This technique has been used for audio signals to adjust the step size with loudness and pitch of the sound.

For video signals we usually are directly interested in the time dependence of the signal and so require means for adjusting the step size according to instantaneous signal values rather than an average value. Suitable methods have been described by M. R. Winkler and J. E. Abate.[7,8] They vary the step size when certain pulse patterns are detected in the digit channel. Thus, steps are increased when a string of consecutive pulses or spaces are detected. This paper describes the design, construction, and performance of such a coder. It differs from earlier coders in the way step sizes increase and decrease and in that the companding is incorporated in a direct feedback coder instead of a delta modulator. Direct feedback coding, which is reviewed in Section II, is an improvement on differential coding.

* Two-bit coders have insufficient levels to permit adequate companding so they usually are inferior to other coders.

II. DIRECTING FEEDBACK CODING

Direct feedback coders are described in Ref. 9. They function almost the same way as differential coders, but the circuit is arranged to allow greater flexibility of filter design. Figure 1 is a block diagram of a one-bit direct feedback coder and Fig. 2 shows some typical filter characteristics. For television signals the de-emphasis filter H_2 is a short time integrator; the pre-emphasis H_1 is a differentiating filter approximately the inverse of H_2; and the filter A in the feedback loop is a long time integrator.

The feedback acts like a servomechanism trying to make the average value of the quantized signal y equal to the pre-emphasized input x. The difference between x and y is accumulated in A and used to correct the quantized output. The quantized signal in a one-bit coder is observed to oscillate between a positive and a negative level in such a way that its average equals x, as Fig. 3 demonstrates. Changing the pattern of oscillation, the coder interpolates values between the quantization levels, but low frequency components of the oscillation appear as granular noise on the output. The filter A is chosen to make these low frequency components small. High frequency components are de-emphasized by the integrating filter H_2 whose output steps up or down in response to a pulse or a space as does the output of a delta modulator. The advantage of direct feedback coding is flexibility in choosing the deemphasis H_2 independent of the interpolation process which is controlled by the feedback loop.

Notice in Fig. 3 how the large voltage spike in x overloads the quantizer by exceeding its quantization level. The coder responds with a string of pulses which is the largest signal it can transmit. The resulting distortion of the signal is called slope overload; it is a

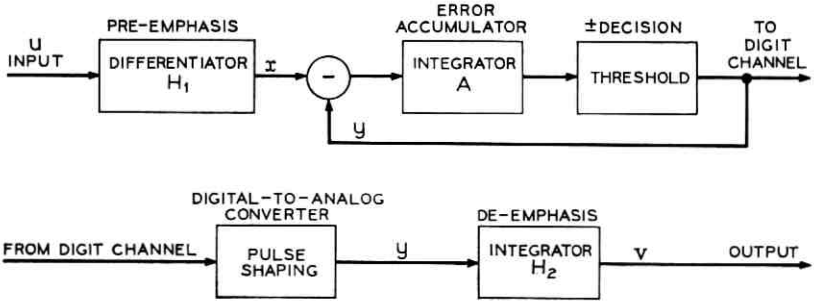


Fig. 1 — A one-bit feedback codec,

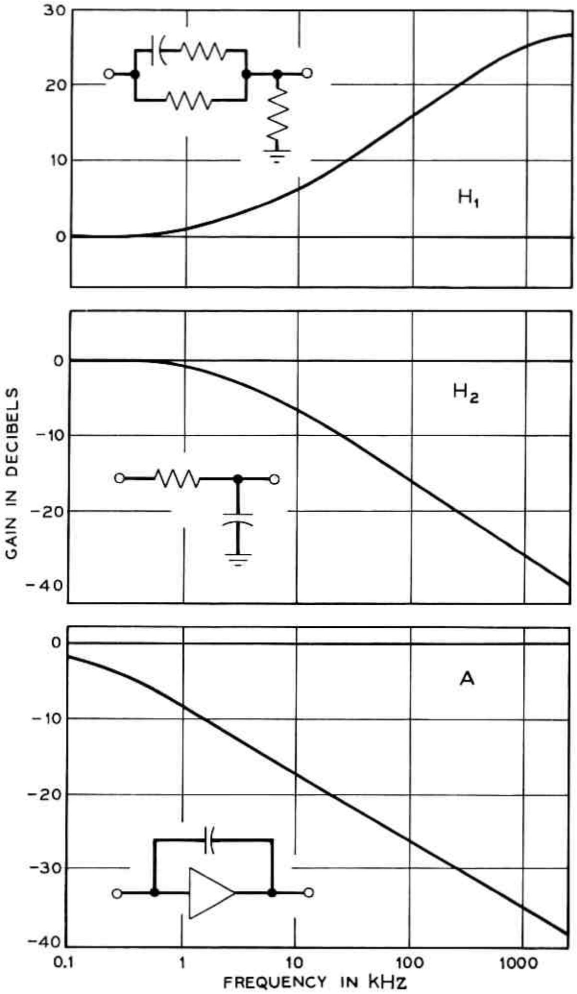


Fig. 2 — Filter characteristics.

characteristic of systems using integrating deemphasis. Usually, the input to the coder is adjusted to a compromise where there is neither too much slope overload at edges nor too much granular noise on "flat" areas.

In the companded coder, overloading is detected by locating strings of pulses or spaces in the code, and then the step size is increased both at the transmitter and at the receiver. This increase extends the range of the coder but increases the interpolation noise in the vicinity of sharp changes. The success of the technique depends on the way step sizes are varied. We have no theoretical criterion for optimizing the formula, instead practical reasons are given in favor of the proposed scheme. The strongest arguments are: the system works, it is easy to implement, and it functions as well as any scheme we have tried for the *Picturephone®* see while you talk service, which is approximately the transmission of a 1 MHz video signal as a 6 MHz binary signal.

III. COMPANDING METHOD

3.1 *The Weights*

Figure 4 shows a block diagram of the proposed coder and decoder (codec). It differs from the ordinary feedback coder by the addition of a weighting circuit in the feedback path at the transmitter and in series with the receiver. The weighting is controlled by a circuit that detects strings of pulses or spaces in the transmitted code. The signal y is then made up of a pulse sequence whose amplitudes depend on the code. The pulses corresponding to the first two bits of each string are left unweighted at the smallest step size. For the third and fourth bits the pulse size is increased two and three times, respectively. For the fifth bit, and all that follow in the string, the pulse size is made five times that of the smallest pulse's value. The string ends when a change of polarity is called for by the appearance of the complementary binary code; then the weight returns to unity. An example of a digit stream and its corresponding quantized signal is given in Fig. 5 which also shows the decoded signal. Compared with Fig. 3 there is a decided improvement in the reproduction of the signal because slowly changing signals are reproduced with smaller steps but the larger signal changes are reproduced with larger steps. Consider the reasons for using this particular set of weights.

3.2 *Choice of Weights*

The plan is to increase the step size when the input changes rapidly. Thus, small steps are used when the differentiated input x is small, and they are increased as x increases. In this way we take advantage of the fact that noise in busy areas of a scene is less noticeable than noise in flat areas.

The step size is left unchanged at its smallest value when no more than two consecutive bits are the same. Such codes are used to transmit

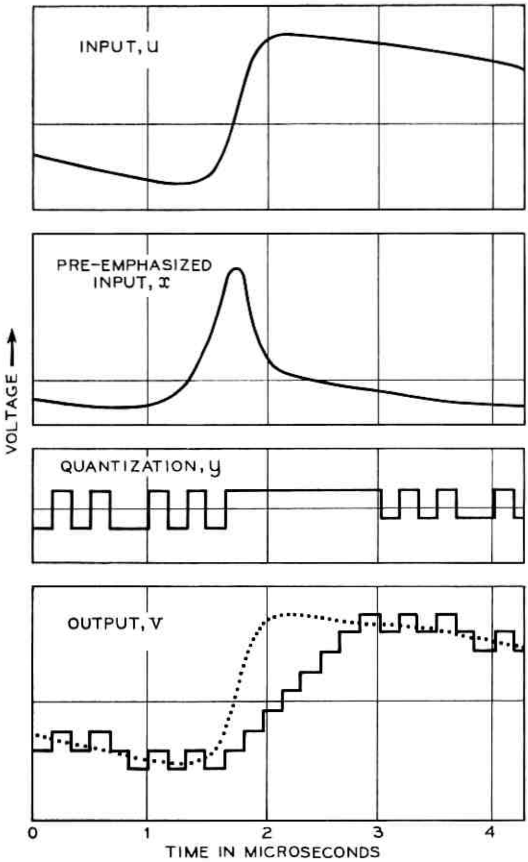


Fig. 3 — Waveforms in the ordinary one-bit direct feedback codec.

the slowly varying inputs that represent the flat areas of a picture. The largest of these codes is $110110 \cdots 110$ and the smallest is $001001 \cdots 001$. They correspond, respectively, to values of y whose average is $+\frac{1}{3}$ and $-\frac{1}{3}$ of the smallest step size. These codes are generated when x has a steady value in the range $\pm\frac{1}{3}$ of a step.

When x exceeds $\frac{1}{3}$ of a step size, codes with more than two repeated bits are generated; the weighting circuit then increases the step size. The signal level in the coder is set so that this occurs only in busy areas of the picture and at edges. Usually, the larger values of x appear as spikes of voltage resembling the one in Fig. 3. Therefore, step sizes should be increased promptly in order to code the transient in a short time; they should be promptly decreased afterwards. Indeed, it is desirable to code sharp changes in video signals in less time than is used to scan three picture elements, otherwise the distortion is objectionable.[9] For *Picturephone®* visual telephone, the edges should be coded in less than 1.5 microseconds, that is, with less than nine bits.

A stability requirement restricts the way weights can be applied. Consider for example, a poor design using a weight sequence 1, 1, 2, 4, 9 · · · 9 to code each string of similar bits. Figure 6 shows an impulse in the voltage x and the subsequent behavior of the quantized signal y: it oscillates continuously between the largest weights after the impulse instead of falling to unity. This oscillation is undesirable because it may increase granular noise in the flat areas of the picture. Figure 6 also shows the response of the proposed coder to an impulse. There is a small undershoot following the representation of the impulse but the step size assumes its smallest value after taking eight bits to code it.

A condition for the weights to fall to their lowest value after any impulse in x is that the weight sequence increase no faster than 1, 1, 2, 4, 8 · · · that is, each weight be no greater than the sum of previous weights in the sequence. The proposed weights 1, 1, 2, 3, 5 · · · 5 satisfy this requirement, giving a safe margin to dampen oscillations.

The weight returns to unity when a string of similar bits end. Then subsequent weight values are independent of the previous code which

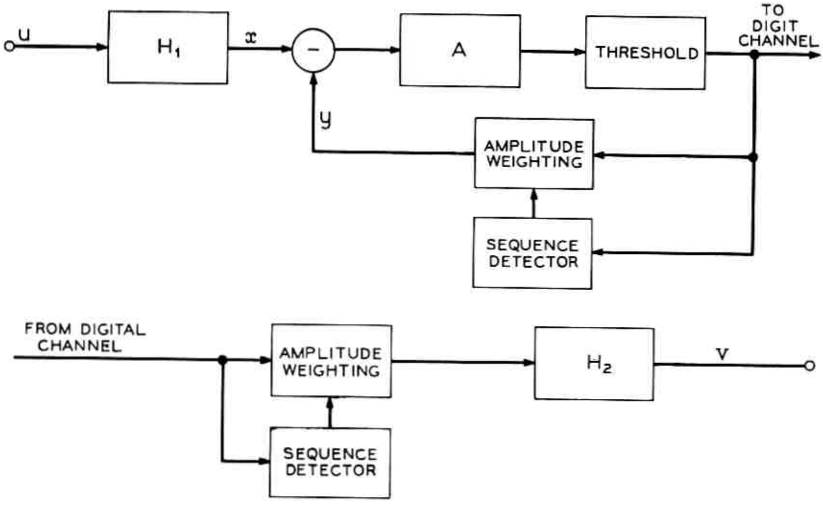


Fig. 4 — The companded codec.

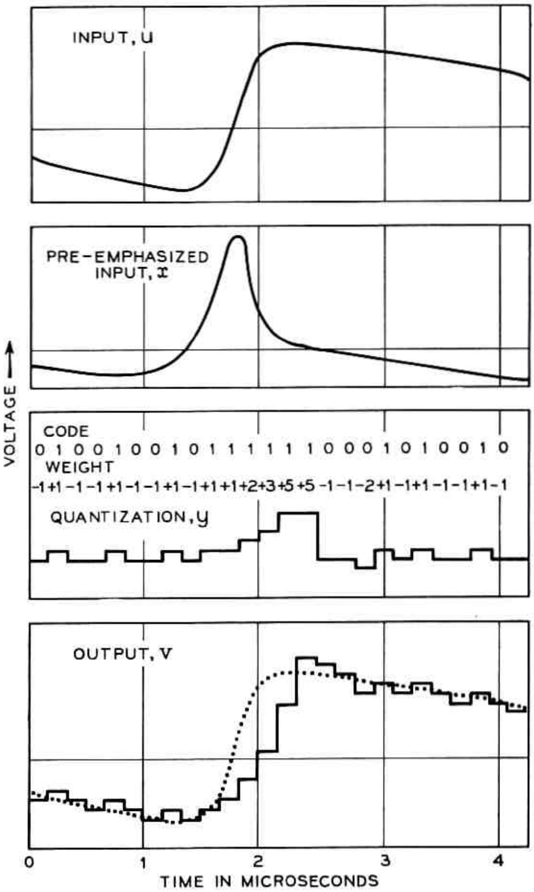


Fig. 5 — Waveforms in the companded one-bit codec.

helps to reduce streaking caused by transmission errors. Properties of a coder are often as dependent on its method of construction as they are on the philosophy of its design. In order that the evaluations be meaningful the circuits are described in the appendix.

IV. EVALUATION OF THE CODER

4.1 *The Test Setup*

For the tests the coder uses a 6.3 MHz sampling rate to code a television signal having 1 MHz bandwidth. This signal represents a 271 line interlaced picture, displaying 30 frames a second. All subjective

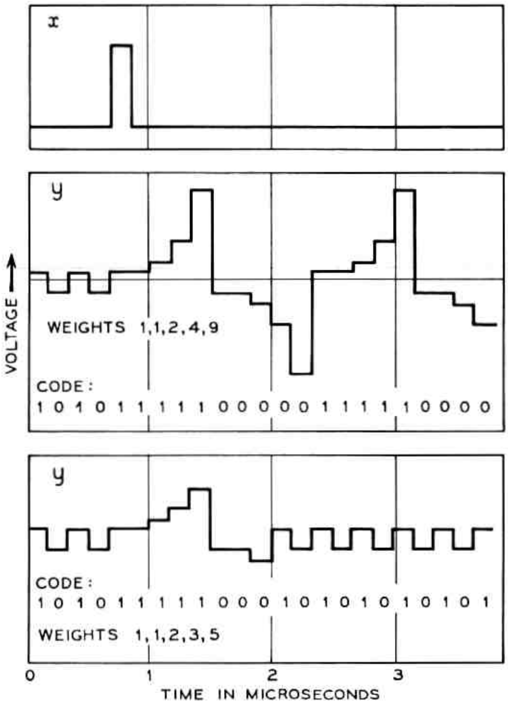


Fig. 6 — Responses to an impulse of x using two different weighting sequences. In normal use the signal will be band-limited so the impulse will be broadened.

tests were carried out using a 5½ by 5 inch display viewed from 3½ feet. The peak luminance was 70 foot lamberts and the room illumination about 100 foot candles.

4.2 *Subjective Tests*

Subjective tests were made by observers who were experienced in picture evaluation and familiar with the coding process. They compared two displays which they switched alternately onto the monitor with equal contrasts. One was the coded picture, the other an uncoded picture with noise added. Each observer varied the noise amplitude until the displays had equal overall quality for him. At this setting the ratio of the signal to the added noise power was recorded as his measure of picture quality. The noise used in these experiments was approximately gaussian with a flat spectrum from 100 Hz to 0.6 MHz as shown in Fig. 7.

The first group of tests concern the signal level in the coder. The

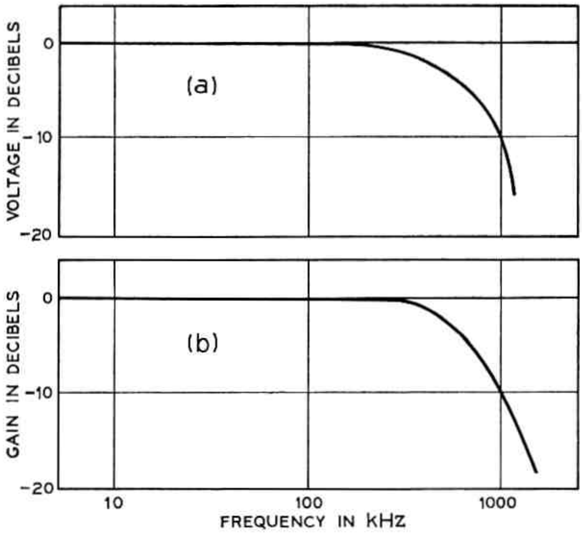


Fig. 7 — (a) Noise source used for testing; (b) Characteristics of filters used for restricting the signal band at the codec input and output.

input amplitude to the coder was varied while a compensating variation at the output maintained a fixed contrast on the monitor. At each amplitude setting an equivalent signal-to-noise ratio was obtained as described. Figure 8 shows the graph of signal-to-noise ratio plotted against signal amplitude into the coder. Results obtained by four observers are given.

Observers agree with one another for small inputs but differ at larger amplitudes where overloading predominates. They all prefer inputs around 70 mV; above this value the quality of the picture falls abruptly because overloading becomes objectional at edges of the scene. When the amplitude of the signal is decreased from 90 mV to 30 mV the picture quality falls slightly as there is a subtle interchange between overloading and granularity. Below 30 mV the granular noise becomes objectional. This graph was obtained using a video signal derived from a back lighted transparency that has unnaturally high contrasts. Figure 9 is a print of this film.

Figure 10 shows an evaluation of a natural live subject. This test was difficult to perform because of the high quality of the coded picture. Observers accept larger inputs (up to 120 mV) because movement makes edge distortion less noticeable. Figure 11 is an evaluation of a transparent resolution chart. This has the lowest signal-to-noise ratio because the peak to root-mean-square value is small, and because a rapid succession of black and white vertical stripes induced oscillation in the weights. But these patterns are unlikely to occur in real scenes: the codes usually transmit pictures of graphic material with little impairment. All of the graphs demonstrate the wide tolerance of the coder to changes of input amplitude.

A result of the second group of tests is shown in Fig. 12 demonstrating the benefits of weighting step sizes. Curve (a) in Fig. 12 is a subjective measure of the ordinary unweighted one-bit coder; the other curves are for the weight values specified on the graph. Notice that at low signal amplitude, where granular noise predominates, the weight has no effect. Weighting only improves the response to large inputs where overloading is important.

The next test concerns the tolerance of the coder to changes of weight values. The sequence 1, 1, 2, 3, 5 ··· 5 was proposed for our application; an attempt was made to find a better sequence experimentally. Figure 13 compares the proposed weights with the best we could find; there is little difference. In fact, the choice of weights is not critical provided they do not cause instability.

The last subjective test concerns the matching of weights at the transmitter and the receiver. Figure 14 shows the equivalent signal-

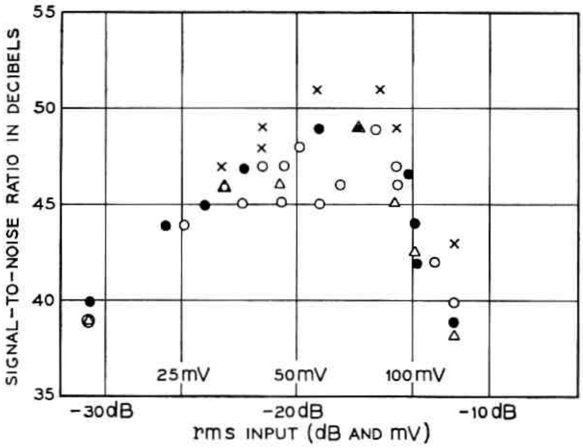


Fig. 8 — A ratio of peak-signal to root mean square-noise plotted against input amplitude. The noise was measured subjectively by four observers. The scene was a photograph of a face.

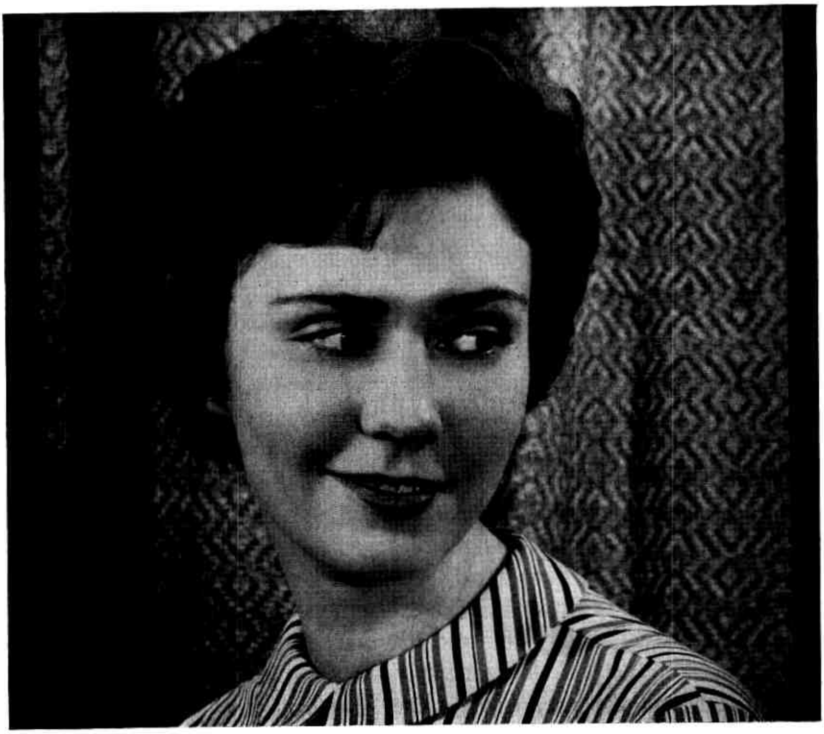


Fig. 9 — The still picture used for the subjective test in Fig. 8.

to-noise ratio of a coder with a 70 mV input and a weighting sequence 1, 1, 2, 3, 5 ⋯ 5, at the transmitter. At the receiver the weighting sequence was*

$$1, 1, (2 + \epsilon), (3 + \epsilon), (5 + \epsilon) \cdots (5 + \epsilon)$$

where ϵ is a controlled variable: it is the absissa of the graph. Graphs for other weight sequences are also given. In all cases the circuit is unusually tolerant of mismatching the transmitter and receiver. Mismatching weights tends to distort the scene in busy areas, rather than introduce noise. This is discussed in the Section A.2 of the appendix.

We have not obtained numerical evaluation of the effect of transmission error. The opinion of most observers is that error probability

* This type of mismatch is consistent with the method of construction where each new weight value is obtained by augmenting the previous one, as shown in Fig. 22.

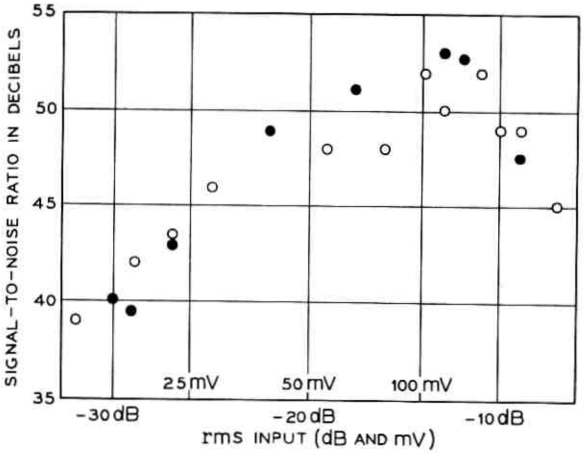


Fig. 10 — Subjective signal-to-noise ratios using a live scene.

less than one in 10^6 is hardly noticeable in a live scene. Errors more frequent than one in 10^5 were troublesome, but the picture was useful with error rates up to one in 10^3. Each error appears as a streak no longer than 0.6 inches with random amplitude. Synchronizing errors were not included because the timing signals were sent on a separate channel.

This subjective measurement is a valuable tool in that it gives more

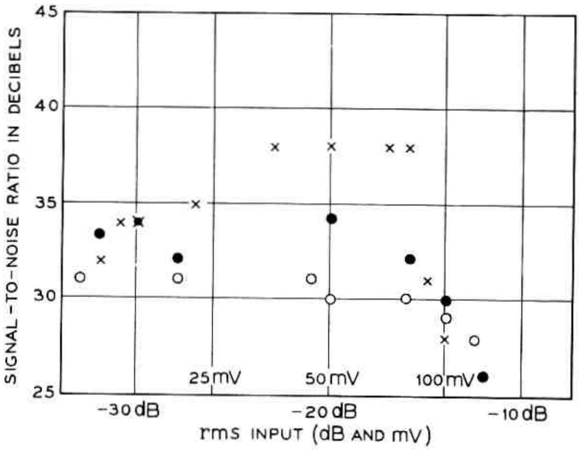


Fig. 11 — Subjective signal-to-noise ratios using a test chart.

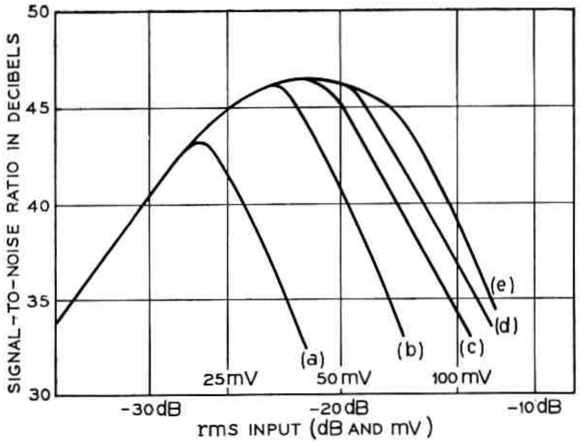


Fig. 12 — Subjective signal-to-noise ratios for various weight sequences by one observer: (a) unweighted; (b) 1, 1, 2 . . . 2; (c) 1, 1, 2, 3 . . . 3; (d) 1, 1, 2, 3, 4 . . . 4; (e) 1, 1, 2, 3, 5 . . . 5.

realistic evaluation of the coder than any objective measure we have used. Objective measurement, however, is needed for commercial evaluations. A useful method is a noise loading test.

4.3 *Noise Loading Tests*

Because of difficulty in characterizing video signals and human observers, some theoreticians have considered gaussian noise as the input when determining a coder's signal-to-noise ratio. Their results can be tested with a noise loading measurement. Such a measurement is described here in order to provide a comparison with published figures for other coders and to provide data for theoretical confirmation.

For these tests, gaussian noise with the spectrum shown in Fig. 7a was the coder input; the resultant output power was measured in selected 1 kHz bands. This power comprises a representation of the input with additional noise generated in the coder itself. A band rejection filter was then inserted before the coder to block the applied noise in the frequency band where the measurement is made; the measured power is therefore the noise generated in the coder alone. A signal-to-noise ratio for the coder can be determined from these two measurements. It is an objective measurement of the coder's properties in the particular band of frequency chosen.

Figure 15 gives the objective signal-to-noise ratio at 14 kHz for various weighting. These curves show that the weights have little ad-

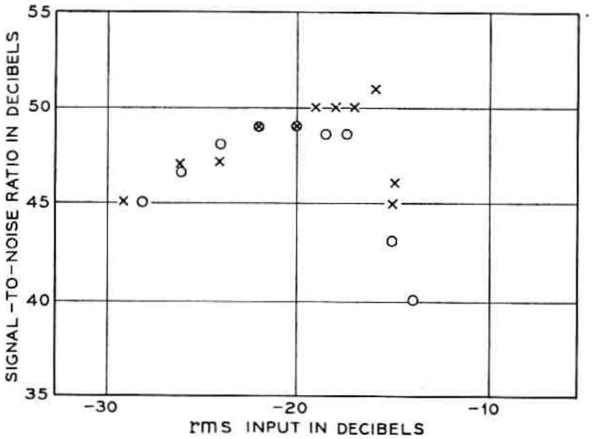


Fig. 13 — Subjective evaluation of two weight sequences: 0 = 1, 1, 2, 3, 5 . . . 5 and x = 1, 1, 2, 3·6, 4·7, . . . 4.7.

vantage for coding this noise. This is not surprising because the weights were chosen to suit the characteristics of video signals—especially the property that large values of the signal derivative occur as a few sharp spikes separated by relatively constant levels, whereas the derivative of the noise has a gaussian distribution. Figure

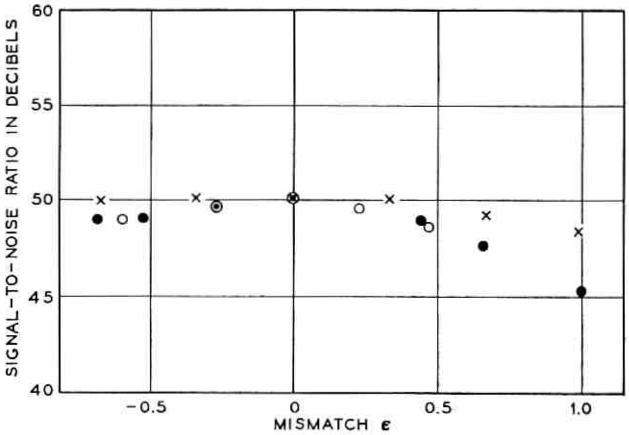


Fig. 14 — Effect of weight mismatch on the signal-to-noise ratio. The weights at the receiver are:
. 1, 1, (2 + e), (3 + e), (5 + e)
x 1, 1, 2, (3 + e), (5 + e)
o 1, 1, 2, 3, (5 + e).

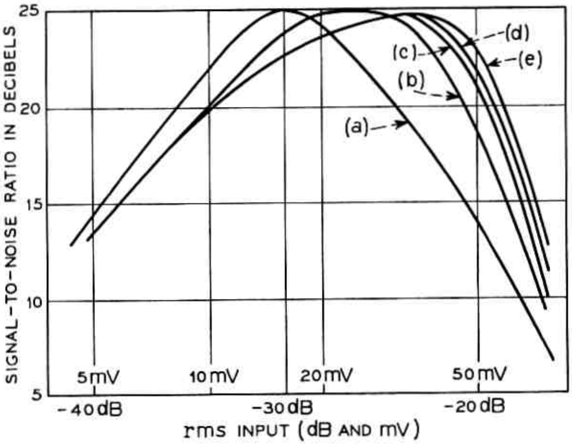


Fig. 15 — Noise loading results at 14 kHz. The weights are: (a) unweighted; (b) 1, 1, 2 . . . 2; (c) 1, 1, 2, 3 . . . 3; (d) 1, 1, 2, 3, 4 . . . 4; (e) 1, 1, 2, 3, 5 . . . 5.

16 shows how the signal-to-noise ratio depends on the frequency at which the measurement is made. By combining this result with the known spectrum of the input, it can be shown that the net signal-to-noise is about 22 dB.

Figure 17 shows signal-to-noise ratios obtained in the same way as those in Fig. 15, but using a video signal as input. These curves more

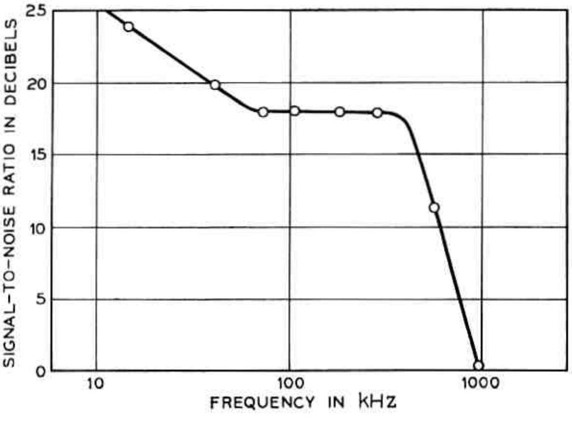


Fig. 16 — Loaded signal-to-noise ratios in 1 kHz slots at various center frequencies.

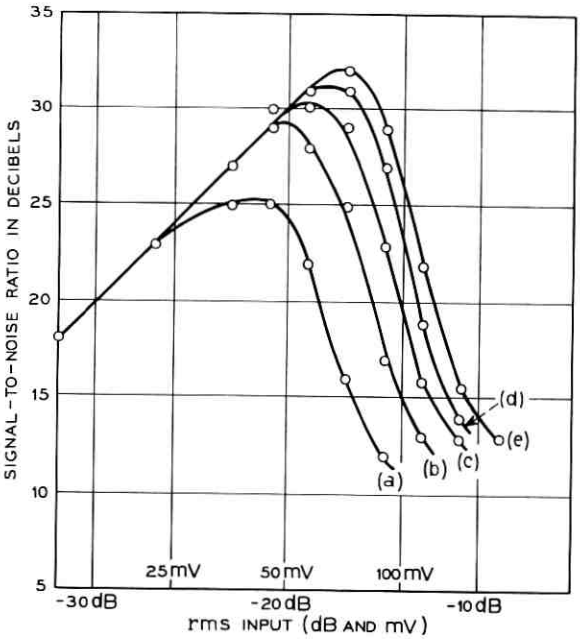


Fig. 17 — Objective signal-to-noise ratios at 14 kHz with video input, using the same weights as in Fig. 13.

nearly resemble the subjective results in Fig. 12. Notice that Fig. 12 refers to peak signals and Fig. 17 to root mean square signals; this accounts for the 11dB difference in the ordinates.

V. CONCLUSIONS

Weighting the step size of a one-bit coder improves the quality of the transmitted signal and broadens its tolerance to changes of input amplitude. The weighting is easily implemented with integrated circuits; in fact, the whole codec need be more complex or expensive than a simple radio receiver. The circuit tolerates up to ±30 percent mismatching of the transmitter and receiver (that is, $\epsilon = 0.3$ in Fig. 14). This is an important property for network applications where each transmission is available to many receivers.

The coder has been presented as a useful circuit for a particular application. No theoretical method for optimizing the companding is known because of difficulty in analyzing a system that incorporates an interaction of a television source, a human observer, quantization, linear filters, and digital processing. Instead, the circuit has been considered intuitively as an extension of the direct feedback coder described in Ref. 9. Indeed, the filters used are those recommended in that work.

We emphasize the tolerance of the circuit to parameter changes because attempts to improve a coder sometimes peak its response about certain parameter values. These parameters are then critical factors in the design. The present coder is very tolerant to changes; this is an important practical advantage.

When each element of a television signal is coded with three bits, the degradation of the picture is subjectively equivalent to about −50 dB of added noise. When the coder was adapted for voice transmission, telephone quality speech could be transmitted using a 50 KHz digit rate. In both examples the coders accepted a wide range (10 dB) of input level.

APPENDIX

The Circuit and Effect of Mismatched Weights

A.1. *The Circuit*

A.1.1. *Circuit Outline*

It is important that the transmission delay around the feedback loop not exceed a sample interval. Otherwise the excess delay will cause a low frequency instability called double moding. Correct operation requires that each decision of the threshold be sent around the feedback in time to fully influence the next decision. Meeting this requirement at high sampling rates is difficult but simplified by moving the weighting circuit, in Fig. 4, outside the feedback loop, as in Fig. 18. Now, each threshold decision activates a switch, S, that sends either of two values to the integrator. These two values have been set up by previous code values held in registers. For this purpose the threshold decision is placed in a flip-flop, F, in readiness for ensuing decisions.

A.1.2 *Circuit Action*

All the components of the feedback loop are dc coupled, enabling the levels in the circuit to be well defined and avoid displacements caused by spurious charges on coupling capacitors.

The timing cycle is given in Fig. 18. When gate T_1 conducts, it

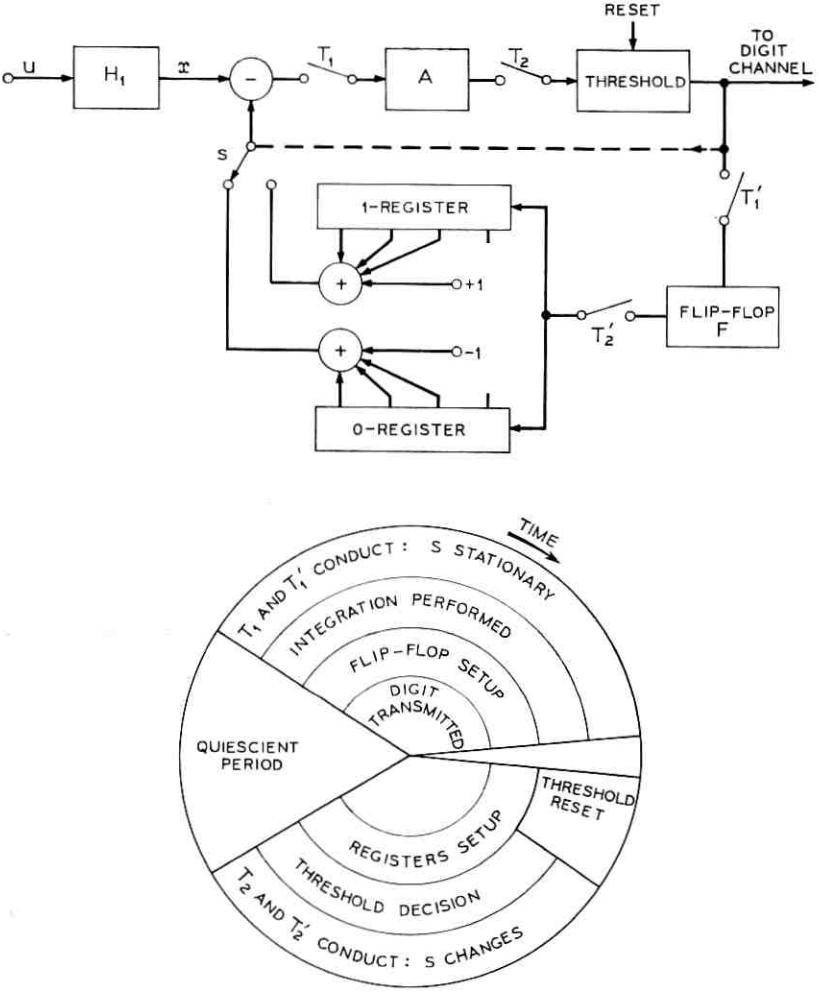


Fig. 18 — The block diagram and timing cycle.

samples the difference between the input and the feedback. It thus defines the pulse width fed to the integrator and isolates the integrator from the input while a polarity decision is being made on its output.

The second gate, T_2, conducts a short while after T_1 switches off. It defines the time in which decisions are made. The threshold circuit is bistable and so holds its decision until reset.[10] Resetting occurs just as T_2 starts conducting. A negative signal applied to the threshold input leaves it in the "off" state; a positive signal switches it on. Once on, the circuit cannot be switched off again from the input terminal.

The output from the threshold circuit sets the switch, S, in readiness for the next conduction of gate T_1. When T_1 conducts, the digit gate T'_1 also conducts, placing the decision in flip-flop F. At this time digit gate T'_2 is off, it conducts at the same time as T_2 transferring the content of F to the registers. A "one" in F resets the 0-register and inserts a "one" into the one-register, shifting up its content. Similarly, a "zero" in F resets the one-register and inserts a "one" into the 0-register. These registers feed signals to adding circuits whose outputs provide the quantized signals, either of which is selected by the next decision, using switch S.

A.2 *Effect of Mismatched Weights*

Any codec needs a digital-to-analog converter at its receiver to assign analog values to the digital code. For the ordinary one-bit codec it is simply a pulse shaping circuit; for multilevel codecs it is more complex, because a variety of different analog values must be generated in response to different code words. The present codec uses a digital-to-analog converter with eight outputs, ± 1, ± 2, ± 3, ± 5 corresponding to different code patterns.

What happens when there is an error in one of the levels generated at the receiver? Every time the code calls for that level, the output will be wrong. When use of each level is completely determined by the instantaneous input the error is a distortion, or nonlinearity, of the output. This is a characteristic of straight pulse code modulation. Conversely, when use of a particular level is not determined by values of the input, but is used almost at random, then errors in it appear as noise on the output. This often happens in multilevel differential and feedback coders. For the companded one-bit coder described, there appears to be a high correlation between amplitudes of the pre-emphasized input x and use of particular levels. Mismatching weights are thus, approximately equivalent to a distortion of x.

Distortion of the pre-emphasized signal appears on the output as a distortion of edges and busy areas. The errors persist for about 3 microseconds which is the time constant of the de-emphasizing filter. The visible effect of small errors is not displeasing; it resembles a change of contrast in the busy areas, and sometimes, a little streaking near the edges.

If the weight sequence used is one that makes the coder unstable, then the correlation between values of x and use of particular weights is lost, and mismatching weights introduces noise.

VI. ACKNOWLEDGMENT

We gratefully acknowledge helpful discussions with our colleagues in Departments 135, 321, and 462 at Bell Telephone Laboratories.

REFERENCES

1. O'Neal, J. B., "Delta Modulation Quantizing Noise, Analytical and Computer Simulation Results for Gaussian and Television Input Signals," B.S.T.J., *45,* No. 1 (January 1966), pp. 117–141.
2. O'Neal, J. B., "Predictive Quantization Systems for the Transmission of Television Signals," B.S.T.J., *45,* No. 5 (May-June 1966), pp. 689–721.
3. O'Neal, J. B., "A Bound on Signal-to-Quantizing Noise Ratio for Digital Encoding Systems," Proc. IEEE, *55,* No. 3 (March 1967), pp. 287–292.
4. Greekes, J. A., and de Jager, F., "Continuous Delta Modulation," Philips Research Report 23, 1968, pp. 233–246.
5. Tomozawa, A., and Kaneko, H., "Companded Delta Modulation for Telephone Transmission," IEEE Trans. on Communication Theory, *16,* No. 2 (February 1968), pp. 149–156.
6. Brolin, S. J., and Brown, J. M., "Companded Delta Modulation for Telephony, IEEE Trans. on Commun. Theory, *16,* No. 2 (February 1968), pp. 157–162.
7. Abate, J. E., "Linear and Adaptive Delta Modulation," Proc. IEEE, *55,* No. 3 (March 1967), pp. 298–307.
8. Winkler, M. R., "Pictorial Transmission with H.I.D.M.," IEEE Int. Conf. Record, Part 1, 1965. pp. 285–290.
9. Brainard, R. C. and Candy, C. J., "Direct Feedback Coders: Design and Performance with Television Signals," Proc. IEEE, *57,* No. 5 (May 1969), pp. 776–786.
10. Candy, C. J., unpublished work.

The Silicon Diode Array Camera Tube

By MERTON H. CROWELL and EDWARD F. LABUDA

(Manuscript received November 26, 1968)

A new electronic camera tube has been developed for Picturephone® *visual telephone applications; with minor modifications it should also be suitable for conventional television systems. The image sensing target of the new camera consists of a planar array of reversed biased silicon photodiodes which are accessed by a low energy scanning electron beam similar to that used in a conventional vidicon. This paper presents a description of the operating principles and an analysis of the sensitivity and resolution capabilities of the new silicon diode array camera tube.*

We also give the detailed experimental results obtained with the tubes. The gamma of a silicon diode array camera tube is unity and its spectral response is virtually uniform over the wavelength range from 0.45 *to* 0.90 *micron with an effective quantum yield greater than* 50 *percent. For a* 13.4 *millimeter square target the silicon diode array camera tube's sensitivity is* 20 *μamp foot-candles of faceplate illumination with normal incandescent illumination or* 1.3 *μamp per foot-candle with fluorescent illumination; with a center-to-center diode spacing of* 15 *micron it's modulation transfer function is greater than* 60 *percent for a spatial frequency of* 14 *cycles per millimeter. Typical dark currents for a* 13.4 *millimeter square target are in the range of* 5 *to* 50 *nanoamperes.*

I. INTRODUCTION

A large number of electronic cameras have been developed for converting an optical image into an electrical signal.[1–3] In many of these, a light-induced charge pattern is stored on a suitable image sensing target and a low velocity scanning electron beam is used to access the charge pattern. One such camera tube, the vidicon, has many desirable characteristics; it has found extensive commercial use partly because of small size and inexpensive construction.[2] However, the vidicon does possess characteristics which, in many applications, can prove undesirable or even detrimental.

Recently there have been several reports of development aimed at obtaining an all solid-state image-sensing system.[4-8] Typically, these systems consist of an array of photosensitive elements scanned by solid-state logic circuits. In general, the technology associated with producing the logic circuits that must duplicate the function of the scanning electron beam is quite complicated. As a result, in all such systems reported to date, the density of photosensitive elements has been rather limited, and the resulting resolution has been small compared with what can be achieved with a vidicon and what would be required in a great many applications of interest.

This paper describes a new camera tube which has the resolution, small size, and inexpensive construction of the vidicon, but not many of its undesirable features. While the vidicon has an evaporated photoconducting film as the image sensing target, the new camera has a planar array of reverse biased silicon photodiodes.[9-12] The diode side of the array is scanned by a low velocity electron beam, and the electron optics are similar to that of a conventional vidicon. Notable improvements in device performance result from the chemical stability of the planar array of silicon photodiodes. This stability insures that the target performance will not be impaired by a high temperature vacuum bake (400°C), necessary for long tube life, or by accidental exposure to intense light images or prolonged exposure to fixed images of normal intensity.

The new silicon diode array camera (SIDAC) tube has three valuable attributes:

(*i*) The spectral response is approximately constant from 0.45μ to approximately 0.90μ with an effective quantum yield of greater than 50 percent.

(*ii*) Electronic zoom can be achieved by varying the size of the raster on the mosiac of diodes since, as discussed in Section 6.2, under the proper operating conditions the scanning beam does not alter the uniformity of the target response.

(*iii*) There is no undesirable image persistance resulting from photoconductive lag.

The first two are unique to the silicon diode array camera tube; the last one is true for the Plumbicon and at high levels of illumination for the vidicon.[3]

Section II discusses the operating principles of the silicon diode array camera tube and Section III analyses its sensitivity and resolution capabilities. The results of the analysis are compared with experiments. The experimental results which are in agreement with the theoretical calculations demonstrate the feasibility of using the tube in systems requiring the quality of entertainment type television. Several alternative modifications of the basic diode array structure that are intended to improve various aspects of its performance are described in Section IV. One of these embodiments, the resistive sea structure, is discussed and analyzed in more detail in Section V. Various other miscellaneous topics, including image lag and dark current are discussed in Section VI.

Details about the target concerning fabrication techniques, X-ray imaging, and other electron imaging applications are described elsewhere.[13,14,15]

II. OPERATING PRINCIPLES OF THE DIODE ARRAY CAMERA TUBE

Figure 1 illustrates the silicon diode array camera tube. The optical image is focused by a lens onto the substrate of the photodiode array. The diode side of the array is scanned by an electron beam that has passed through the appropriate electron optics for focusing and deflection. Deflection is achieved magnetically; focus is achieved either electrostatically or magnetically. In all the experiments to be reported, the interlaced raster scanning period was 1/30 second.

Most of the experimental results given in this paper were obtained

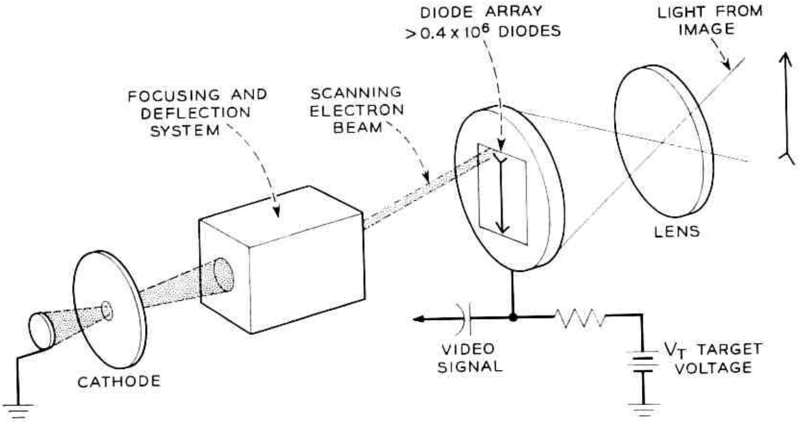


Fig. 1 — Schematic of a diode array camera tube. The electron beam scans the diode side of the array, and the optical image is focused onto the substrate of the array.

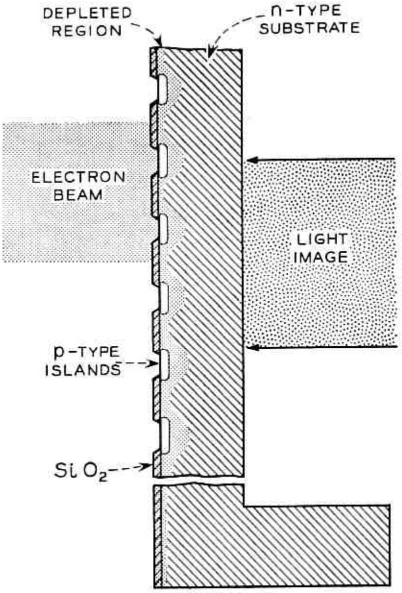


Fig. 2 — Schematic of a diode array target. To obtain a self-supporting structure, the perimeter of the wafer is left much thicker than the substrate in the area of the diode array.

with the target geometry illustrated in Fig. 2. These arrays typically consisted of a matrix of 660 by 660 diodes—about 436,000 diodes within a 0.528-inch square. The substrate is nominally 10 Ω-cm, *n*-type silicon with a diameter of 0.85 inch. The substrate in the area of the diode array is uniformly thick—0.2 to 2.0 mils (5–50μ)—while the perimeter of the wafer is thicker—4 mils—to ensure a self-supporting structure. The diodes, consisting of *p*-type islands in the *n*-type substrate, are formed by standard photolithographic and planar processing techniques.[13] The 660 by 660 array has a center-to-center diode spacing of 20μ and an oxide hole diameter of 8μ. In the early models gold was evaporated over a separately diffused n^+ region to ensure good electrical contact to the substrate. Subsequent experimental results have indicated that a satisfactory contact can be obtained without the evaporated gold.

In normal operation the substrate of the diode array is biased positively with respect to the cathode of the electron gun. The substrate potential relative to cathode potential is called the target voltage and is typically 10 volts. The impinging electron beam thus strikes the mosaic with a maximum energy of 10 electron volts and deposits electrons on both the p-type islands and the silicon dioxide film surrounding the diodes, which isolates the substrate from the beam. Since the resistivity of the silicon dioxide film is very high, the electronic charge accumulates on this surface and charges it to some voltage very close to cathode potential where it remains.

The beam diameter, as indicated in Fig. 2, is generally larger than the diode spacing to eliminate any need for registration between the beam and the mosaic. The electronic charge deposited by a sufficiently intense beam will place a reverse bias of 10 volts on the diodes as it scans over the array. This bias will create a depletion width of approximately 5μ with a 10 Ω-cm substrate giving a junction capacitance that results in an effective charge storage capacitance for the target of approximately 2,000 pF per cm^2. Notice that, at this bias, the silicon surface under the oxide will normally be depleted as indicated in Fig. 2. With very low values of diode leakage currents (less than 10^{-13} amperes per diode) the diodes remain in the full reverse biased condition throughout the entire frame period, if they are not illuminated. The usable values of target capacitance are limited to a narrow range by several factors.[3] For example, the minimum useful target capacitance is determined from the ratio of the required peak video current to the permissible swing in voltage on the scanned side of the array. The maximum voltage swing of the scanned surface is limited by the allowable amount of beam bending which results from transverse (that is, parallel to the surface) electric fields. On the other hand, the maximum capacitance is limited by the charging ability of the electron beam and the image lag requirements placed on the camera. The charging ability of the beam is substantially greater for a higher positive surface potential which is inversely proportional to the target capacitance. In addition, in the diode array camera tube the maximum amount of charge that can be stored is limited by the breakdown voltage of the diodes.[12]

Almost all of the incident light associated with the image is absorbed in the n-type region, each absorbed photon giving rise to one hole-electron pair. Since the absorption coefficient for visible light in silicon is greater than 3000 cm^{-1}, the majority of the photon-generated carriers will be created near the illuminated surface.[16] This will increase the minority carrier (that is, hole) density above its thermal equilibrium value and cause a net diffusion of holes toward the reverse biased diodes. If the lifetime of the holes is sufficiently long and the illuminated surface has been treated properly to reduce recombination effects, a large fraction of the photon-generated holes will diffuse to the electric fields associated with the depletion regions of the diodes and will contribute to the junction current. The light-induced junction current will continue to flow and discharge the junction capacitance throughout the frame period as long as the diodes remain in the reverse-biased condition. Thus, high light levels require high values of reverse-bias voltage or high values of junction capacitance to avoid saturation. The video output signal from each diode is created when the electron beam returns to a diode and restores the original charge by re-establishing the full value of reverse bias. The sensitivity and resolution capabilities of the basic diode array structure are considered in Section III.

It has been found experimentally that the basic diode array structure indicated in Fig. 2 has one rather undesirable characteristic: the silicon dioxide film which insulates the substrate from the electron beam can exhibit uncontrollable charging effects. In some cases the film will accumulate enough negative charge to repell the electron beam and prevent it from impinging on the p-regions. Several alternative modifications of the basic target structure which prevent this charging phenomenon and which improve the performance of the array in other respects are discussed in Section IV.

III. SENSITIVITY AND RESOLUTION CAPABILITIES OF A DIODE ARRAY TARGET

As described in Section II the light associated with the optical image is absorbed in the n-type substrate of the diode array creating hole-electron pairs. The photo-generated holes then diffuse from their point of generation to the depletion regions of the reverse-biased diodes. This section considers the sensitivity and resolution capabilities of the diode array target as determined by hole diffusion and the discrete nature of the diode array, but it does not consider any limitations in resolution resulting from the finite size of the electron beam, aberrations in the light optics, or frequency response of the video amplifiers.

3.1 *Diffusion of Minority Carriers*

An analytical evaluation of the diffusion process in a mosaic target would be quite complicated. In fact, an exact solution would require detailed knowledge of the shape of the depletion regions. However, to estimate the light sensitivity and resolving ability from the simplified model in Fig. 3 is quite straightforward. In this figure,

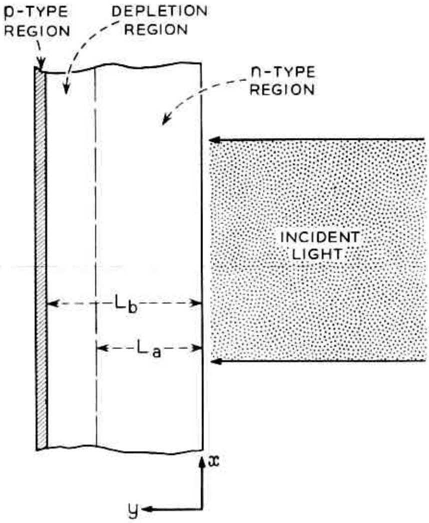


Fig. 3 — Schematic of the simplified model used to estimate the light sensitivity and resolving ability of a diode array target.

the isolated p-regions have been replaced by one homogeneous p-region in which there is no lateral conductivity. This is equivalent to a mosaic structure with zero spacing between diodes. With the low surface recombination velocity normally achieved at the silicon dioxide-silicon interface between diodes or with a fully depleted surface as shown in Fig. 2, the theoretical results obtained from the simplified model should accurately predict the sensitivity of the silicon diode array camera tube. Since the response of the tube is proportional to the incident light level (that is, the gamma is unity) camera sensitivity may be determined by calculating the ratio of the flux of optically generated holes entering the p-region to the incident photon flux.

The steady state diffusion of optically excited holes in the substrate from their point of generation to the depletion regions of the diodes will be governed by the time independent continuity equation[17]

$$-D\nabla^2 p + p/\tau = G(x, y, z) \tag{1}$$

where

p = hole density in excess of thermal equilibrium
τ = minority carrier lifetime

D = hole diffusion constant in n-type silicon
G = hole generation rate per unit volume.

For the model of Fig. 3, the appropriate boundary conditions are

$$\begin{aligned} Sp &= D\frac{\partial p}{\partial y} \qquad \text{at} \quad y = 0 \\ p &= 0 \qquad\qquad \text{at} \quad y = L_a \end{aligned} \tag{2}$$

where S is the surface recombination velocity for holes at the illuminated surface. Setting $p = 0$ at $L = L_a$, the edge of the depletion region, is valid since the electric field prevents any accumulation of holes by quickly sweeping the holes across the depletion region.

The problem being considered here is similar to the one analyzed by Buck and others;[13] however, it does differ in two significant respects. First, our calculation takes into account carrier generation in the depletion regions of the diodes whereas Buck's analysis, intended for short circuit current measurements, does not include carrier generation in the junction space charge region. Second, the hole generation rate is permitted to vary in the transverse direction (the x direction of Fig. 3) so that nonuniform incident light intensities can be considered. This permits evaluation of the loss in resolution caused by lateral diffusion of the holes.

If it is assumed that the light incident on the target is stationary, monochromatic, parallel, and varying in intensity only in the transverse direction as

$$(N_0/2)(1 + \cos kx),$$

then the generation function $G(x, y, z)$ will be given by

$$G(x, y) = \frac{N_0}{2}\alpha(1 - R)(1 + \cos kx)e^{-\alpha y} \tag{3}$$

in which

N_o = peak incident photon flux,
α = silicon absorption coefficient at the optical wavelength of interest,
R = silicon reflectivity at the optical wavelength of interest,
k = 2π/[spatial period of the intensity variation in the transverse direction].

This equation does not include the response to infrared light that may be multiple reflected when the absorption coefficient becomes very small (that is, $\alpha L_b < 2$ corresponding to optical wavelengths greater than approximately 0.8μ). With the above generation function, equation (1) can be solved, subject to the boundary conditions given in equation (2), for the hole distribution in the substrate. The hole flux entering the p-region, $J_p(x)$, can then be obtained by evaluating the hole diffusion current density that enters the depletion region and adding to this the number of holes per unit time and area created by photons absorbed in the depletion region. The result may be written in the form

$$J_p(x) = (N_0/2)\{\eta_0 + \eta_k \cos kx\} \tag{4}$$

with

$$\eta_k = \frac{\alpha L(1-R)}{\alpha^2 L^2 - 1}\left[\frac{2(\alpha L + SL/D) - (\beta_+ - \beta_-)\exp(-\alpha L_a)}{\beta_+ + \beta_-} - (\alpha L)^{-1}\exp(-\alpha L_a)\right] - (1-R)\exp(-\alpha L_b), \tag{5}$$

$$\eta_0 = \eta_k \,|_{\,k=0}$$

and in which

$$\begin{aligned}
\beta_\pm &= (1 \pm SL/D)\exp \pm (L_a/L),\\
1/L^2(k) &= 1/L_0^2 + k^2,\\
L_o &= \text{diffusion length} = (D\tau)^{\frac{1}{2}},\\
L_a &= \text{thickness of undepleted region},\\
L_b &= \text{thickness of the } n\text{-type region plus the width of the depletion region.}
\end{aligned}$$

Notice that η_o is the ratio of the flux of optically generated holes entering the p-region to the incident photon flux for uniform illumination ($k = 0$).

The existence of a "dead layer" and an electric field associated with the illuminated surface, as discussed by Buck and others, invalidates the field-free continuity equation in a small region near the illuminated surface.[13] Consequently, at the shorter wavelengths ($< 0.5\mu$), measured sensitivities may be less than that predicted by equation (5).

With the above reservation in mind, η_o versus optical wavelength for various values of the target parameters can be obtained from equation (5). For the results to be presented, the thickness of the undepleted portion of the substrate L_a was assumed to be 15μ. As Section 3.1 shows, this is a practical value since the maximum value of

L_a for an operating camera tube will be determined by the resolution requirements. The width of the depletion region was assumed to be 5μ which is appropriate for a 10 Ω-cm substrate with a target bias of approximately 10 volts. The wavelength dependence of α, the absorption coefficient, was obtained from the data of Dash and Neumann while the measured wavelength dependence of R, the reflectivity as given by the solid curve in Fig. 4, was used.[16]

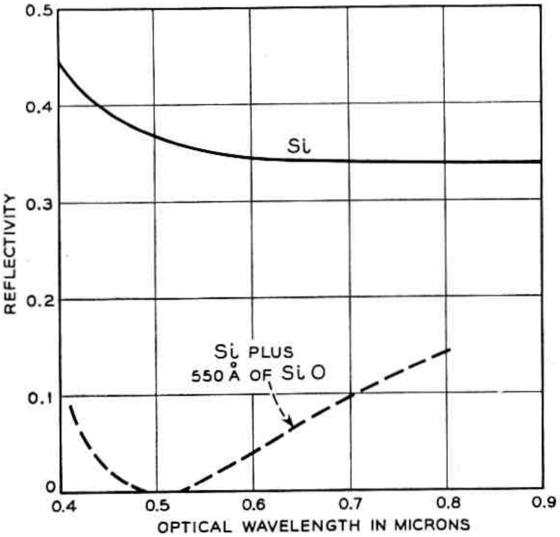


Fig. 4 — Reflectivity versus optical wavelength of a bare, polished silicon surface and of a polished silicon surface with an evaporated layer of silicon monoxide. In both cases the silicon was n-type with a resistivity of approximately 10 Ω-cm.

In Figs. 5 and 6, η_o is plotted versus wavelength for various values of L_o/L_a (or equivalently lifetime τ) for two values of S.* As expected, the curves of Fig. 6, corresponding to a surface with a relatively low recombination velocity, are much higher at the shorter wavelengths than those of Fig. 5 which correspond to a surface with a high recombination velocity. Also as expected, η_o becomes independent of τ for $L_o > L_a$.

An inspection of equation (5) in the wavelength range where $\alpha L_a \gg 1$

* The apparent discontinuity in the curves near 0.5μ results from a discontinuity in the dependence of absorption coefficient upon optical wavelength as reported in the literature and is probably spurious.

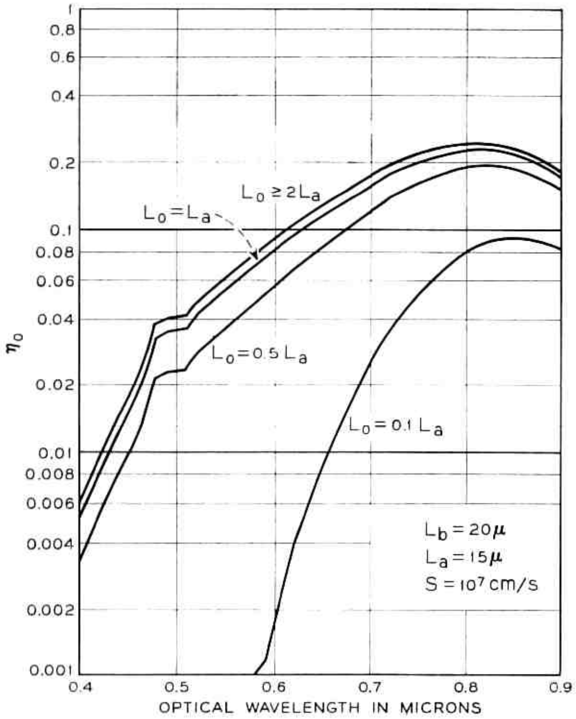


Fig. 5 — Calculated plots of η_o versus optical wavelength for different values of minority carrier lifetime or equivalently different values of the diffusion length L_o and for a high value of surface recombination velocity S.

indicates that η_o will be virtually independent of wavelength, except for the slight wavelength dependence of the reflectivity, if $S/D < \alpha$. This is illustrated by the curves of Fig. 7 which give η_o versus wavelength for various values of S and a given value of τ or equivalently L_o. At the shorter wavelengths (increasing α) the curves are essentially independent of wavelength for values of S less than 10^3cm per second.

Measured values of η_o for three diode array camera tubes with different targets are shown in Fig. 8. These results are in qualitative agreement with the above considerations. The arrays with a low recombination velocity which provide the best response for short wavelengths were obtained by the formation of an n^+ region on the light incident side while the array with a high recombination velocity had an untreated etched surface. In the near infrared (wavelengths > 0.90μ) the thicker array had a higher response. This is not surprising

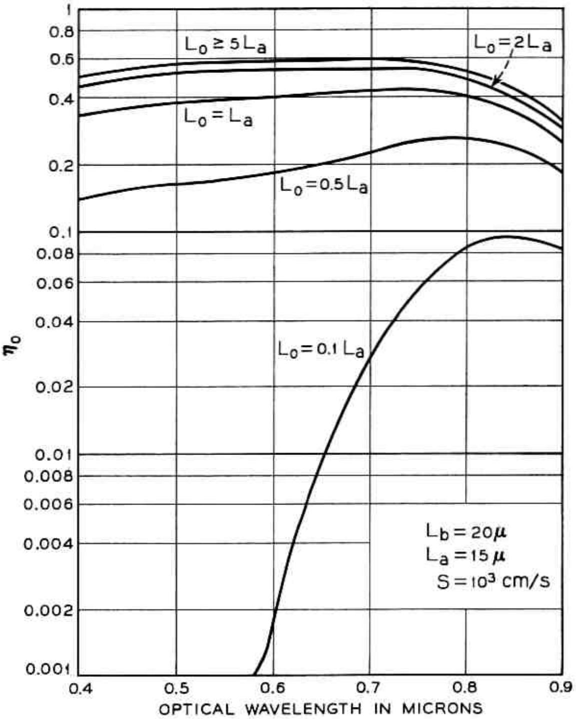


Fig. 6 — Calculated plots of η_o versus optical wavelength for different values of minority carrier lifetime or equivalently different values of the diffusion length L_o and for a low value of surface recombination velocity S.

since the only reason η_o is falling with increasing wavelength in this range is because the photon energy is approaching the bandgap energy of silicon and the substrate is becoming transparent.

The sensitivity of a diode array camera tube is compared with that of a commercially available vidicon and the response of a unity quantum efficiency ($\eta_o = 1$) ideal detector is shown in Fig. 9a. The target of the diode array tube was approximately 20μ thick and had a low recombination velocity on the light incident surface. For the comparison, both tubes were operated with comparable dark currents. The dark current of a typical diode array camera tube is in the range from 5 to 50 nanoamperes. This upper value of dark current is obtained at a target bias of 30 volts in a typical vidicon with visible light response. The vidicon response curves shown in Figure 9 were obtained with this target bias. The light power incident on the tubes was adjusted so that the video output current was approximately equal to the dark current. The conclusion that follows from the curves given in Fig. 9a is that the diode array camera tube has a much broader and a much higher sensitivity than that of a vidicon.

The sensitivity of the diode array camera represented by the curve given in Fig. 9a may also be expressed as approximately 20 μamps per ft-cd of faceplate illumination when the scene is illuminated with an incandescent lamp operating at a normal temperature. The corresponding response of a vidicon with 50 nanoamperes of dark current may be written as approximately 0.6 μamp per ft-cd at a faceplate illumination of 0.1 ft-cd. At this light level, the video signal current of the vidicon is comparable to that of the dark current; because of the photoconductive decay characteristics the image lag in the displayed video may be excessive. For fluorescent illumination the

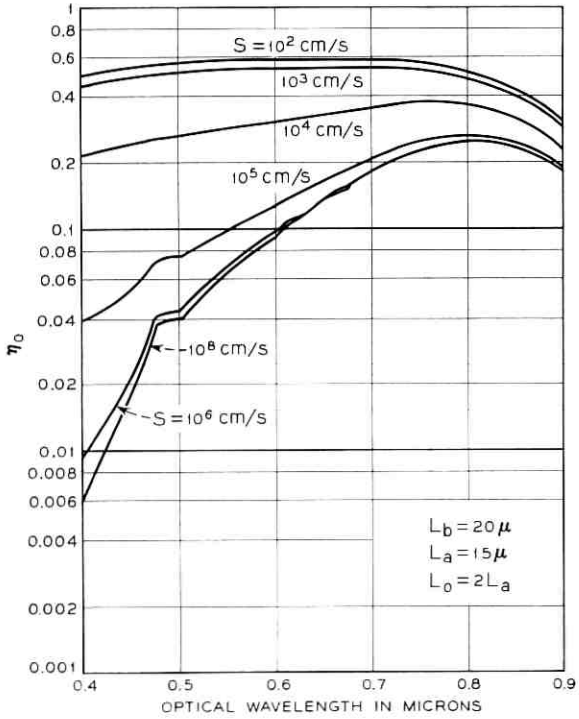


Fig. 7 — Calculated plots of η_o versus optical wavelength for various values of surface recombination velocity.

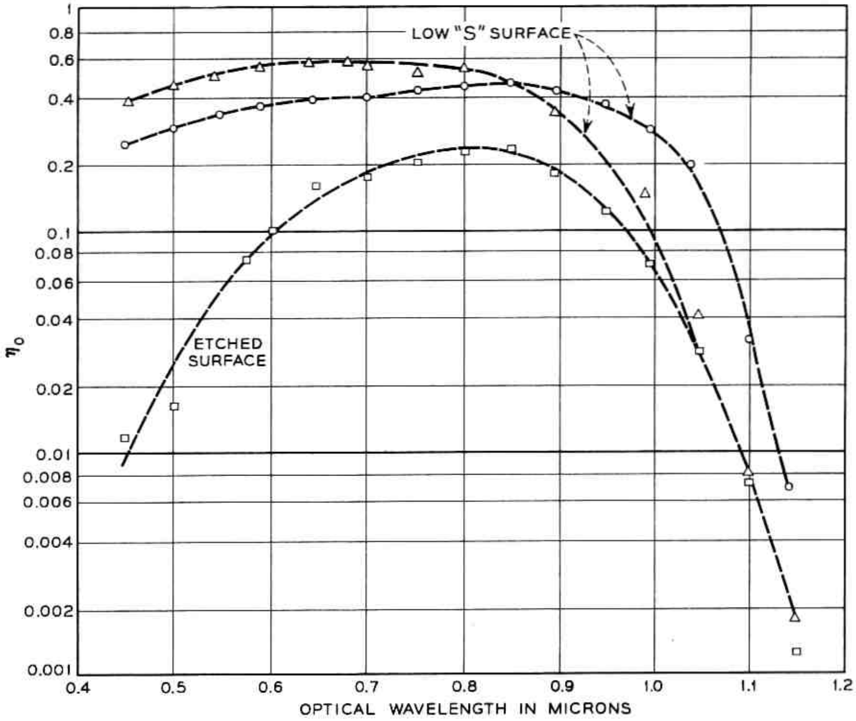


Fig. 8 — Measured values of η_o as a function of optical wavelength for different substrate thicknesses and for targets with a low and high (etched) surface recombination velocity at the light incident surface. *Target Thickness:* ○ Approximately 4.3 mils, △ Approximately 1.0 mils, and □ Approximately 1.0 mils.

sensitivity of the diode array camera is approximately 1.3 μamps per ft-cd of faceplate illumination.

The sensitivity of a vidicon is less at higher levels of illumination since its gamma is approximately 0.65. This is ilustrated in Fig. 9b in which the sensitivity versus output signal current is plotted for a diode array camera tube and a vidicon. These curves were obtained with monochromatic illumination at a wavelength of 0.55μ when both tubes were operated at a dark current of approximately 0.02 μamp. The zero slope of the diode array camera tube results from a unity gamma; whereas the slope for the vidicon corresponds to the value of $(\gamma - 1)$.

The cross-hatched area below 0.02 μamp of output signal current is the region where the dark current is greater than the signal cur-

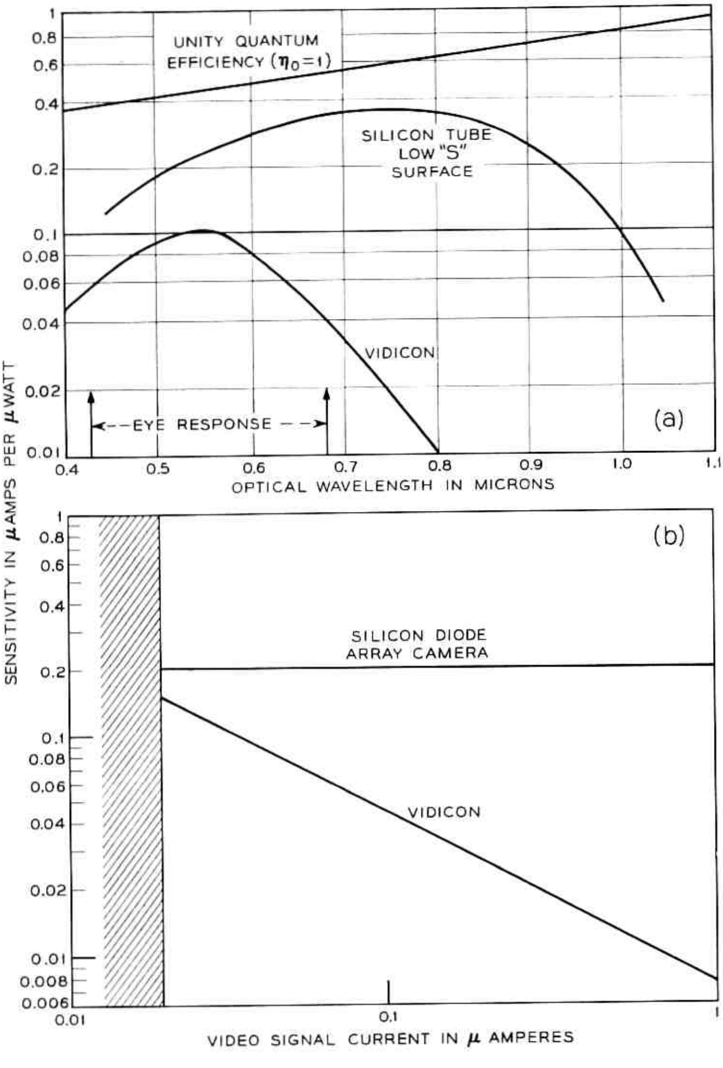


Fig. 9 — (a) Sensitivity of a silicon diode array camera tube as a function of optical wavelength. As a comparison the sensitivity of a commercially available vidicon is also plotted. For the vidicon the signal current was equal to the dark current (0.05 μa) at all wavelengths. (b) Sensitivity of a silicon diode array camera and a vidicon as a function of video signal current: Optical wavelength = 0.55 μ; dark current = 0.02 μ amps.

rent and is considered to be below the operating range of both camera tubes. As the signal current approaches 1.0 μamp, the sensitivity of a vidicon is reduced to a value approximately 1/25 of a diode array camera tube. Notice that the optical wavelength of 0.55μ used for the above comparison corresponds to the peak of the response for both the vidicon and the normal eye.

For the experimental curves of Fig. 8 no effort was made to reduce the reflectivity of the substrate below that given in Fig. 4 for bare silicon which is approximately 0.34 throughout the visible portion of the spectrum. This reflectivity can be significantly reduced by a single-layer antireflection film. As illustrated by the dotted curve of Fig. 4, a film of evaporated silicon monoxide, 550 Å thick, will reduce the reflectivity to less than 0.10 throughout the visible portion of the spectrum. When such a film is used on a diode array target, the target sensitivity is increased by an amount corresponding to the reduction in reflectivity for wavelengths greater than approximately 0.55μ. For wavelengths less than 0.55μ, the sensitivity is also increased but not as much as would be expected from the reduction in reflectivity. The reason is not fully understood but it may be that light absorption in the evaporated silicon monoxide layer is appreciable at these shorter wavelengths.

In the diode array camera, the video signal is normally obtained from the target lead as in a conventional vidicon; as a result, the lowest usable light level will be determined by thermal noise sources in the video preamplifier. This means that in spite of the high sensitivity of the basic silicon diode array camera, its use will be restricted to relatively bright light with presently available commercial preamplifiers. If it is desired to operate at extremely low light levels, the use of return beam reading with secondary emission amplification may improve matters. The minimum detectable light level of an image tube depends upon a number of factors, and the actual determination of this level is beyond the scope of this paper. With return beam reading, however, the minimum detectable light level of the silicon diode array camera would probably be limited by the presently achievable room temperature dark current of 5 to 50 nanoamperes.[18] A modest amount of cooling could be used to reduce the dark current considerably since the dark current drops an order of magnitude for a reduction in temperature of about 25°C.

Consider how lateral diffusion of the photo-generated holes affects the resolution capabilities of the model depicted in Fig. 3. The resolution capabilities of a camera tube are usually evaluated by illuminating the tube with a sinusoidal light pattern, measuring the peak-to-peak video response as a function of the spatial wavelength or frequency of the light pattern, and normalizing with respect to the response for uniform light. For targets with low values of diode leakage currents it is reasonable to assume that the peak-to-peak video signal is proportional to the peak-to-peak hole flux entering the p-region of Fig. 3. Therefore, the modulation transfer function resulting from the hole diffusion process, R_D, is readily obtained from equation (4), the result is

$$R_D(k) = \eta_k/\eta_o .$$

Values of R_D for various values of the target parameters can be obtained from equation (5).

The response $R_D(k)$ will be a function of the wavelength of the incident light pattern, increasing with increasing wavelength as long as multiple reflections in the substrate are not significant. This increase results from the fact that at the longer wavelengths more of the photo-generated holes are created closer to the edge of the depletion region and thus they do not have as far to diffuse. In addition, the existence of a dead layer and associated electric field may result in greater resolution capabilities than predicted by equation (5) when the illumination is restricted to wavelengths less than 0.5μ.

As illustrated by Fig. 10, where R_D is plotted versus spatial frequency $(k/2\pi)$ for various values of L_b, the degradation in resolution contributed by lateral diffusion is a strong function of target thickness. For these curves the width of the depletion region $(L_b - L_a)$ was kept constant at a value of 5μ and the wavelength of the incident light was assumed to be 0.55μ. The quantity $R_D(k)$ will also be a decreasing function of the minority carrier lifetime τ. This is illustrated by the curves of Fig. 11 for which τ has been increased an order of magnitude over the value used for Fig. 10. For the curves of both of these figures a low surface recombination velocity was used because this is a necessity for adequate sensitivity in the visible portion of the spectrum. The resolution and sensitivity will be relatively independent of τ or L_o when $L_a \ll L_o$. However, if $L_o < L_a$, then the sensitivity will increase and the resolution will decrease with increasing L_o, and vice versa.

3.2 *Image Detection with a Mosaic*

The discrete nature of a diode array target places a limit on its resolution capabilities; an estimate of this limit can be obtained if the model depicted in Fig. 3 is modified so that the homogeneous

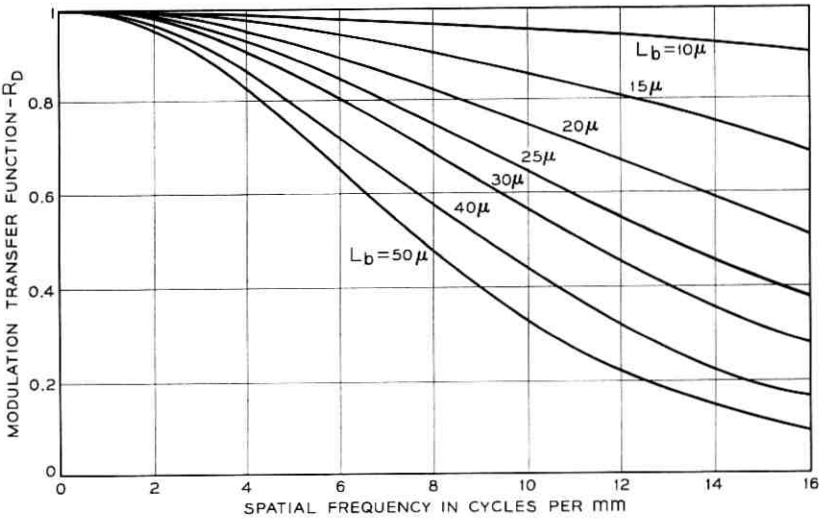


Fig. 10 — Calculated values of the modulation transfer function resulting from lateral hole diffusion, R_D, as a function of spatial frequency for various values of target thickness, L_b, and for a minority carrier lifetime corresponding to a diffusion length of 15μ. ($L_o = 15\mu$; $S = 10^3$ cm/s; $L_b - L_a = 5\mu$; optical wavelength $= 0.55\ \mu$)

p-region is divided into discrete p-islands as indicated in Fig. 12. It will be assumed that the lateral conductivity of each p-type island is infinite.

The resolution capabilities of the mosaic will be obtained by evaluating the peak-to-peak response obtained on the p-type islands for a given incident hole flux. The response of the nth island r_n will be proportional to the total number of holes collected by this p-region; if $J_p'(x)$ is the hole flux, then

$$r_n \propto \int_{(n-1)d_p}^{(n+1)d_p} J_p'(x)\, dx \tag{6}$$

in which $2d_p$ is the center-to-center spacing of the islands. Assuming a sinusoidal variation in the incident light pattern, it follows from equation (4) that

$$J_p'(x) = (N_o/2)[\eta_0 + \eta_k \cos(kx + \varphi)] \tag{7}$$

where φ is a spatial phase factor that accounts for the relative orientation between the mosaic and the light pattern. The peak-to-peak response will be a function of the phase relationship φ between the

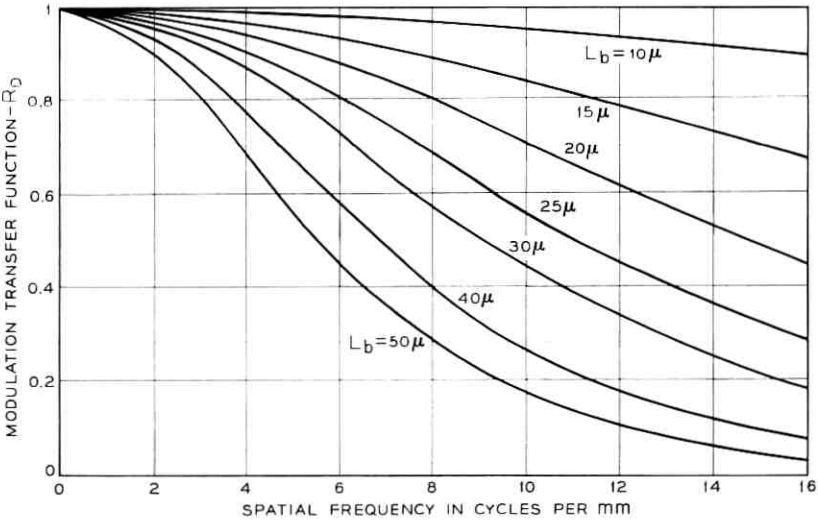


Fig. 11 — Calculated values of the modulation transfer function caused by lateral hole diffusion R_D as a function of spatial frequency for various values of target thickness L_b and for a minority carrier lifetime ten times that used for Fig. 10. ($L_o = 47.3\mu$; $S = 10^3$ cm/s; $L_b - L_a = 5\mu$; optical wavelength 0.55μ)

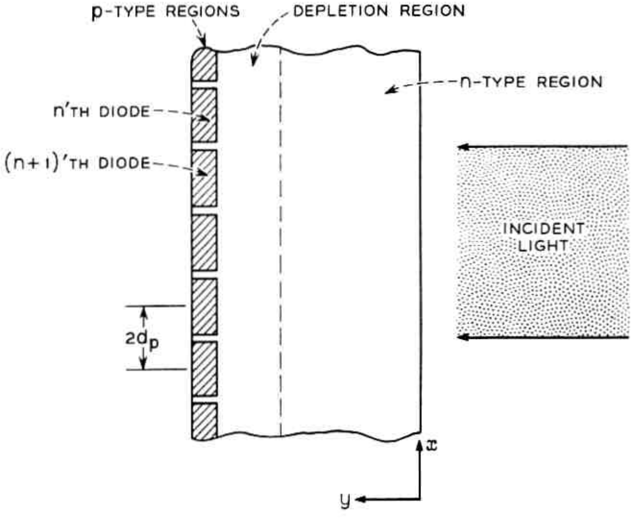


Fig. 12 — Model used to estimate the loss in resolution caused by the finite diode spacing.

light pattern and the mosaic; however, if we restrict our considerations to spatial wavelengths greater than two or three times d_p, the response will be virtually independent of φ. With equation (6) and (7), the peak-to-peak response of the mosaic can be evaluated and if this is normalized with respect to the response for uniform light, the resulting modulation transfer function $R(k)$ is given by

$$R(k) = \frac{\eta_k}{\eta_0}\left(\frac{\sin kd_p}{kd_p}\right) = R_D(k)\left(\frac{\sin kd_p}{kd_p}\right) \qquad \text{for} \quad kd_p \ll 2\pi. \qquad (8)$$

Thus because of the discrete nature of the diode array target, its resolution capabilities are reduced by the factor

$$\sin kd_p/kd_p \,.$$

The effect of this function on the curve of Fig. 10 corresponding to $L_b = 20\mu$ is shown in Fig. 13 for various values of the diode spacing $2d_p$.

In addition to lateral diffusion and the discrete nature of the target, the resolution capabilities of an operating camera tube will be degraded by the finite size of the electron beam. Measured modulation

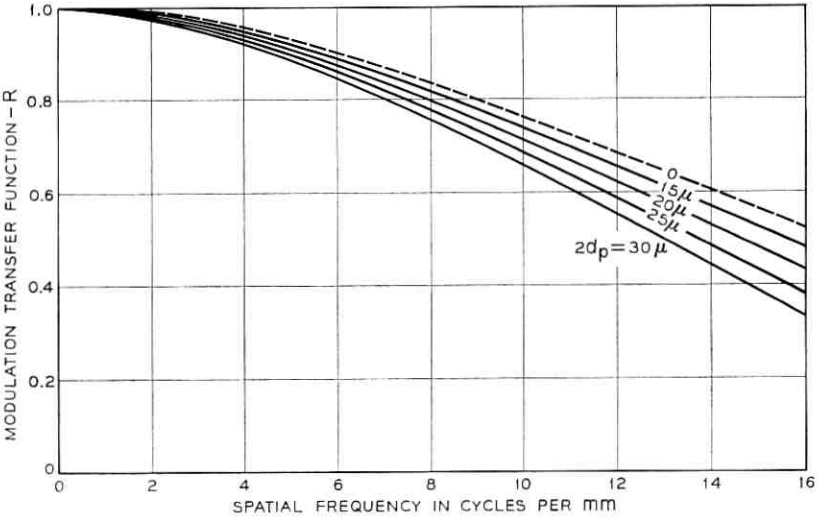


Fig. 13 — Calculated values of the modulation transfer function due to lateral hole diffusion and the finite diode spacing as a function of spatial frequency for various diode spacings. ($L_b = 20\mu$; $L_a = 15\mu = L_e$; $S = 10^3$ cm/s; optical wavelength $= 0.55\mu$).

transfer functions of a typical diode array camera tube for electrostatically and magnetically focused electron beams are given in Fig. 14. Using reasonable estimates of the unknown target parameters, the modulation transfer function can be calculated from equation (18); the results of such a calculation are also given in Fig. 14. The agreement between calculation and experiment is fairly good when a magnetically focused electron beam is used. The increased resolution obtained with magnetic focus compared with electrostatic focus results from a smaller electron beam size.

IV. MODIFICATIONS OF THE BASIC TARGET STRUCTURE

In the basic diode array structure, the silicon dioxide is exposed directly to the scanning electron beam; it has been found that sufficient negative charge can accumulate on the insulating silicon dioxide layer to prevent the beam from striking the recessed p-type islands. The effect of the silicon dioxide film is analogous to that of a control grid in a triode. This section discusses three modifications of the basic diode array structure that will prevent this charging behavior and will improve the performance of the array in other respects.

4.1 *Enlarged Islands*

One modification of the basic diode array structure, identified as a conducting island structure, is shown in Fig. 15. In this structure, electrically isolated conducting islands are placed over each *p*-type region. If the spacing between islands is small enough, most of the silicon dioxide film will be covered with a conducting material so that charging of this surface should be reduced if not eliminated.

Another advantage of the island structure is that the electron beam current is utilized more efficiently. With the typical diode spacing of 20μ and the typical diode diameter of 8μ only approximately $\frac{1}{8}$ of the total beam current is available for producing an output signal if beam pulling effects are neglected. This is simply the ratio of the total exposed area of all of the *p*-regions to the total target area. With the conducting islands, the beam landing area of each *p*-type region is greatly increased and more of the beam current can be used. Reducing the required beam current permits smaller beam diameters to be achieved and as a result the degradation in resolution because of the size of the electron beam may be reduced and possibly the cathode loading may be reduced.

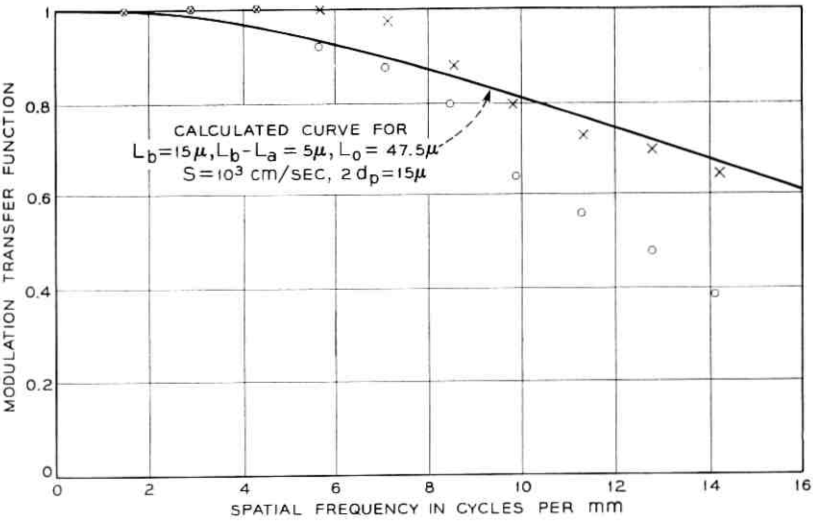


Fig. 14 — Measured values of the modulation transfer function of a diode array camera tube as a function of spatial frequency. Results are given for both a magnetically (x) and electrostatically focused (o) electron beam (optical wavelength = 0.55μ).

The conducting islands will also increase the capacitance shunting the diodes without a corresponding increase in the diode leakage current. Thus the time constant and the charge storage characteristic or dynamic range of each diode will be increased. However, this increase in capacitance must be consistent with the image lag requirements because if the capacitance becomes too large the electron beam may not be able to fully recharge the diode in one scan.

Another potential advantage of metallic conducting islands is that the infrared sensitivity will be increased at wavelengths where a significant amount of light can pass through the substrate. The metallic islands will reflect most of the transmitted light back into the substrate and thus effectively double the absorption path. Furthermore, the metallic islands will also shield the substrate from stray light emitted by the cathode.

Several diode array camera tubes with gold conducting islands have been fabricated. The thickness of the gold islands was about 0.5μ; the minimum separation between islands that has been successfully achieved to date is approximately 3μ. These arrays when examined in a camera tube still showed significant charging effects. These results indicate that with an island thickness of approximately 0.5μ, an island separation of less than 2μ would be required to eliminate the charging behavior. However, stringent requirements must be placed on the photolithographic processes in order to obtain this small separation over the entire array. On the other hand, if the thickness of the islands is considerably increased, it might be possible to use a larger island separation with no deleterious charging effects.

4.2 *Conductive Sea Surrounding the p-Type Islands*

Another attractive target structure, called a conducting sea structure, is illustrated in Fig. 16. In this embodiment the silicon dioxide is covered by a conducting material which surrounds the diodes without contacting the *p*-type islands. This structure should also eliminate charging effects since the silicon dioxide is shielded from the electron beam.

An attractive feature of the conducting sea is that the potential between the sea and the *n*-type substrate can be varied. Thus the silicon surface potential at the silicon-silicon dioxide interface can be controlled and, more important, it can be optimized so as to minimize the leakage current resulting from generation centers at the interface. These centers are the dominant source of dark current in an operating camera tube. Notice that the capacitance between the sea and the substrate is rather large (approximately 6000 pF per cm^2 for an oxide thickness of 0.5μ, assuming no depletion at the interface), and the

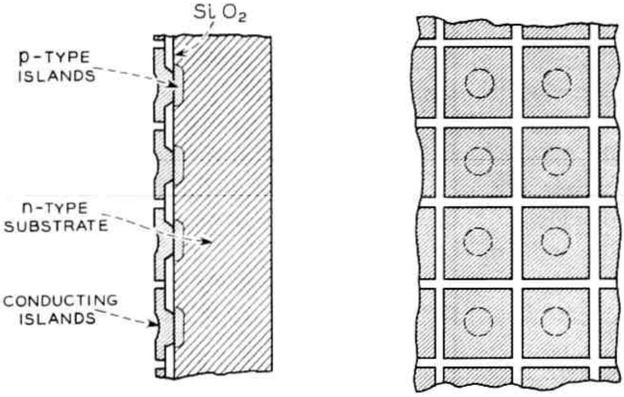


Fig. 15 — Conducting island structure in which electrically isolated conducting islands are placed over each *p*-type region.

high frequency shunting effects of this capacitance must be reduced by the use of a high frequency blocking filter in the bias lead for the sea as shown in Fig. 16.

The conducting sea also has the potential advantage of providing electronic gain. Gain may be obtained by adjusting the bias applied to the conductive sea so that the fraction of beam current which can strike the sea will be modulated by the charge pattern stored on the p-type islands. The video signal is obtained from the parallel combination of the conductive sea and the n-type substrate. When the target is operated in this mode, the performance should be similar to a triode with zero spacing between the control grid and the plate.

The practicality of these advantages depends upon the development of successful fabrication techniques for creating the conductive sea. Thus far inadvertent shorts between the sea and the substrate or between the sea and some of the p-type islands have prevented actual evaluation of a conducting sea structure.

4.3 *Resistive Sea in Contact with the p-Type Islands*

Another technique for eliminating the uncontrolled charging of the silicon dioxide film is illustrated in Fig. 17. In this case, a resistive film or sea covers both the silicon dioxide film and the p-type islands. This resistive sea prevents any build-up of excess charge in the regions between p-type islands by providing a controlled leakage path to the individual diodes. The resistance (that is, ohms per square) of the resistive sea must be chosen judiciously in order to provide this leakage path without impairing the resolution capabilities of the basic diode array target. This implies that there should not be a significant amount of charge leakage between picture elements during a frame period (that is, 1/30 second).

Section V shows that for a silicon dioxide film thickness of approximately 0.5μ, the resistivity of the resistive sea must be greater than approximately 10^{13} ohms per square. This sheet resistivity has been obtained with thin films formed by evaporation or sputtering. Table I lists some of the source materials that have been tried. Since sometimes the process was performed in the presence of a background gas, the composition of the resulting resistive film is not precisely known. The required film resistivity is rather high and one of the biggest problems in obtaining suitable resistive sea structures has been reproducibility. Comments about the reproducibility of the different materials are given in Table I.

Although in many respects Sb_2S_3 works well as a resistive film, it

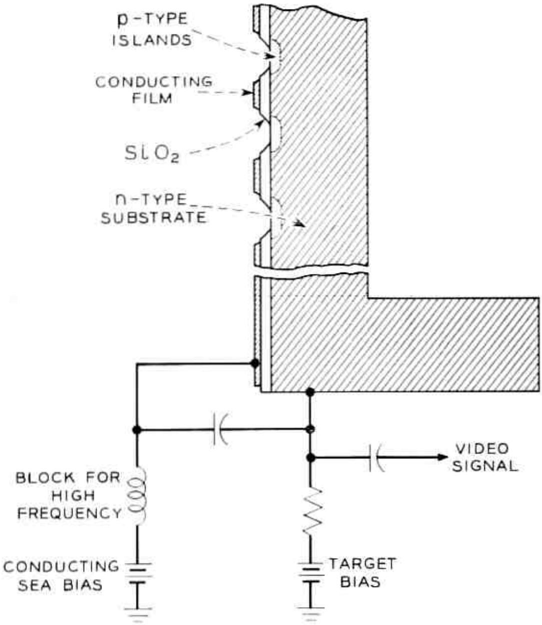


Fig. 16 — Conducting sea structure with suitable bias network. In this structure, the silicon dioxide surrounding the *p*-type islands is covered with a conducting material.

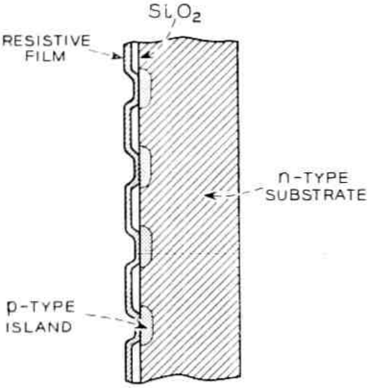


Fig. 17 — Resistive sea structure.

TABLE I — SOURCE MATERIALS

Material	Method of deposition	Obtainable resistivity	Reproducibility
Sb_2S_3	Evaporated	High enough	Small problem
GaAs	Evaporated	High enough	Problem
SiN*	Sputtered	High enough	?
Si*	Evaporated Sputtered	Marginal	Extreme problem
Ha (Ta) N†	Sputtered	High enough	?

* These films were provided by E. N. Fuls.
† This film was provided by F. Vratny.

has one serious drawback. That is, the completed camera tubes with an Sb_2S_3 film cannot be vacuum baked at high temperature. The other materials listed in Table I result in films which permit the camera tube to be baked at 400°C.

Of the many structures and techniques proposed to eliminate the charging problem associated with the basic diode array structure, we have found the resistive sea structure to be the simplest to implement and to date it has given the best results. All of the experimental results presented in this paper were obtained with diode arrays which had a resistive sea.

V. RESISTIVE SEA STRUCTURE

It is quite clear that a resistive film covering the diodes and the silicon dioxide can affect the resolution capabilities of the basic diode array structure and, as pointed out previously, in order for the film not to impair these capabilities, its resistance should be such that there is not a significant amount of charge leakage between picture elements during a frame period. In addition to affecting the resolution, the amount of lateral charge spreading permitted by the film during a frame period will influence many of the other electrical properties of the basic diode array structure.

The following model of the resistive film will be used to establish the required sheet resistivity and to provide a basis for interpretation of experimental results. It will be assumed that the film can be characterized by an effective sheet resistance R_f. This sheet resistance could be a function of the free carrier density spatial distribution in the film if the current flow is space charge limited or if interface states are present at the oxide-film interface. To simplify the present considerations, we assume that R_f is independent of the lateral current flow in the film and that the free carriers in the film are negatively charged and reside at the oxide interface.

Under these assumptions the voltage distribution with respect to the cathode V on the resistive film as a function of time will satisfy the following equation

$$\nabla^2 V = R_f C \frac{\partial V}{\partial t} \tag{9}$$

in which C is the capacitance per unit area between the oxide-film interface and the substrate. This capacitance, which consists of the series combination of the oxide capacitance and the capacitance of the depletion region formed at the oxide-silicon interface, will generally be a function of the difference between the substrate voltage and the film voltage. To obtain a model amenable to analysis the capacitance C will be assumed to be independent of V. This assumption will be valid for a target in an operating camera tube if the maximum amplitude of V is small compared with the target (substrate) voltage.

To obtain some idea of the minimum film resistivity or charge spreading behavior required to prevent a loss in resolution, consider a simple model of the resistive sea structure that neglects the discrete nature of the target. The oxide layer is assumed to be uniform in the lateral direction as indicated in Fig. 18. With such a model, it is possible to determine, for a given stored charge pattern, a minimum

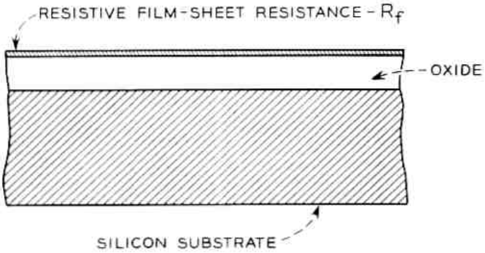


Fig. 18 — The model used for analyzing the resistive sea structure,

value for R_f for which there is no appreciable lateral charge leakage during a frame time.

Since the purpose of the resistive film is to provide a controlled charge leakage path without introducing a concurrent loss in resolution, the decay rate of a given initial charge distribution is the parameter of interest. If we restrict our considerations to a one dimensional charge distribution and a uniform oxide layer, equation (9) becomes

$$\frac{\partial^2 V(x,\, t)}{\partial x^2} = R_f C \frac{\partial V(x,\, t)}{\partial t} \tag{10}$$

in which x is a lateral coordinate parallel to the film.

Let the initial charge distribution at the resistive film-oxide interface be given by

$$q_k(x,\, t = 0) = q_0 \cos kx$$

in which

$$k = 2\pi(\text{spatial wavelength})^{-1}.$$

This charge distribution will create a voltage profile which may be approximated as

$$V_k(x,\, t = 0) = (q_0/C) \cos kx,$$

provided $k \ll C/\epsilon_o$, where ϵ_o is the permittivity of free space. It follows from equation (10) that such an initial voltage profile will decay exponentially with time with a time constant τ_k that is given by

$$\tau_k = R_f C/k^2. \tag{11}$$

Thus the time interval over which a sinusoidal charge pattern may be stored without smearing is proportional to the square of the spatial wavelength of the pattern.

If the decay time of the voltage profile is required to be 10 times the frame period of 1/30 second so that there is only a 10 percent loss in resolution resulting from charge spreading, then the value of R_f must be such that

$$R_f > k^2/3C.$$

For a spatial frequency of 14 cycles per mm, the largest spatial frequency of interest, and assuming $C \cong 4000$ pF per cm^2 (a value which lies between the oxide capacitance and the capacitance of the depletion region under the oxide), it follows that

$$R_f \gtrsim 5 \times 10^{13} \text{ ohms per square.}$$

This implies that for a resistive sea thickness of approximately 0.1μ, a bulk resistivity of at least 10^8 ohm-cm is required for the material of the resistive layer.

Measurements of the decay of an initial voltage distribution have been used to obtain estimates of the sheet resistances of the resistive films. A voltage distribution is created on the resistive sea by focusing onto the camera target a bar pattern (resolution chart) that is illuminated by a light pulse, the duration of which is much shorter than a frame period. The light induced charge pattern and resulting voltage pattern is introduced to the resistive film in discrete areas corresponding to the p-regions of the diodes. If the spatial wavelength of the illuminated bar pattern is much greater than the diode spacing, then the charge over the p-regions will relax into the surrounding areas in a time that is short compared with the relaxation time of the overall light induced charge pattern. For times longer than the relaxation time between diodes, the simple model discussed above should be valid. In the measurements, the peak-to-peak video response is measured as a function of the time between when the bar pattern is illuminated with the light pulse and when the electron beam scans the light induced charge pattern produced on the resistive sea.

Some results obtained from this type of measurement are given in Fig. 19 for targets with different film resistances. During the time between writing and reading, the electron beam was blanked so that no electrons were hitting the target. A square wave bar pattern was used and the dotted lines are calculated curves for the decay of an initial square wave voltage profile using the simple model discussed above. Except for very short times, the agreement between calculation and experiment is very good. At long enough times, only the fundamental component of the square wave contributes to the video signal and the decay is then truly an exponential, that is a straight line on the semilog plot of Fig. 19. The sheet resistances indicated in the figure were calculated from the decay times of the various curves by assuming the effective capacitance between the film and the substrate was the same for all curves and was equal to 4000 pF per cm^2. In the remainder of the paper, when a value of sheet resistance is given, it refers to a value obtained from decay curves as plotted in Fig. 19.

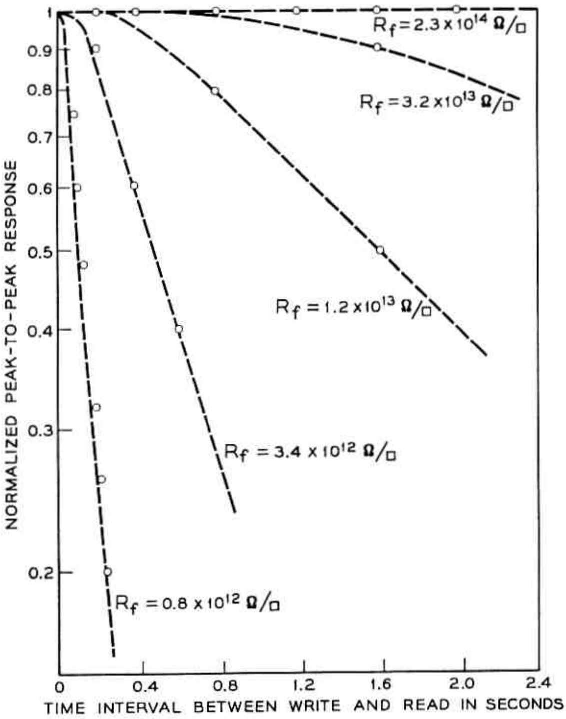


Fig. 19 — Peak-to-peak video response obtained from a bar pattern as a function of the interval between the time when the bar pattern is optically written onto the target with a light pulse and the time when the pattern is read with the electron beam. During the interval between writing and reading the electron beam was not scanning the target. The various curves are for different film resistances. (Spatial wavelength = 0.04 cm; C = 4,000 pF/cm^2) ○ experimental points, - - - calculated curves.

It has been found that targets with low resistivity films (sheet resistances $<10^{13}$ ohms per square) will have certain distinguishing characteristics that are quite different from those observed on targets with high resistivity films (sheet resistances $>10^{14}$ ohms per square). That is, the resistance region between 10^{13} ohms per square and 10^{14} ohms per square is a transition region for the typical diode arrays with a diode spacing in the range of 15 to 20μ. One of the most striking contrasts between targets with a high resistivity film and those with a low resistivity film occurs when the video current through a white defect is observed as the substrate voltage is increased. Most arrays fabricated to date have isolated diodes which exhibit higher values of dark current than their neighboring diodes. This higher value of dark current manifests itself in the displayed video as an isolated bright spot or equivalently a white defect.

The behavior of a target with a low resistivity film is illustrated in Fig. 20. The pictures in this figure are of the video display of the dark current pattern at different target voltages obtained from a camera tube which had many white defects. With the low resistivity film, as the target voltage is increased, the video current through the defects increases, that is, the white spots get brighter and also tend to enlarge only slightly.

Compare this behavior with that exhibited by a target with a high resistivity film as shown in Fig. 21. Here, as the target voltage is increased the video current through the defects again increases, but now when the target voltage reaches a certain critical voltage, the

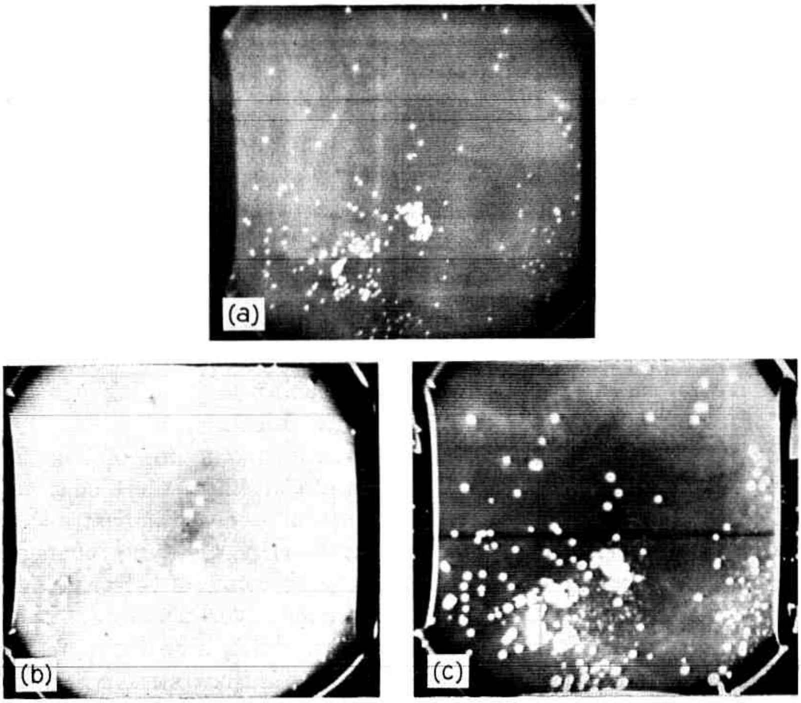


Fig. 20 — Photographs of the video display of the dark current pattern at different target voltages obtained from a camera tube in which the target had a low resistivity resistive film: (a) $V_T = 1.5$ volts, (b) $V_T = 5.0$ volts, (c) $V_T = 12$ volts. (Video display scan lines and printing screens cause moiré patterns in some figures that are not in the originals.)

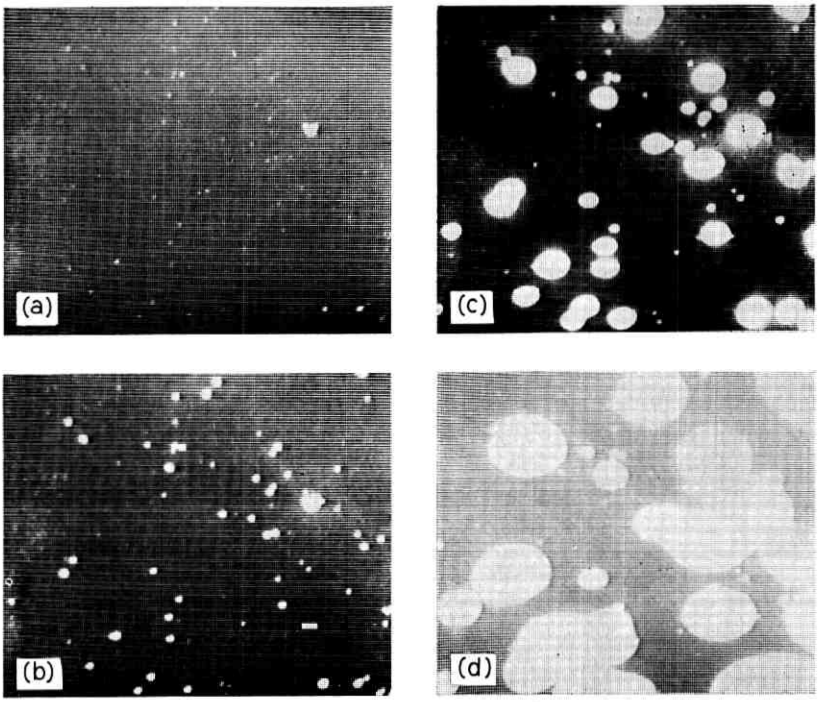


Fig. 21 — Photographs of the video display of the dark current pattern at different target voltages obtained from a camera tube in which the target had a high resistivity resistive film: (a) $V_T = 2.5$ volts, (b) $V_T = 3.0$ volts, (c) $V_T = 3.5$ volts, (d) $V_T = 4.0$ volts.

defects grow larger in the lateral direction very rapidly and eventually envelop the entire target. This enveloping or "whiting out" of the target can result from only one single defect.

The large white regions in the last two photographs of Fig. 21 cover many diodes and correspond to areas in which the diodes are all electrically shorted together. Experimental evidence indicates that these diodes are electrically shorted together by a p-type inversion layer which forms under the oxide and which connects the originally isolated p-regions. The fact that an inversion layer can form with a high resistivity film but not with a low resistivity film turns out to be what one would expect; the reason for this is illustrated in Fig. 22.

In the top part of the figure, the area around one diode is schematically indicated just after the electron beam has recharged the diode. The film potential will be at cathode potential; assuming the target voltage or the potential of the n-region is high enough, approximately 5 to 8 volts for a 10 Ω-cm substrate and an oxide thickness of approximately 0.5μ, the area under the oxide will also be depleted as indicated. Let us further assume that in the vicinity of this diode there is for some reason a high generation rate of minority carriers. This situation could arise for example as the result of some sort of defect in the vicinity of the diode.

The charge collected on the *p*-region resulting from the large generation rate will cause both the potential of the *p*-type island and the potential of the film over the *p*-region to increase from cathode potential towards target potential. What happens now depends upon the charge spreading behavior of the film.

As indicated in the lower left half of Fig. 22, the rise in potential of the *p*-type region for a low resistivity film will be communicated laterally a significant distance during a frame period. Thus the film potential over the oxide increases and as a result both the diode depletion region and the depletion region under the oxide directly surrounding the

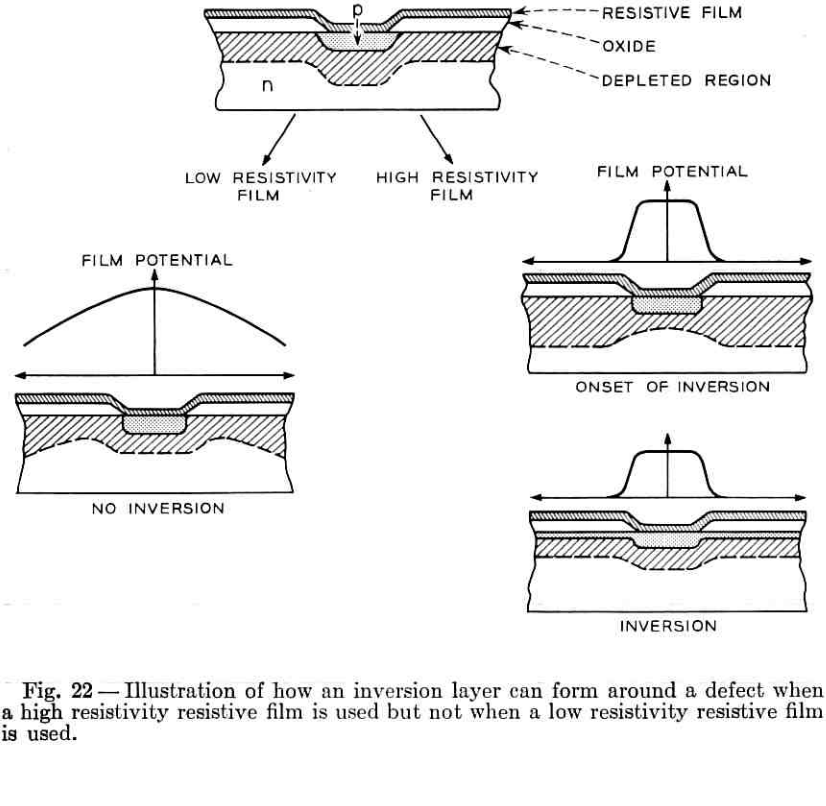


Fig. 22 — Illustration of how an inversion layer can form around a defect when a high resistivity resistive film is used but not when a low resistivity resistive film is used.

p-region will be reduced. The reduction of the depletion region under the oxide inhibits the formation of an inversion layer and no inversion occurs in this case.

On the other hand, as indicated in the lower right half of Fig. 22, with a high resistivity film the rise in potential of the p-region is not accompanied by a rise in potential of the film out over the oxide. Therefore, the depletion region under the oxide will not immediately collapse along with the diode depletion region and an electric field in the lateral direction will be produced which forces holes from the p-region into the depletion region under the oxide, resulting in the formation of an inversion layer. As the experimental results have indicated, the inversion layer can cause many diodes to be shorted together. This behavior is similar to the shorting together of the source and drain of an insulated gate field effect transistor by the application of the appropriate voltage to the gate electrode.

Thus with a high resistivity film we have the possibility of inversion layers forming at a defect whereas with a low resistivity film the lateral charge spreading inhibits the formation of an inversion layer.*

Besides influencing the target properties discussed above, the resistive sea also affects the ability of the electron beam to re-establish the full value of the reverse bias on a diode during one scan.[15] Some insight into this problem can be obtained from the equivalent circuit shown in Fig. 23 which approximates one of the diodes. In this figure the p-n junction is represented by a schematic diode which is shunted with an effective junction capacitance, C_j, and a current generator. The equivalent circuit is valid only if the charge stored on the oxide surrounding the diode is negligible compared with that stored on the diode. The resistive sea immediately over the p-region is represented by the parallel combination of R_s and C_s. The time-constant for this combination is the intrinsic time-constant for the resistive sea (that is, $R_sC_s = \rho_s\epsilon_s\epsilon_o$ where ϵ_s is the relative dielectric constant and ρ_s is the volume resistivity of the resistive film).

A qualitative estimate of the charge storage properties of the

* These conclusions are consistent with the results obtained by Grove and Fitzgerald[19] on a gate-controlled diode structure. They show that for inversion to occur, the difference between the silicon surface potential at the oxide interface and the reverse bias voltage of the diode must be less than twice the fermi potential of the substrate. Because of the lateral charge spreading in a low resistivity sea, this inequality is never satisfied whereas with a high resistivity sea it can be satisfied in a region where there is a high generation rate of minority carriers.

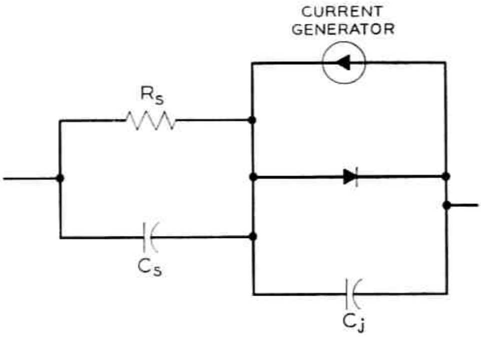


Fig. 23 — Equivalent circuit used to represent the area around one diode or p-type region.

equivalent circuit can be obtained from intuition. Consider first the fact that the electron beam will charge both the resistive sea and the beam side of capacitor C_s down to zero potential. If the reverse-bias leakage current of the diode may be neglected, the resistor R_s will discharge any voltage difference across C_s at a rate related to the time-constant R_sC_s. From the previous discussion it is estimated that ρ_s is approximately 10^8 ohm-cm. Therefore, R_sC_s may be estimated to be 35 μsec by assuming ϵ_s to be 4. Thus, without illumination, the p-regions are quickly charged to cathode potential and the full value of reverse-bias is placed across the diode.

The current from the photoresponse is represented by the current generator in Fig. 23. Since the capacitance from the film surface to electrical ground is very small, any change in reverse-bias voltage across the diode caused by photoresponse throughout the frame period appears very quickly on the electron beam side of the resistive sea. Furthermore, this process does not create a significant voltage drop across R_s. However, when the full value of reverse bias is re-established by the scanning electron beam, a significant voltage drop may appear across the parallel combination of R_s and C_s since the beam is on a diode for less than 0.3 μsec. For example, if the photoresponse has created a reduction in diode reverse-bias of ΔV_1 volts, the process of charging the beam side of the resistive sea down to zero volts will increase the reverse-bias by the amount ΔV_2 where

$$\Delta V_2 = \frac{C_s}{C_s + C_j} \Delta V_1.$$

The ratio of ΔV_2 to ΔV_1 may be estimated by assuming equal values for the relative dielectric constants for the resistive sea and the depletion region. Thus, if the thickness of the resistive sea is 1/6 of the depletion width, then

$$\Delta V_2/\Delta V_1 = 6/7.$$

The significance of this voltage ratio is that the scanning beam cannot re-establish the full value of reverse bias across the diode in one sweep even with arbitrarily large beam currents. While the charge stored on the oxide surrounding the diode has been neglected in this discussion, a similar conclusion would result from a calculation which included this additional charge.

One important question about the resistive sea structure that has not yet been answered is whether a film resistivity can be chosen which will lead to an increase in the effective beam landing area of each p-type region without significantly affecting the resolution capabilities of the basic diode array. Answering this question requires an evaluation of the amount of charge stored on the resistive film over the oxide surrounding the diode relative to the amount of charge stored on the diode. This evaluation in turn requires a complicated model which includes the effects of the isolated p-regions and is beyond the scope of this paper. However, preliminary calculations indicate that there is a value of R_f which will preserve the resolution capabilities of the diode array and will also lead to a significant increase in the beam landing area of each p-type island. The optimum value of R_f is a strong function of the target geometry but will always be in the range of 10^{12} to 10^{14} ohms per square for practical geometries.

VI. MISCELLANEOUS TOPICS

The dark current characteristics of a diode array target are predominantly determined by the surface states at the silicon-silicon dioxide interface, as discussed by Buck and others.[13] However the detailed behavior of the dark current versus target voltage depends upon many other factors some of which are discussed in this section.

6.1 *Effect of Resistivity Striations on Dark Current*

A large number of the silicon diode array camera tubes fabricated to date have exhibited a phenomenon called "coring." Coring manifests itself as a modulation of the dark current pattern as illustrated in Fig. 24. The photographs in the figure are of the video display of the dark current pattern of a diode array camera tube at different target

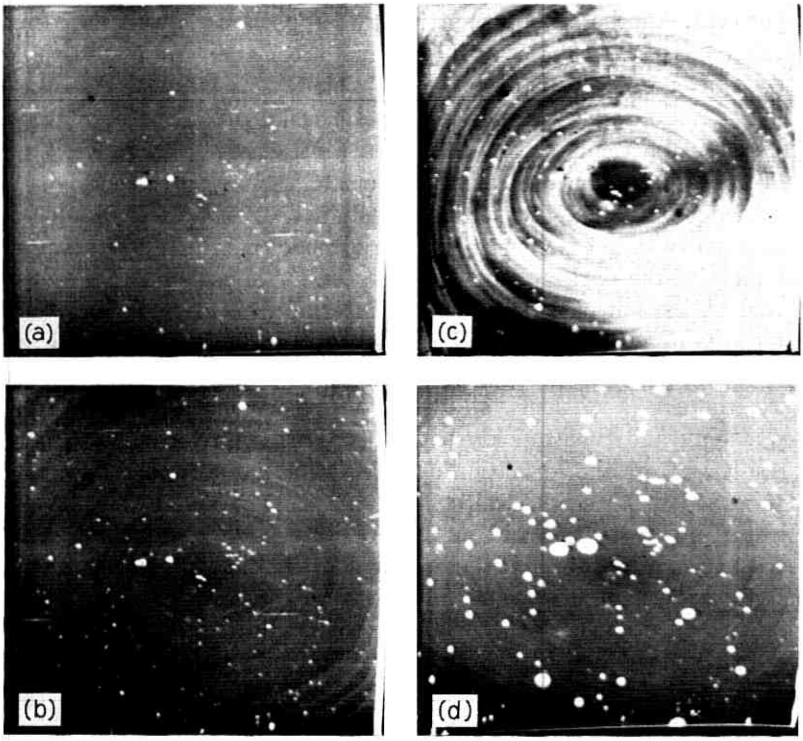


Fig. 24 — Photographs of the video display of the dark current pattern of a camera tube which exhibits "coring": (a) $V_T = 4$ volts, (b) $V_T = 6$ volts, (c) $V_T = 8$ volts, (d) $V_T = 12$ volts.

voltages. The modulation introduced by the coring pattern is seen to be a strong function of the target voltage; its maximum amplitude can be as high as 40 per cent. The spatial wavelength of the coring pattern is typically of order 500μ. The term modulation as used here means the ratio of the peak-to-peak modulation of the dark current, introduced by the coring, to the average dark current.

One possible cause of the coring, consistent with experimental results, is resistivity striations produced in the silicon substrate during crystal growth. The standard methods used for growing silicon crystals would result in circular striations.[20] In addition, a silicon crystal in which resistivity striations had been purposely introduced yielded targets which exhibited coring patterns that corresponded to the resistivity striations. A variation in resistivity of approximately 25 per cent yielded coring patterns with a modulation of approximately

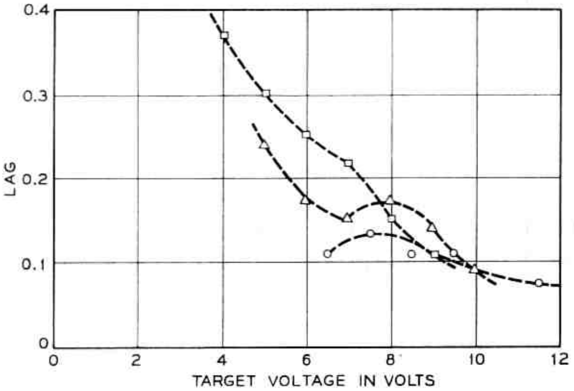


Fig. 29 — Image lag as a function of target voltage for different video signal levels: □ = 50 nanoamps, △ = 200 nanoamps, ○ = 500 nanoamps.

VII. CONCLUSION

The preliminary results reported in this paper indicate the silicon diode array camera tube has improved lag and spectral response and comparable resolution capabilities when compared with commercially available vidicons. In addition, the unity gamma of a diode array camera would be a significant advantage for color television cameras.

Two of the outstanding features of the silicon diode array camera are its wide spctral response (0.4 to 1.0μ) and its high effective quantum yield (approximately 50 percent). For fluorescent illumination these provide a sensitivity of approximately 1.3 μamp per ft-cd of faceplate illumination with an image sensing area of 1.8 sq-cm.

The expected operating life of a silicon diode array camera should exceed that a vidicon for at least two reasons. First, the image sensing target is not damaged by intense light images (for example, the noonday sun has been imaged with a F:1.5 lens on the silicon target without damage) Second, the completely assembled tube can be vacuum baked at 400°C provided an appropriate resistive film is used. This vacuum bake should provide a longer cathode life.

Typical video performance of a diode array tube is illustrated by Figs. 30, 31, and 32. These photographs were obtained from a 525-line monitor when the image of a black and white transparency was focused onto the camera. The photograph in Fig. 32 was obtained by reducing the size of the raster on the diode array so that only a small portion of the array was scanned. This electronic zooming permits the individual diodes to be observed for detailed study. For example, the one very bright spot or white defect on the number 300 results from a single defective diode. The two bright spots on the extreme left represent two defective diodes separated by a good diode. For this photograph the optical magnification was adjusted so that the black and white wedge pattern created 300 cycles per inch at the center of the display. Since the diodes are located on 20μ centers, only two diodes are fully illuminated by a white bar near the numeral 300.

While the bright defects depicted in the photographs of Figs. 30 to 32 impair the image quality and would in some instances prevent this tube from being used, the small size and number of defects would be

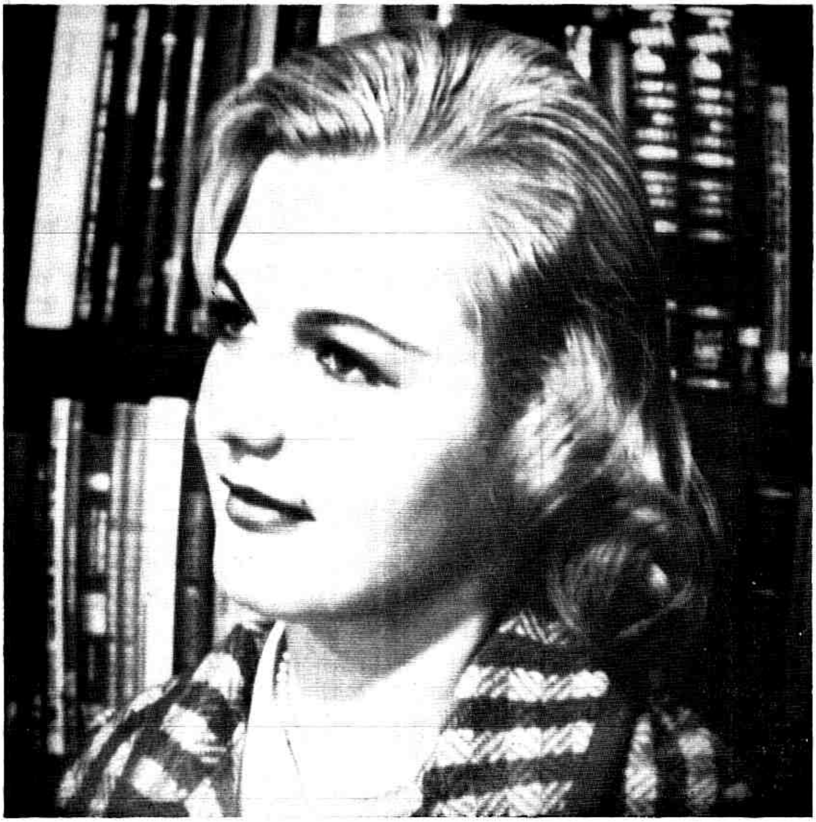


Fig. 30 — A video display obtained with a typical silicon diode array camera tube. The subject was a black and white transparency.

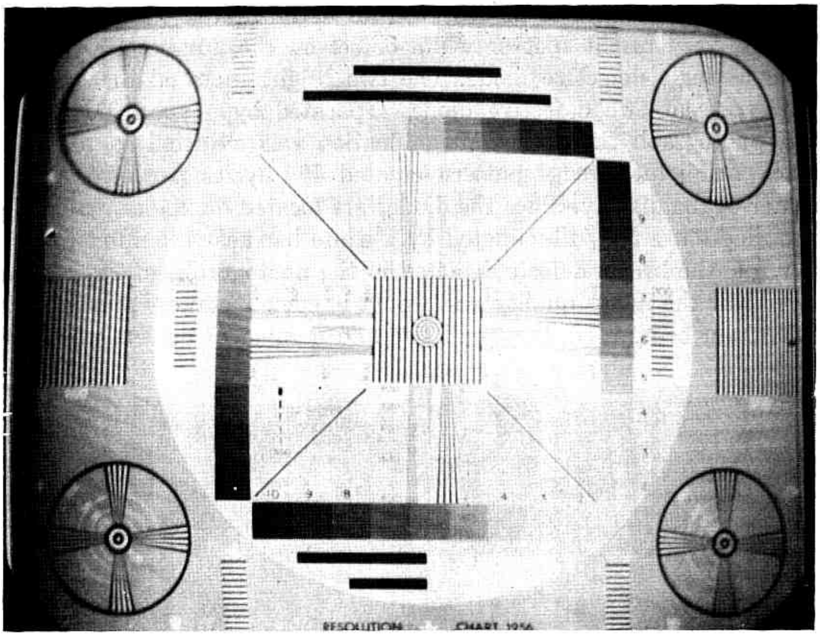


Fig. 31 — Video display of a resolution chart obtained with a silicon diode array camera tube.

acceptable in a number of applications. Although most of the arrays fabricated to date have exhibited bright defects, considerable progress has been made in reducing their number; improved technology should permit fabrication of defect-free arrays with moderately good yield.

VIII. ACKNOWLEDGMENTS

Although it is impossible to recognize individually all those who contributed to the work reported in this paper, we gratefully acknowledge the effective cooperation provided by everyone associated with the diode array camera tube program. The contributions of the people mentioned below were essential to the success of this program. Expert technical assistance was provided by P. M. Ryan and E. J. Zimany, Jr. The processing of the diode array targets was directed by T. M. Buck and J. V. Dalton; the mechanical design of the tubes was handled by E. J. Walsh, R. P. Hynes, and E. G. Olsen. Computational assistance was provided by A. J. Chick. Finally, the authors thank E. I. Gordon for the valuable stimulation he provided.

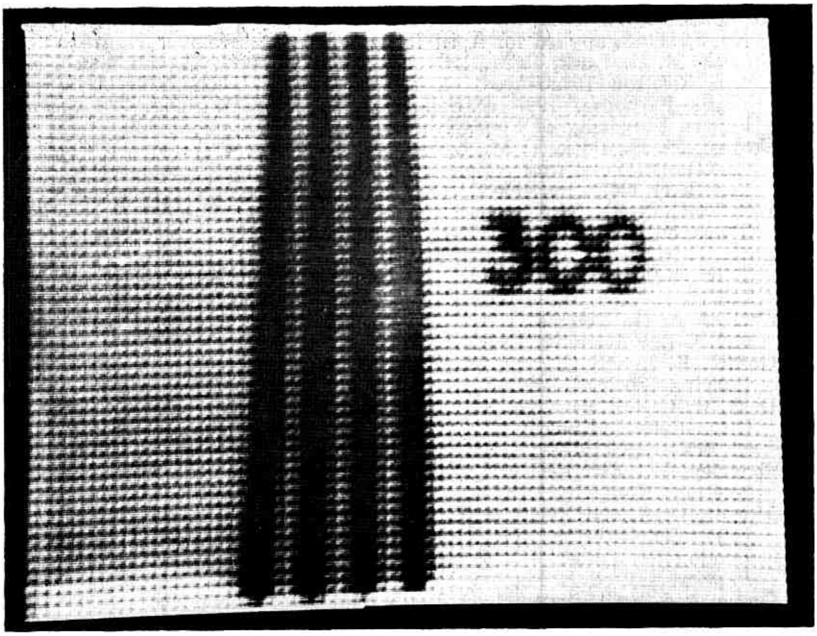


Fig. 32 — The video display of a small portion of the resolution chart shown in Fig. 30 obtained by electronically zooming the diode array camera. The white spots or defects correspond to diodes with a high value of reverse bias leakage current.

REFERENCES

1. Zworykin, V. K. and Morton, J. A., *Television,* New York: John Wiley and Sons, 2nd ed., 1954.
2. Weimer, P. K., Forgue, J. V., and Goodrich, R. R., "The Vidicon Photoconductive Camera Tube," RCA Rev., *12,* No. 1 (September 1951), pp. 306–313.
3. de Haan, E. F., van der Drift, A., and Schampers, P. P. M., "The 'Plumbicon,' a New Television Camera Tube," Philips Technical Rev., *25,* No. 6 and 7, (1963 and 64), pp. 133–155.
4. Horton, J. W., Mazza, R. V., and Dym, H., "The Scanistor—A Solid State Image Scanner," Proc. IEEE, *52,* No. 12 (December 1964), pp. 1513–1528.
5. Weimer, P. K., Sadasiy, G., Borkan, H., Meray-Horvath, L., Meyer, J., Jr., and Schallcross, F. V., "A Thin Film Solid-State Image Sensor," 1966 Int. Solid State Circuits Conf., University of Pennsylvania, Digest of Technical Papers, pp. 122–123.
6. Weckler, G. P., "Storage Mode Operation of Phototransistor and Its Adaption to Integrated Arrays for Image Detection," 1966 Int. Electron Device Meeting, Washington, D. C., October 26-28, 1966, p. 34.
7. Schuster M. A. and List, W. F., "Fabrication Considerations for Monolithic Electrooptical Mosaics," Trans. of the Metallurgical Soc. of AIME, *236,* No. 3 (March 1966), p. 375–378.
8. Papers in "Special Issue on Solid State Imaging," IEEE Trans. on Elec. Devices, *ED-15,* No. 4 (April 1968).

9. Reynolds, F. W., "Solid State Light Sensitive Storage Device," U. S. Patent No. 3,011,089, applied for April 15, 1958, issued November 21, 1961.
10. Crowell, M. H., Buck, T. M., Labuda, E. F., Dalton, J. V., and Walsh, E. J., "An Electron Beam-Accessed, Image-Sensing Silicon-Diode Array with Visible Response," 1967 Int. Solid State Circuits Conf., Digest of Technical Papers, University of Pennsylvania, March 1967, pp. 128–130.
11. Crowell, M. H., Buck, T. M., Labuda, E. F., Dalton, J. V., and Walsh, E. J., "A Camera Tube with a Silicon Diode Array Target," B.S.T.J., *46,* No. 2 (February 1967), pp. 491–495.
12. Wendland, P. H., "A Charge-Storage Diode Vidicon Camera Tube," IEEE Trans. on Elec. Devices, *ED-14,* No. 9 (June 1967), p. 285–291.
13. Buck, T. M., Casey, H. C., Jr., Dalton, J. V., and Yamin, M., "Influence of Bulk and Surface Properties on Image Sensing Silicon Diode Arrays," B.S.T.J., *47,* No. 9 (November 1968), pp. 1827–1854.
14. Chester, A. N., Loomis, T. C., Weiss, M. M., "Diode Array Camera Tubes and X-Ray Imaging," B.S.T.J., *48,* No. 2 (February 1969), pp. 345–381.
15. Gordon E. I. and Crowell, M. H., "A Charge Storage Target for Electron Imaging Sensing," B.S.T.J., *47,* No. 9 (November 1968), pp. 1855–1873.
16. Dash, W. C. and Newman, R., "Intrinsic Optical Absorption in Single-Crystal Germanium and Silicon at 77°K and 300°K," Phys. Rev., *99,* No. 4 (August 15, 1955), pp. 1151–1155.
17. Shockley, W., *Electrons and Holes in Semiconductors,* New York: D. van Nostrand Company, Inc., 1950.
18. Morton, G. A. and Ruedy, J. E., "The Low Light Level Performance of the Intensifier Orthicon," in *Photo-Electronic Image Devices,* symposium at London September 3–5, 1958, in *Advances in Electronics and Electron Physics,* XII, ed. L. Marton, New York: Academic Press, 1960, pp. 183–193.
19. Grove, A. S. and Fitzgerald, D. J., "Surface Effects on p-n Junctions: Characteristics of Surface Space-Charge Regions Under Non-Equilibrium Conditions," Solid State Elec., *9,* No. 8 (August 1966), p. 783–806.
20. Dikhoff, J. A. M., "Inhomogeneities in Doped Germanium and Silicon Crystals," Philips Technical Rev., *25,* No. 8 (1963 and 64), pp. 195–206.
21. Redington, R. W., "The Transient Response of Photoconductive Camera Tubes Employing Low Velocity Scanning," IRE Trans. on Elec. Devices, *ED-4,* No. 3 (July 1957), pp. 220–225.

Television Transmission of Holograms With Reduced Resolution Requirements on the Camera Tube

By C. B. BURCKHARDT and L. H. ENLOE

(Manuscript received November 21, 1968)

This paper proposes a technique for the television transmission of a hologram of a two-dimensional transparency. The spatial resolution required on the camera tube is reduced by a factor of four compared with the transmission of a conventional off-axis reference beam hologram. The resolution required is therefore no higher than that required for the direct transmission of the transparency itself. Implementation of the proposed arrangement should be easy. Three holograms formed with an on-axis reference beam are transmitted. The phase of the reference beam assumes the values 0°, 120°, *and* 240° *for the first, second, and third hologram, respectively. The carrier-frequency hologram is "synthesized" from these three on-axis holograms at the receiver. The technique has the further advantage that the undesirable zero-order terms are eliminated.*

Holograms of two-dimensional transparencies have been transmitted via television.[1] The hologram is first formed on the face of the camera tube with an off-axis reference beam and is then transmitted. The main difficulty with this scheme is the high spatial resolution requirement for the camera tube. If the object wavefront has spatial frequencies between $-W$ and $+W$, then the spatial frequencies of the unwanted zero-order terms extend from $-2W$ to $2W$. The spatial frequency of the reference beam therefore has to be at least $3W$; the highest spatial frequency to be resolved by the camera tube is $4W$. (The conditions mentioned, and further discussed in Ref. 2, are well known.) This is higher by a factor of 4 than the highest spatial frequency of the original two-dimensional transparency which is unfortunate because television camera tubes are of rather limited resolution.

Two scanning schemes have recently been proposed which reduce the resolution requirement for the camera tube by a factor of 2 and 4.[3] It is the purpose of this paper to point out that a reduction by a factor of 4 can also be achieved by the adaption of a scheme described by Burckhardt and Doherty.[4] The idea to be described should be easier to implement than the heterodyne scanners proposed in Ref. 3 and has the advantage that it allows the use of charge storage camera tubes. Since a factor of 4 is saved in resolution requirement, the resolution of the camera tube has to be no higher than that required for the direct transmission of the original transparency.

Figure 1 shows the adaption of the idea of Reference 4 to hologram transmission via television. The hologram is formed with an on-axis reference beam on the camera tube. This hologram is scanned and transmitted; at the receiver, the received electrical signal is multiplied by a cosinusoidal signal and displayed on a kinescope. The phase plate at the transmitter is then switched electro-optically to give a phase shift of 120° in the reference beam; correspondingly the cosinusoidal signal at the receiver is shifted by 120° in temporal phase. The hologram is again scanned, transmitted, multiplied by the cosinusoidal signal, and displayed. This procedure is repeated once more. It will now be shown that the intensities of the three scans add up to give a carrier frequency hologram on the kinescope.

Let the complex-valued amplitude of the subject wavefront be called A and the real-valued amplitude of the reference beam be called B. For the intensity I_1 on the camera tube during the first scan we then have

$$I_1 = (A + B)(A^* + B) = AA^* + B^2 + AB + A^*B. \qquad (1)$$

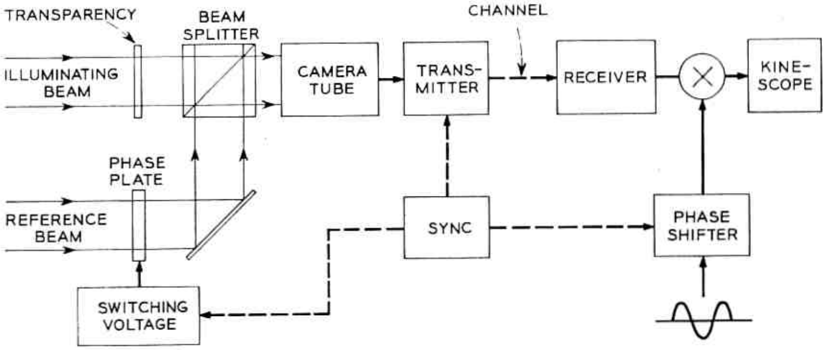


Fig. 1 — Hologram transmission via television, with reduced resolution requirement on the camera tube.

Suppose now that the transmitter is linear and that the voltage U_1 arriving at the receiver is proportional to I_1,

$$U_1 = K_1 I_1 , \tag{2}$$

where K_1 is the constant of proportionality. The voltage U_1 is now multiplied by a cosinusoidal signal to give the voltage U_1' ,

$$\begin{aligned} U_1' &= U_1 \cos \omega t = K_1 I_1 \cos \omega t \\ &= K_1 \cos \omega t (AA^* + B^2 + AB + A^*B). \end{aligned} \tag{3}$$

We now assume that the display is also linear and that the intensity I_{K1} on the kinescope is given by

$$I_{K1} = K_o + K_2 U_1' . \tag{4}$$

The constant bias term K_o is necessary because U_1' assumes both positive and negative values. The term K_2 is a constant of proportionality. Combining equations (3) and (4) we obtain for the intensity I_{K1} on the kinescope

$$I_{K1} = K_0 + K_1 K_2 \cos \omega_s x (AA^* + B^2 + AB + A^*B). \tag{5}$$

The term ω_s is the spatial frequency which corresponds to the temporal frequency ω in equation (3).

During the second scan the total amplitude on the camera tube is $A + B \exp (j2\pi/3)$ because the phase of the reference beam is now shifted by 120°. The intensity I_2 therefore is

$$\begin{aligned} I_2 &= [A + B \exp (j2\pi/3)] \cdot [A^* + B \exp (-j2\pi/3)] \\ &= AA^* + B^2 + AB \exp (-j2\pi/3) + A^*B \exp (j2\pi/3). \end{aligned} \tag{6}$$

We now multiply the voltage arriving at the receiver by $\cos (\omega t + 2\pi/3)$ and obtain for the intensity I_{K2} on the kinescope

$$\begin{aligned} I_{K2} &= K_0 + K_1 K_2 I_2 \cos (\omega_s x + 2\pi/3) \\ &= K_0 + \tfrac{1}{2} K_1 K_2 I_2 [\exp (j\omega_s x + j2\pi/3) + \exp (-j\omega_s x - j2\pi/3)] \\ &= K_0 + \tfrac{1}{2} K_1 K_2 [\exp (j\omega_s x + j2\pi/3) + \exp (-j\omega_s x - j2\pi/3)] \\ &\quad \cdot [AA^* + B^2 + AB \exp (-j2\pi/3) + A^*B \exp (j2\pi/3)]. \end{aligned} \tag{7}$$

The intensity I_{K3} on the kinescope during the third scan is obtained in an analogous way. For the total intensity $I_{K\text{tot}}$ we then obtain

$$\begin{aligned} I_{K\text{tot}} &= I_{K1} + I_{K2} + I_{K3} \\ &= 3K_0 + \tfrac{1}{2}K_1K_2 \sum_{n=0}^{2} \{[\exp(j\omega_s x + jn2\pi/3) \\ &\qquad + \exp(-j\omega_s x - jn2\pi/3)] \\ &\qquad \cdot [AA^* + B^2 + AB \exp(-jn2\pi/3) + A^*B \exp(jn2\pi/3)]\} \\ &= 3K_0 + (\tfrac{3}{2})ABK_1K_2 \exp(j\omega_s x) + (\tfrac{3}{2})A^*BK_1K_2 \exp(-j\omega_s x). \end{aligned} \tag{8}$$

The last two terms of this expression are the real and virtual image terms modulated onto different spatial carriers. Notice that the undesirable zero-order terms do not occur in equation (8). This is because we multiplied the voltage arriving at the receiver with a bipolar electrical signal. If the subject wavefront at the camera tube has spatial frequencies extending from $-W$ to W, the spatial carrier frequency at the kinescope can be chosen as W. The positive spatial frequencies of the kinescope display then extend from 0 to $2W$. (Since the intensity on the kinescope is a real function, a knowledge of the positive frequencies is sufficient.)

Some bandwidth considerations are appropriate. If the positive spatial frequencies of the amplitude transmitted through the original transparency extend from 0 to W, the hologram displayed at the receiver has a bandwidth of $2W$. This increase by a factor of 2 occurs because the hologram contains information about amplitude and phase. The system just described transmits three holograms, each with a bandwidth W. This is equivalent to transmitting one hologram with a bandwidth $3W$. The minimum bandwidth of an off-axis hologram is $4W$; therefore, our scheme requires less bandwidth than transmitting an off-axis hologram. Since the bandwidth of the hologram on the kinescope is $2W$, the amount of information to be transmitted in our scheme is still higher by a factor 3/2 than what it necessarily has to be. The scheme described in Section IV of Ref. 3 only transmits a hologram of bandwidth $2W$ therefore avoiding this increase.

In our discussion we have used three subholograms and phase shifts of 120°. In the Appendix we derive the general equations and show that three subholograms is the minimum required number.

It might be mentioned that our scheme can be modified such that it only transmits one hologram of bandwidth $2W$. In this case all the processing is done at the transmitter and the final hologram of bandwidth $2W$ is transmitted. A scheme for doing this is shown in Fig. 2.

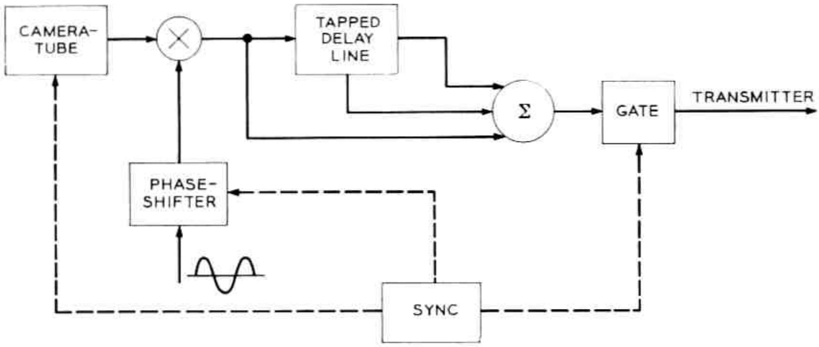


Fig. 2 — Modification of the arrangement of Fig. 1 to reduce the amount of information to be transmitted by one third.

The voltage from the scan is multiplied by the sinusoidal signal and then stored in a tapped delay line. The delay line delays the first scan by the time needed for two scans and the second scan by one scanning time. The gate opens during the third scan. The channel is only used during one third of the time. In order that anything be gained the channel has, of course, to be used for something else during the remaining two thirds of the time. Alternatively, the output of the gate can be stored in a buffer memory (for example, magnetic tape) and transmitted at a slower rate. It is seen that the scheme of Fig. 2 is quite a bit more complex than the scheme of Fig. 1.

APPENDIX

General equations

Here we present the general equations which must be satisfied for the N subhologram case and show that the least number required is $N = 3$.

The general expression for the spatially varying part of the intensity on the kinescope corresponding to equation (8) is

$$\begin{aligned} I_{\text{tot}} &= \operatorname{Re} \sum_{n=0}^{N-1} [\{1 + AA^* + A \exp(-j\beta_n) \\ &\qquad + A^* \exp(j\beta_n)\} \exp(j\omega_s x + j\gamma_n)] \\ &= \operatorname{Re} \sum_{n=0}^{N-1} [\{[1 + AA^*] \exp(j\gamma_n) + A \exp(j\gamma_n - j\beta_n) \\ &\qquad + A^* \exp(j\gamma_n + j\beta_n)\} \exp(j\omega_s x)] \end{aligned} \tag{9}$$

where β_n is the relative phase-shift of the plane-wave reference beam; ω_s and γ_n are the spatial frequency and phase of the grating produced by the electrical carrier introduced at the receiver. In order to simplify expressions, we have equated the multiplying factors K_1, K_2 and the magnitude of the reference beam to unity. Notice that the quantity within the { } braces in the second equation of (9) represents the beam which would be diffracted at the angle ω_s when a hologram is reconstructed. This term is the coefficient of $\exp(j\omega_s x)$. We desire to superimpose N exposures, each having the form of Eq. (9), to accomplish the following:

(*i*) Force the complex coefficient of $1 + AA^*$ to zero. This prevents components from the direct beam, during reconstruction, from being diffracted at angle ω_s.

(*ii*) Force the complex coefficient of A^* to zero. This prevents components of the conjugate wave from being diffracted at ω_s.

(*iii*) Force the complex coefficient of A to some nonzero value. This reconstructs the desired object wavefront at angle ω_s.

In order to control these 3 complex coefficients, we need a minimum of 6 independent variables to adjust.† Each exposure of the form equation (9) has 2 variables to adjust, β_n and γ_n. Thus, we need a minimum of 3 subholograms.

The equations which must be satisfied are

$$\sum_{n=0}^{N-1} \exp(j\gamma_n) = 0 \tag{10a}$$

$$\sum_{n=0}^{N-1} \exp(j\gamma_n - j\beta_n) \neq 0 \tag{10b}$$

$$\sum_{n=0}^{N-1} \exp(j\gamma_n + j\beta_n) = 0, \tag{10c}$$

where $N = 3$ is the minimum value. Equation (10a) can be satisfied if for the γ_n's we simply pick the N roots of $\exp(j\theta)$ according to the well-known theorem of De Moivre, that is,

$$\gamma_n = \frac{\theta + 2\pi n}{N}$$

† There is always a possibility that the equations for these three complex coefficients are not themselves independent, and that as a consequence only four independent variables are required to control them. In order to rule out this case, we let $N = 2$ in equations (10) and define $s_n \equiv \exp(j\gamma_n)$ and $z_n \equiv \exp(j\beta_n)$. Equations (10) then reduce to $s_0 = -s_1$, $s_0^*(z_0 - z_1) \neq 0$ and $s_0(z_0 - z_1) = 0$. We see that these equations cannot be satisfied simultaneously.

where $n = 0, 1, 2, \ldots N - 1$. If we then set $\beta_n = \gamma_n$, equations (10b) and (10c) are then automatically satisfied.

As an example, for the minimum number of subholograms $N = 3$, we may pick $\theta = 0$ without loss of generality since the absolute phase of the reference beam is unimportant. Then we have from De Moivre's theorem $\gamma_0 = \beta_0 = 0$, $\gamma_1 = \beta_1 = 2\pi/3$, $\dot{\gamma}_2 = \beta_2 = 4\pi/3$. Thus, for three subholograms we shift the reference beam and grating producing electrical carrier phase by 120°

REFERENCES

1. Enloe, L. H., Murphy, J. A. and Rubinstein, C. B., "Hologram Transmission via Television," B.S.T.J., *45*, No. 2 (February 1966), pp. 333–335.
2. Leith, E. N. and Upatnieks, J., "Reconstructed Wavefronts and Communication Theory," J. Opt. Soc. Amer., *52*, No. 10 (October 1962), pp. 1123–1130.
3. Enloe, L. H., Jakes, W. C., Jr., and Rubinstein, C. B., "Hologram Heterodyne Scanners," B.S.T.J., *47*, No. 9 (November 1968), pp. 1875–1882.
4. Burckhardt, C. B. and Doherty, E. T., "Formation of Carrier-Frequency Holograms with an On-Axis Reference Beam," Appl. Opt., *7*, No. 6 (June 1968), pp. 1191–1192.

A Sliding-Scale Direct-Feedback PCM Coder for Television

By EARL F. BROWN

(Manuscript received December 19, 1968)

A sliding-scale coder for television signals was built which extends the range of the quantizing scale by processing the input signal twice when the input signal exceeds a prescribed threshold. On the second pass the quantizing range is effectively moved outward to reduce the errors in coding large signals. Double processing nearly triples the number of quantizing levels of a basic three-bit coder. Measurements of the number of extra bits required, that is, those in excess of three-bits per sample show that they may be accomodated on a three-bit per sample transmission channel by reducing the sampling rate five percent. The experimental coder generates 19 *quantizing levels. Its performance approaches that of a seven-bit pulse code modulation coder. Busyness or streaking, common to most three-bit differential type coders, is eliminated. Acceptable pictures are reproduced with* ±5 *dB changes in the input signal's range. Over this range the signal-to-noise ratio of the reproduced pictures varies from* 47 *dB to* 54 *dB and the rise-time of a regenerated step-signal varies from* 1 *microsecond to* 1.45 *microsecond when the input signals rise-time is limited to* 1 *microsecond.*

I. INTRODUCTION

Differential, direct-feedback, and delta-modulation pulse code modulation systems take advantage of the television viewer's tolerance to brightness errors, especially in high detail areas of the picture.[1–5*] Analog signals must be quantized into a finite number of levels for conversion to digital signals. This quantization introduces errors in the reconstructed picture. These errors are lumped together under the name of quantizing noise which for differential pulse code modulation (PCM) systems is a function of the quantizer step size(s), the sampling rate, channel capacity, and filter characteristics. Quantizing noise may be classified into six visually subjective catagories: granular

[*] This family of coders are hereafter referred to as differential coders.

noise, streaking, contouring, slope-overload, edge-busyness, and edge-stepping.

Granular noise is a high frequency noise, caused by individual sample errors, whose visibility is increased by amplitude differences from frame to frame. Contouring produces brightness steps in flat regions of the picture. Both of these defects may be decreased with proper filtering and decreasing the smaller step sizes. A reduction in contouring is usually made at the expense of increased granular noise.

Streaking results from mistracking between the coder and the decoder. The length is determined by the decoder's time constant.

Slope-overload, edge-busyness, and edge-stepping occur at large brightness boundaries which are not parallel to the scanning lines. These defects become increasingly visible as the brightness boundaries approach the vertical. Slope-overload appears as a smearing effect. This may be reduced by increasing the step size for large difference signals at the expense of increasing edge-busyness and edge-stepping. Edge-busyness appears as relatively large brightness errors jumping back and forth along the scanning line. This defect results from large errors at brightness boundaries whose jitter is increased with frame-to-frame amplitude differences and when the sweep rates are not locked to the digital processing rates. Edge-stepping appears as discontinuities in brightness boundaries because of amplitude differences along the continuum. This defect appears to crawl up and down the boundary when the sweep rates are not locked to the digital processing rates.

Some or all of these defects may be reduced, if not eliminated, through one or more of the following procedures:

(*i*) Companding the signal,
(*ii*) Increasing the number of levels and length of PCM words, and
(*iii*) Increasing the sampling and bit rate.

When the bandwidth and bit rate are fixed, more sophisticated techniques are required such as:

(*i*) Optimizing sampling rate or coder processing rate as a function of spatial frequency,

(*ii*) Adding levels as a function of slope amplitude.

(*iii*) Efficiently using time slots such as redundant signal areas and blanking periods.

Two types of sliding-scale differential coders were simulated on a computer.[6] The excellent results obtained in the simulation encouraged the building of a real-time sliding-scale coder.

The real-time sliding-scale coder was designed to process the input signal twice and introduce additional levels when the input signal exceeded a threshold. Double processing can increase the number of effective levels in a three-bit system to 22 at a moderate increase in circuit complexity. The additional information may be handled by reducing the sampling rate; or the additional information could be transmitted during the blanking period or in place of redundant signal components.

The experimental coder was limited to 19 levels. Its performance approached that of a 7-bit straight PCM coder. Slope-overload, edge-stepping, and granular noise were minimized. Edge-stepping was just perceptible when the sweep rates and the digital processing rates were unlocked. Edge-busyness and contouring were eliminated. Input signals varying over a ±5 dB range produced acceptable pictures. At midrange the peak-to-peak signal to root mean square noise was 50 dB; and with a 75 percent change in signal level, the rise-time increased from 1 to 1.15 μs. Over the input signal operating range of ±5 dB, a signal-to-noise ratio of 54 to 47 dB was obtained. Over the same operating range, and with a 75 percent change in signal level occurring in 1 μs, the rise-time of the output signal varied from 1 to 1.45 μs.

II. DIRECT-FEEDBACK CODING

The sliding-scale coder was built around a direct-feedback coder configuration. Briefly, direct-feedback coders function the same as DPCM coders, but the circuit is arranged to allow greater flexibility of filter design. Figure 1 is a schematic diagram of a direct-feedback

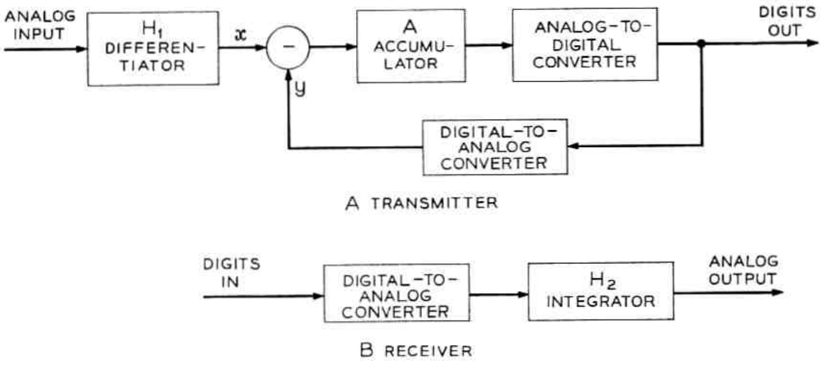


Fig. 1 — Block diagram of a direct-feedback PCM coder,

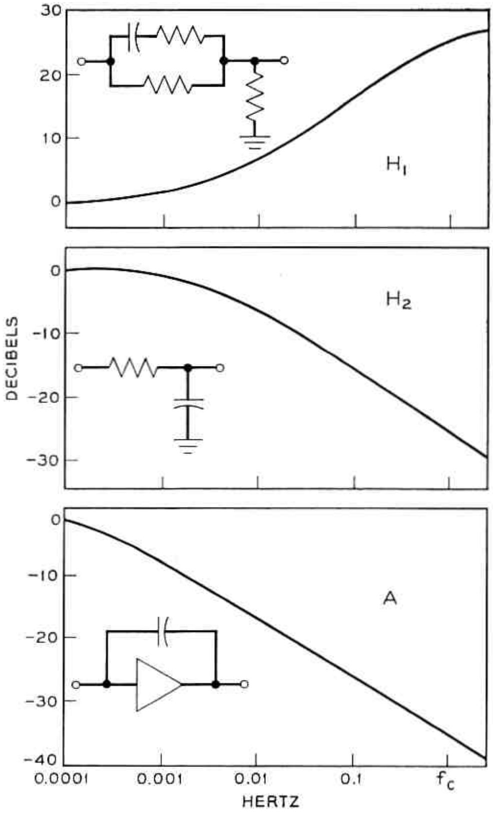


Fig. 2 — Filter characteristics of a direct-feedback PCM coder.

coder, and Fig. 2 shows typical filter characteristics. For television signals the preemphasis filter H_1 is a differentiating filter. The deemphasis filter H_2 is a short time integrator, approximately the inverse of H_1. The accumulator filter A in the feedback loop has a long time constant.

The feedback acts like a servomechanism trying to make the average value of the quantized signal, y, equal to the pre-emphasized input signal, x. The difference between x and y is accumulated in A and used to correct the quantized output.

Figure 3 shows a typical 8-level companded quantizer scale. The quantizer is tailored to the observer's perception; that is, fine quantum steps are used for small signal errors and coarse steps for large error signals. Optimally designed companded quantizers adhere to Max's rule for minimum distortion.[7] Max states that the decision levels must fall midway between the quantizer levels. In such a quantizer the error amplitude ranges between plus and minus half a quantum step over the range of the quantizer.

Even so, minimum distortion quantizers of three bits per sample or less are subject to considerable noise and are only marginally acceptable. The sliding-scale coder is an attempt to increase the subjective acceptability of predictive coders.

III. PRINCIPLES OF THE CODER

3.1 *Transmitter Coder*:

Figure 4a is a block diagram of a sliding-scale direct-feedback coder. It has the same functional blocks as a direct-feedback coder except for an AND gate and an elastic store. Assume a three-bit coder with the quantizer levels shown in Fig. 5. Switches S_1 and S_2 of Fig. 4a are closed at time t_1. The error signal out of the accumulator at time t_1 is quantized and fed back to the accumulator. If the quantizer output stays within the bounds of decision levels $+c$ and $-c$, the processing during that sample period is complete and one word describing

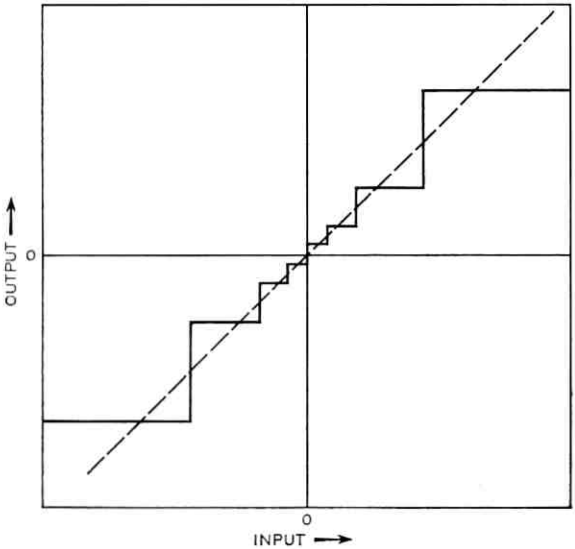


Fig. 3 — Typical eight-level companded quantizer scale.

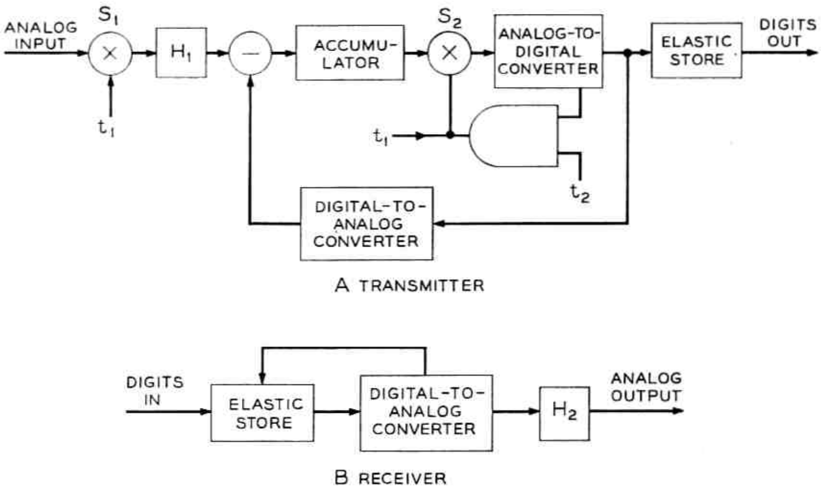


Fig. 4 — A sliding-scale direct-feedback PCM coder.

the error signal is generated. If decision levels $+d$ or $-d$ are generated they are also fed back to the accumulator; they reduce its error signal. In addition, levels $+d$ and $-d$ are used to open the AND gate. A pulse then passes through the AND gate at time t_2 closing switch S_2 during the same sample period. In so doing, the output of the accumulator, already coarsely corrected by level $+d$ or $-d$, has a fine correction applied to it during the sample period.

When levels $+d$ or $-d$ are generated, two words are produced in one sample period. To facilitate a uniform transmission rate the words are fed into an elastic store which feeds the transmission channel at a constant rate. The sampling rate may be reduced by an amount proportionate to the number of additional words so as not to exceed the channel bit rate capacity. When the sampling rate is reduced, the cutoff frequency of the low-pass filter must be reduced proportionately so as to reduce the effects of foldover (aliasing) and granular noise.

3.2 *Receiver Decoder*

Figure 4b is a functional block diagram of the receiver decoder. The output of the receiver elastic store is applied to the digital-to-analog converter. When a $+d$ or $-d$ level is detected by the digital-to-analog converter, a second word is taken out of the elastic store during that sampling period. The output of the decoder after integration by filter H_2 is a replica of the input analog signal at the transmitter.

3.3 *Quantizer Levels*

The technique of double processing the error signal may be thought of as one operation in which the quantizing scale's midpoint may occupy one of three positions: centered around zero, $+d$, or $-d$. Thus the midpoint of the quantizer scale slides up and down the scale as a function of the amplitude of the accumulator's error signal. For an 8-level quantizer six levels are available when the midpoint of the scale is centered around zero and 8 each when centered around level $+d$ or $-d$ as illustrated in Fig. 5b. For a three-bit coder operating in this mode, 22 levels are available during one sample interval if level $+d,-d$ is counted twice. Although the quantizing scale of Fig. 5b is not optimized, it is adequate for most television applications. The effectiveness of all 22 available levels may be increased by additional companding of the error signal, approximately the inverse of the initial companding, on the second pass.

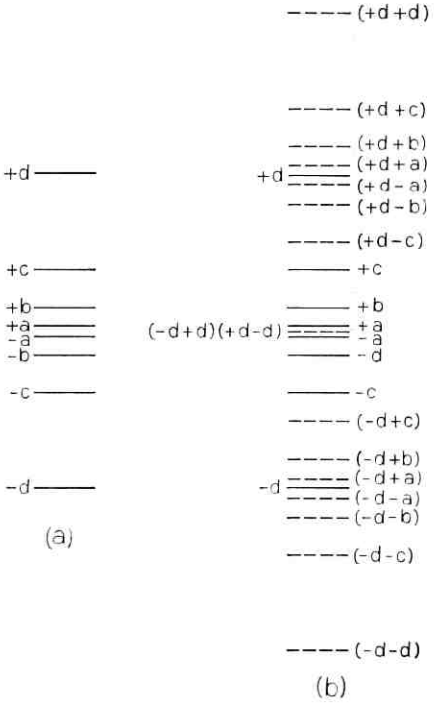


Fig. 5 — Quantizing scales: (a) an eight-level companded quantizing scale; (b) typical levels of a sliding-scale coder derived from the eight-level quantizing scale.

A general expression for the number of effective quantizer levels in a sliding-scale coder is

$$Q_L = (2^n - K) + (K2^m) \tag{1}$$

where 2^n is the number of levels the quantizer can generate, K is the number of levels which causes a double processing, and 2^m is the number of levels used when the midpoint of the quantizer scale is shifted from zero and where m usually equals n.

IV. EXPERIMENTAL SLIDING-SCALE CODER

Figure 6 is a block diagram of the experimental coder. This is a direct-feedback coder with the sliding-scale features added to it. This arrangement of the sliding-scale coder was used to increase its experimental versatility. The elastic stores were omitted since they do not directly relate to the quality of the picture if they have suf-

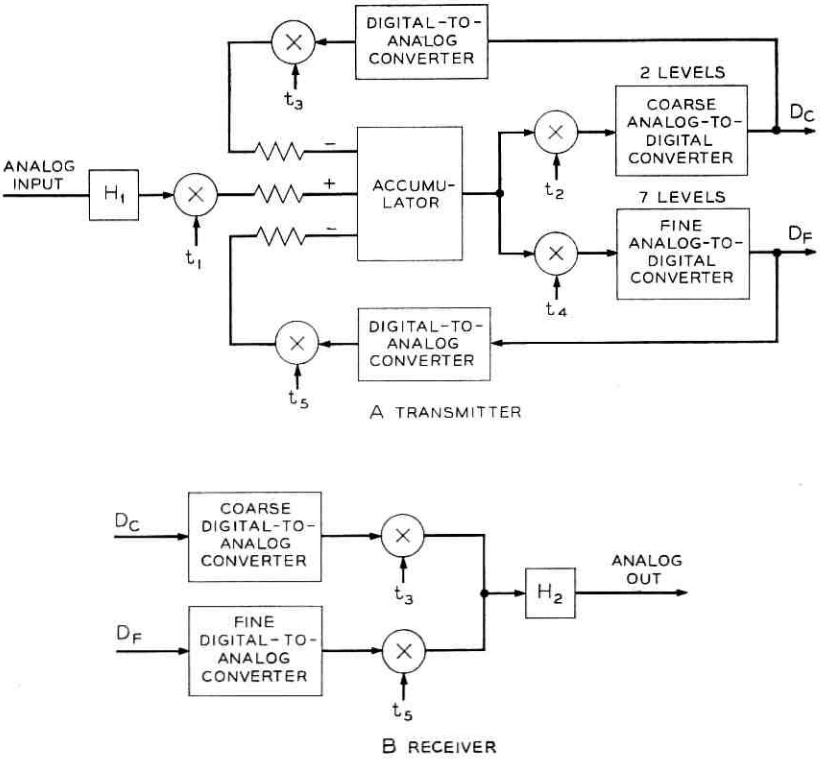

DIGITAL-TO-ANALOG CONVERTER
t₃
2 LEVELS
COARSE ANALOG-TO-DIGITAL CONVERTER
Dc
ANALOG INPUT
H₁
t₁
+
−
−
ACCUMULATOR
t₂
7 LEVELS
FINE ANALOG-TO-DIGITAL CONVERTER
DF
t₄
DIGITAL-TO-ANALOG CONVERTER
t₅
A TRANSMITTER
Dc
COARSE DIGITAL-TO-ANALOG CONVERTER
t₃
ANALOG OUT
H₂
DF
FINE DIGITAL-TO-ANALOG CONVERTER
t₅
B RECEIVER

Fig. 6 — Experimental sliding-scale direct-feedback PCM coder.

ficient capacity. The design procedure of Brainard and Candy was followed for the direct-feedback coder.[3]

Two feedback paths were used, one for coarse and one for fine quantized levels. Several switches were used to control the timing of the operations. Each switch was closed for 50 ns with a 50 ns space between successive operations so that all operations were completed in the sampling period of 0.5 μs.

The companded quantizing scale for the experimental coder is shown in Fig. 7. The fine quantizer was designed with seven levels. Level a was set at zero volts. The five inner levels a, $\pm b$, and $\pm c$ satisfied Max's first rule for minimum distortion. The two outer levels, $+d$ and $-d$, of the fine quantizer were assigned the same code words as the two coarse quantizer levels, $+d'$ and $-d'$. (Notice that in Section III levels $+d$ and $-d$, and $+d'$ and $-d'$, respectively, have the same value).

The decision levels for $+d$ and $-d$ were set slightly higher than the decision levels for $+d'$ and $-d'$. Thus code words for levels $\pm d'$ will always preceed the code words for levels $\pm d$. This permits the receiver to identify and assign the correct level to the d words. Although the optimum quantizing scale was not determined, some information in this direction was obtained. The coder was not sensitive to changes in the fine quantizing scale when the $+d$ and $-d$ levels did not exceed ten percent of the peak input signal. The coarse levels $+d'$ and $-d'$ prefer to be slightly more than twice the value of $+d$ and $-d$.

Examination of the quantizing scale, Fig. 7, shows that the 19 levels are not efficiently used. For instance, levels $+d' \pm b$ and $-d' \pm b$ produce substantially the same results as the $\pm d' + a$ level with signal changes of this magnitude. Therefore, the effective number of levels is more like 15 instead of 19. Since excellent results were obtained with the fifteen "effective" levels the techniques which would permit the effective use of all 19 levels were not tried.

V. LARGE SIGNAL CHANGES

5.1 *Frequency of Occurrence*

The frequency of occurrence of the two outer levels, $\pm d'$, was measured for the two still pictures shown in Figs. 8a and c. The results are listed in Table I for three levels of input signal. Picture A refers to the picture shown in Fig. 8a and Picture C to the picture shown in Fig. 8c. The position of levels $\pm d'$ for picture A is shown in Fig. 9 for three levels of input signal.

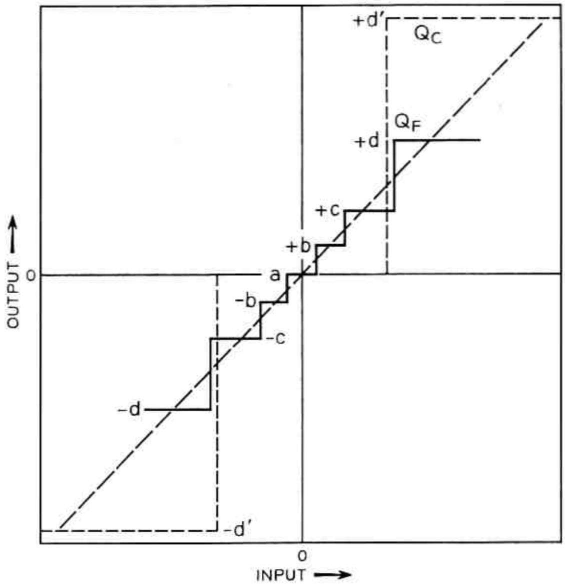


Fig. 7 — Companded quantizing scale of the experimental sliding-scale direct-feedback PCM coder.

5.2 *Capacity of Elastic Store*

The data of Table I provide a measure of the size of the elastic store that is required. Assuming that the average proportion of large changes per picture is the same as the average proportion for each line, an elastic store with a capacity of ten percent of the bits per line would be adequate to handle the three signal levels listed in Table

TABLE I—FREQUENCY OF OCCURRENCE OF LEVELS $\pm d'$ IN PERCENT OF TOTAL CHANGES FOR THREE INPUT SIGNAL LEVELS

Change in signal level	Picture A			Picture C		
	$-d'$	$+d'$	Total	$-d'$	$+d'$	Total
+5 dB	4.6	4.7	9.3	5.8	4.7	10.5
0 dB	1.1	1.2	2.3	1.6	1.8	3.4
−5 dB	0.4	0.6	1.0	0.4	0.4	0.8

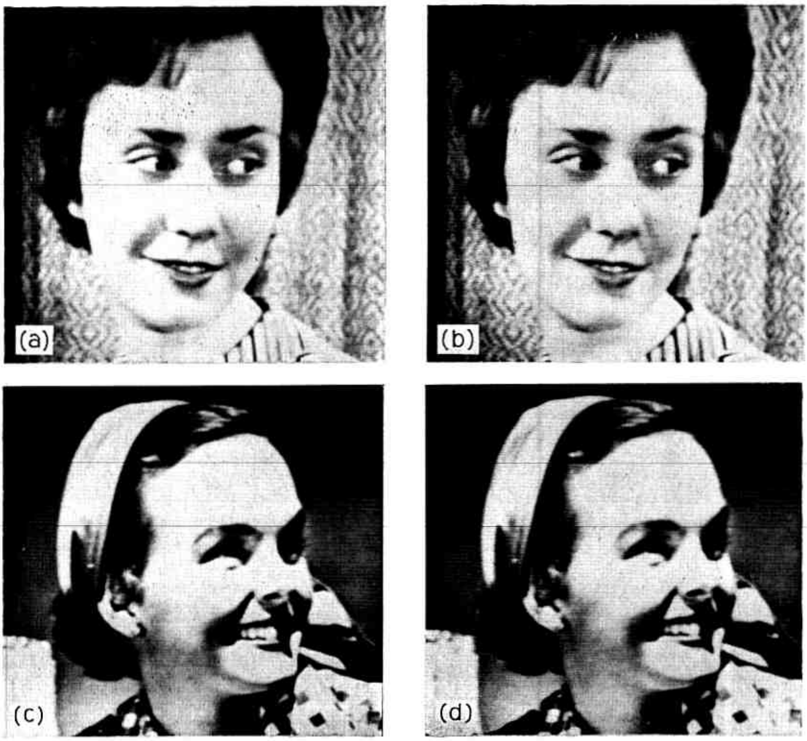


Fig. 8 — Photographs of television pictures bandlimited to 1.0 MHz: (a) high detail picture without coding, (b) high detail coded picture with optimum input signal level, (c) low detail picture without coding, and (d) low detail coded picture with optimum input signal level. (The scan lines and printing screen in Figs. 8 through 11 cause moiré patterns that are not in the originals.)

I. If the average number of large changes per line exceeds ten percent, the coder degrades to an eight-level coder. This "graceful" degradation occurs at the edge of the picture which in most cases will not be noticed.

VI. EVALUATION OF CODER

6.1 *Coder Environment*

The signal source was a television system consisting of a 275 line, 2:1 line interlaced picture, displaying 30 frames per second. The television signal was bandlimited to 1 MHz and sampled at a 2 MHz rate. The transmission bit rate was 6 MHz with 3-bits per sample. The picture display was 5½ inches by 5 inches and was viewed from 3½

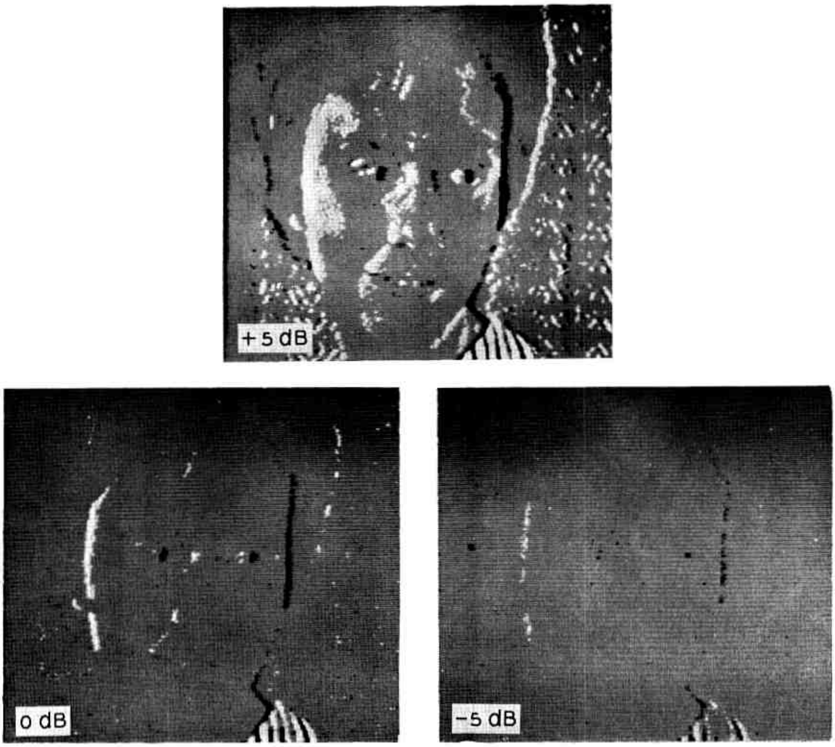


Fig. 9 — Television pictures showing where $\pm d'$ occurred for input signal level changes of —5dB, 0dB (optimum input level), and +5dB.

feet. The peak luminance was 70 foot lamberts and the room illumination was 100 foot candles.

6.2 *Picture Material*

Two types of still pictures (see Fig. 8) were used for the subjective evaluation; one with great detail and one with little detail. Evaluations were obtained for input signals which varied over a ±5 dB range. Figs. 8a and c show the uncoded pictures passed through the same low-pass filters as the coded pictures. Figs. 8b and d show coded pictures at the optimum input signal level. Fig. 10 shows the detailed picture with a −5 dB(a) and a +5 dB(b) change in input signal.

Photographs should be used with care in evaluating television presentations. Long exposures of photographs, compared with television frame time, will integrate noise and motion defects out of television pictures. The photographs of Fig. 8 may be compared except for the granular noise defects which appeared in Figs. 8b and c. The granular signal-to-noise ratio of these two pictures was 50 dB. Fig. 10a shows some of the effects of granular noise that appeared in the television picture. For this picture the input signal was reduced 5 dB from the reference level and had a granular signal-to-noise ratio of 47 dB. Fig. 10b illustrates slope-overload defects which occurred when the input signal was increased by 5 dB from the reference value. This defect is most apparent in the young woman's blouse. The granular signal-to-noise ratio in this case was 54 dB.

6.3 *Evaluation*

Evaluation of the coder using live subjects indicated that the defects listed in this section were more severe for the two still subjects. Therefore only the still subjects were used in the evaluation.

The six types of noise associated with differential type PCM coders were evaluated. The six types of noise are: granular noise, streaking, contouring, slope-overload, edge-busyness, and edge-stepping. These were evaluated subjectively by the author, except for slope-overload. The subjective evaluation was conducted on pictures when the sweep rates were locked to the digital processing rates and when they were not.

Contouring, edge-busyness, and streaking were not perceptible in either picture whether or not the sweep rates and the digital processing rates were locked. However, when the ratios between the outer levels,

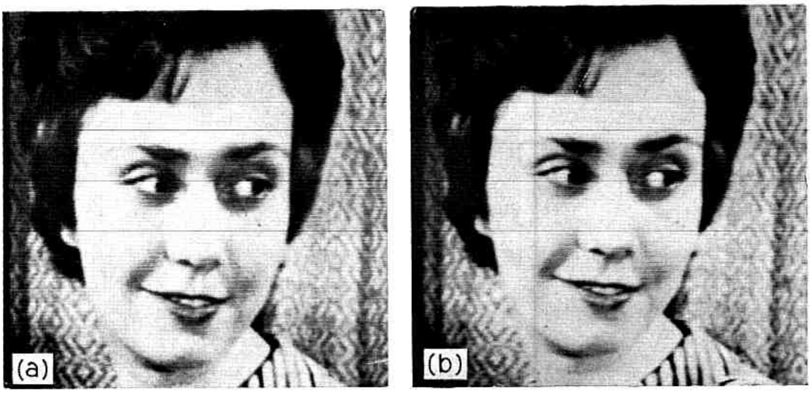


Fig. 10 — Detailed coded picture at two input signal levels: (a) decrease of 5 dB from optimum; (b) increase of 5 dB from optimum.

$+d'$ and $-d'$, and the inner levels were different between the transmitter and receiver, streaking did occur. The threshold for streaking permitted a difference in ratios of ±4 percent. Worst case transmission errors which occur in the d' and $-d'$ words produce an error signal which decays exponentially to zero in 6 μs and appears as a streak about 0.05 inches long on the picture. The photographs in Fig. 11 show the effect of an error locked to the line rate in a $-d'$ and a $+d'$ word.

Edge-stepping was not perceptible when the sweep rates and the digital processing rates were locked. When they were unlocked, edge-stepping was just perceptible at large brightness boundaries.

Slope-overload was measured objectively. A slide which provided a 75 percent white to black transition along the scanning lines was placed in front of the camera. An oscilloscope was used to measure the transition time from white to black for several input levels at the input to the monitor. The results are shown in Fig. 12 and a typical waveform shown in Fig. 13. The rise-time varied from 1.0 to 1.45 μs over a 12.5 dB range of input signals, where the input signal was limited to a rise-time of 1.0 μs. At the optimum input signal level (0 dB), the rise-time increased to 1.15 μs. There was no measurable difference in the slope response when the sweep rates and the digital processing rates were unlocked.

Granular noise was measured subjectively by comparing the coded picture with an uncoded picture to which gaussian noise had been added. The granular noise resulting from coding was a high frequency

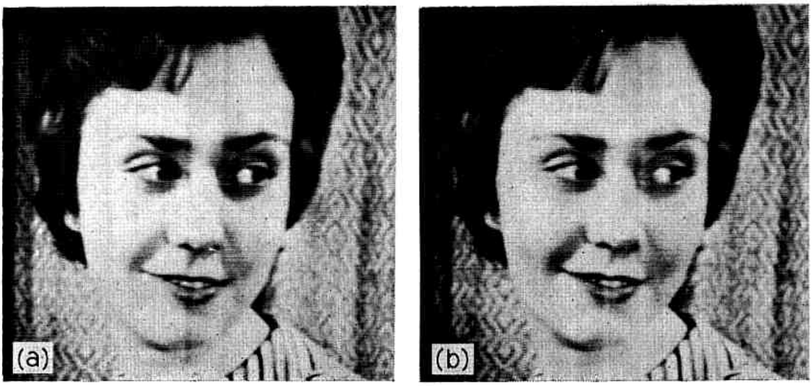


Fig. 11 — Effect of transmitting erroneous $\pm d'$ word when the erroneous word is locked to the scanning rates: (a) $-d'$ error, (d) $+d'$ error.

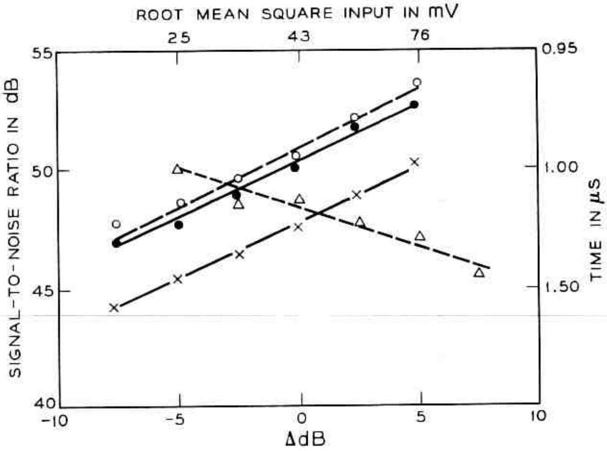


Fig. 12 — Measurements of signal-to-noise ratio and slope-overload as a function of input signal level for the sliding-scale coder.

noise occurring at or near the sampling rate. Since the gaussian noise occupied the full band, some subjective weighting was necessary. The uncoded picture with added noise was passed through the same low-pass filters as the coded picture. The two pictures with equalized contrast were alternately switched onto the monitor and the added noise adjusted until they were judged subjectively equal. The signal-to-noise was measured in terms of peak-to-peak signal to root mean square noise on the uncoded picture with added noise. The results are shown in Fig. 12. The equivalent signal-to-noise of the two test pictures was substantially the same, ranging from 47 dB to 54 dB over an input signal range of 10 dB. With the input signal level optimized (0 dB) the equivalent signal-to-noise was 50 dB. The signal-to-noise of a companded 3-bit differential pulse code modulation coder, using the same measuring technique, was 45 dB. When the picture sweep rates and the digital processig rates were unlocked, the signal-to-noise was decreased by 3 dB. This decrease in signal-to-noise is caused by sampling position differences from frame to frame at the smaller brightness boundaries.

VII. CONCLUSIONS

This experiment, with the sliding-scale coder, demonstrated that 15 "effective" levels are sufficient to produce a high quality television

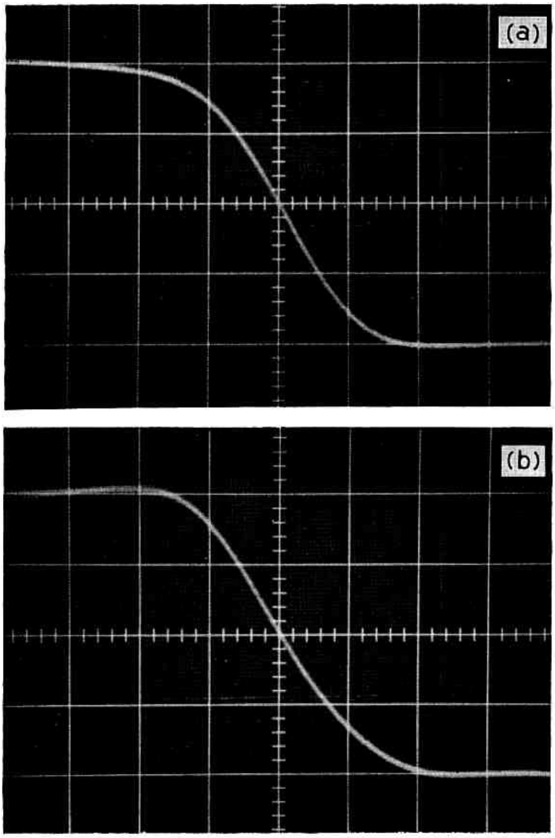


Fig. 13 — Waveform response to a 75 percent change in brightness level when bandlimited to 1.0 MHz: (a) analog response; (b) response of sliding-scale coder.

picture with a small increase in circuit cost and complexity. The increase in circuit cost and complexity is offset by the double processing technique which reduces the requirements on the number of threshold and quantizing circuits.

The most significant improvement offered by the sliding-scale coder is in the rendition of the subjectively critical large brightness changes. The coder performance approaches that of a seven-bit PCM system.

A reduction in the sampling rate of about five percent permits the sliding-scale coder to nearly triple the number of quantizer levels without an increase in channel bit rate.

VIII. ACKNOWLEDGMENTS

Discussions with J. C. Candy, C. C. Cutler, H. Kogelnik, and J. O. Limb were most helpful. Mr. R. K. Jackson assembled the coder.

REFERENCES

1. Cutler, C. C., "Differential Quantization of Communication Signals," U. S. Patent 2-605-361, applied for June 29, 1950; issued July 29, 1952.
2. Graham, R. E., "Predictive Quantizing of Television Signals," IRE Wescon Conv. Record, Part 4, August 1958, pp. 147–157.
3. Brainard, R. C., and Candy, J. C., "Direct-Feedback Coders: Design and Performance with Television Signals," Proc. IEEE, 57, No. 5 (May 1969), pp. 776–786.
4. deJager, F., "Delta Modulation, A Method of PCM Transmission Using a 1-Unit Code," Philips Res. Report, *7*, No. 6 (December 1952), pp. 442–466.
5. O'Neal, J. B., "Delta Modulation Quantizing Noise, Analytical and Computer Simulation Results for Gaussian and Television Input Signals," B.S.T.J., *45*, No. 1 (January 1966), pp. 117–151.
6. Brown, E. F., unpublished work.
7. J. Max, "Quantizing for Minimum Distortion," IRE Trans. Inform. Theor., *IT 6*, No. 1 (March 1960), pp. 7–12.

Contributors to This Issue

M. R. Aaron, B.S.E.E., 1949, M.S., 1951, University of Pennsylvania; Bell Telephone Laboratories, 1951—. His work was initially concerned with the design of networks for various transmission systems including the first transatlantic submarine cable system. Since 1956 he has been involved in analytical work on various PCM systems. Fellow, IEEE, American Association for the Advancement of Science.

R. T. Aiken, B.S., 1957, M.S., 1959, and Ph.D., 1962, Carnegie-Mellon University; U. S. Army, 1961–1963; Bell Telephone Laboratories, 1963—. Mr. Aiken has been concerned with problems in radar, sonar, and communication theory, with emphasis on the effects of random media. He is a supervisor in the Outside Plant Engineering Department. Member, IEEE, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

Raymond H. Bosworth, graduate, Air Force Radio Technical School, 1950; attended Union Junior College, 1952–56; R.C.A. Institutes 1962–64; Bell Telephone Laboratories, 1952—. Mr. Bosworth has worked on negative impedance repeaters, repertory dial telephones, and PCM coders, and now is investigating color television. He holds a patent on a universal printed circuit card.

Earl F. Brown, RCA Institutes, Inc., 1955; Bell Telephone Laboratories, 1955—. Since joining Bell Telephone Laboratories Mr. Brown has been engaged in developing techniques for enchancing the quality of television pictures, subjectively evaluating television pictures, and discovering means to compress the bandwidth of television pictures.

Morgan M. Buchner, Jr., B.E.S., 1961, Ph.D., 1965, The Johns Hopkins University; Bell Telephone Laboratories, 1965—. Mr. Buchner has been interested in problems related to data transmission. Member, IEEE, Tau Beta Pi, Sigma Xi, Eta Kappa Nu.

Christoph B. Burckhardt, Dipl.-Ing., 1959, Dr. sc. techn., 1963, Swiss Federal Institute of Technology; Bell Telephone Laboratories 1963—. Initially Mr. Burckhardt was engaged in the analysis of varactor frequency multipliers. Since 1965, he has been working in holography. Member IEEE, Optical Society of America.

J. C. CANDY, B.Sc., 1951, Ph.D., 1954, University of Wales Bangor; Bell Telephone Laboratories, 1960—. Mr. Candy has worked on digital circuits and pulse transmission schemes. He is concerned with video signal processing methods.

MERTON H. CROWELL, B.S. (E.E.), 1956, Pennsylvania State University; M.S. (E.E.), 1960, New York University; Ph.D. (Electrophysics), 1966, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1956—. Mr. Crowell initially worked on the development of the high-speed pulse code modulation coding tube. He later studied optical modulators and detectors in relation to optical maser communication systems, and recently has investigated the interaction of electron beams with semiconductor devices. Member, IEEE, Eta Kappa Nu, Tau Beta Pi.

HANS G. DANIELMEYER, Dipl. Phys., 1962, Dr. rer. nat., 1965, Stuttgart University, Stuttgart, Germany, Research Staff, Stuttgart University, 1962–1966; Bell Telephone Laboratories, 1966—. In Stuttgart, Mr. Danielmeyer did research in ultrasonics on molecular liquids and solids. At Bell Telephone Laboratories he is doing research on light scattering and developing suitable lasers.

LOUIS H. ENLOE, B.S.E.E., 1955, M.S.E.E., 1956, and Ph.D. (E.E.), 1959, University of Arizona; instructor in electrical engineering and a member of the technical staff of the Applied Research Laboratory of the University of Arizona, 1956–1959; Bell Telephone Laboratories, 1959—. Mr. Enloe's early work was in modulation and noise theory in connection with space communications. Later work has been with lasers, coherent light, and holography, with emphasis on communication and display. He is Head of the Opto-Electronics Research Department. Member, IEEE, Phi Kappa Phi, Sigma Xi, Tau Beta Pi, Pi Mu Epsilon, Sigma Pi Sigma.

JAMES G. EVANS, B.E.E., 1963, M.E., 1964, Cornell University; Bell Telephone Laboratories, 1963—. Mr. Evans has been designing laboratory instrumentation to support transmission systems development. Now he is working on computer-operated instrumentation for automatically measuring network characteristics in the 1 to 12 GHz range. Member, IEEE, Eta Kappa Nu, Tau Beta Pi.

Jerome S. Fleischman, B.E.E., 1963, Cooper Union; M.S.E.E., 1964, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1963—. His work has concerned the design of various circuits for high speed PCM and differential PCM encoding terminals. He has taught the Transmission Systems Design course given in the GSP program. Member, IEEE, Tau Beta Pi, Eta Kappa Nu.

Walter J. Geldart, B.Eng. in E.E., 1958, McGill University; M.Eng. in E.E., 1962, McMaster University; Bell Telephone Laboratories, 1962—. Mr. Geldart has been engaged in the design and development of computer controlled transmission measuring systems. He is a supervisor in the Measuring Systems Design Department. Member, IEEE.

Detlef Gloge, Dipl. Ing., 1961, D.E.E., 1964, Braunschweig Technische Hochschule (Germany); research staff, Braunschweig Technische Hochschule 1961–1965; Bell Telephone Laboratories, 1965—. In Braunschweig, Mr. Gloge was engaged in research on lasers and optical components. At Bell Telephone Laboratories, he has concentrated in the study of optical transmission techniques. Member, VDE, IEEE.

David J. Goodman, B.E.E., 1960, Rensselaer Polytechnic Institute; M.E.E., 1962, New York University; Ph.D., 1967, University of London; Bell Telephone Laboratories 1960–62, 1967—. Mr. Goodman has performed analytic studies of digital communication systems and digital signal processing techniques. Member, IEEE, Eta Kappa Nu, Tau Beta Pi.

B. Gopinath, M.S. (mathematical physics), 1964, University of Bombay, India; M.S.(E.E.) and Ph.D.(E.E.), 1968, Stanford University; postdoctoral research associate at Stanford from November 1967 to April 1968; Bell Telephone Laboratories, 1968—. Mr. Gopinath's primary interest, as a member of the Systems Theory Research Group, is in the applications of mathematical methods to physical problems.

GERALD D. HAYNIE, B.S.E.E., 1956, Virginia Polytechnic Institute; M.E.E., 1961, New York University; Bell Telephone Laboratories, 1956—. Mr. Haynie has been engaged in the development of precision measuring systems used in transmission systems development. Recent emphasis has been in computer operated systems. He is head of the Measuring Systems Design Department. Member, IEEE, Eta Kappa Nu, Tau Beta Pi.

EDWARD F. LABUDA, B.S. (Physics), 1959, Case Western Reserve University; M.S.E.E., 1961, New York University; Ph.D. (Electrophysics), 1967, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1959—. Mr. Labuda was initially concerned with the development of low noise microwave tubes and later became involved in research and development of gas lasers. Recently he has participated in the development of the silicon diode array camera tube for *Picturephone*® video telephone systems. He now supervises a group concerned with semiconductor device physics. Member, AAAS.

RICHARD A. MCDONALD, B.E. (E.E.), 1956, M.E. (E.E.), 1957, and D.E. (E.E.), 1961, Yale University; Assistant Professor of Engineering and Applied Science, Yale University, 1961–1964; Bell Telephone Laboratories, 1964—. Mr. McDonald has been concerned with aspects of the development of digital communication systems such as PCM, differential PCM, and delta modulation for speech and video systems. More recently he has been in charge of a group engaged in system engineering studies of video network service, metropolitan telephone service, maintenance systems, and network interconnections. Member, IEEE, Sigma Xi, Tau Beta Pi, American Society for Engineering Education.

E. N. PROTONOTARIOS, Electrical Engineer, 1963, National Technical University of Athens, Greece; Eng. Sc.D., 1966, Columbia University; Bell Telephone Laboratories, 1966–1968. Mr. Protonotarios was engaged in analytical studies of digital communications systems. He is presently on leave of absence from the Laboratories to teach at Columbia University. Member, Sigma Xi, IEEE, American Association for the Advancement of Science.

LAWRENCE R. RABINER, S.B. and S.M., 1964, Ph.D. (E.E.) 1967, Massachusetts Institute of Technology. From 1962 through 1964 he participated in the cooperative plan in electrical engineering at Bell

Telephone Laboratories, Whippany and Murray Hill, New Jersey. He worked on digital circuitry, military communications problems, and problems in binaural hearing. He joined the staff of Bell Laboratories in 1967 and has been engaged in research on speech communication, signal analysis, and techniques for waveform processing. Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, IEEE, Acoustical Society of America.

Charles M. Rader, B.E.E., 1960, M.E.E., 1961, Brooklyn Polytechnic Institute. He joined the staff of Lincoln Laboratory, Massachusetts Institute of Technology, in 1961. He has worked in the areas of speech compression, system simulation, and digital signal processing. He is coauthor of a book on modern techniques for signal processing. Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, Acoustical Society of America, IEEE.

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.

Ronald W. Schafer, B.S.(E.E.), 1961, M.S.(E.E.), 1962, University of Nebraska; Ph.D., 1968, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1968—. Mr. Schafer has been concerned with nonlinear digital waveform processing techniques for deconvolution. He is doing research on signal processing techniques for speech communication systems. Member, Phi Eta Sigma, Eta Kappa Nu, Sigma Xi, IEEE, Acoustical Society of America.

R. G. Schleich, A.A.S., 1959, Mohawk Valley Community College; Bell Telephone Laboratories, 1959—. Mr. Schleich has been engaged in the development of laboratory transmission measuring systems. His work has included a video frequency distortion measuring set and an automated 20 Hz to 20 KHz transmission measuring set. He is engaged in Holmdel Measurement Center development.

M. V. Schneider, M.S., 1956, and Ph.D., 1959, Swiss Federal Institute of Technology, Zurich, Switzerland; Bell Telephone Laboratories, 1962—. Mr. Schneider has been engaged in experimental work on thin film solid state devices, optical detectors, and microwave integrated circuits. Member, IEEE, American Vacuum Society.

Peter W. Smith, B.S., 1958, M.S., 1961, Ph.D., 1964, McGill University (Canada); Bell Telephone Laboratories, 1963—. Mr. Smith has done research on gas laser output power and stability problems; he has also investigated a number of systems for obtaining single-frequency laser operation. Now he is studying pulse propagation in gas laser amplifiers. Member, Canadian Association of Physicists, American Physical Society, IEEE.

William H. Steier, B.S.E.E., 1955, Evansville College; M.S.E.E., 1957, and Ph.D.(E.E.), 1960, University of Illinois; Bell Telephone Laboratories, 1962–1968. Mr. Steier first worked on the millimeter wave circular waveguide transmission system. More recently he had worked on optical transmission lines and gas lenses. He is now with the Department of Electrical Engineering at the University of Southern California. Member, IEEE, Sigma Xi.

Jeffrey D. Ullman, B.S., 1963, Columbia University; Ph.D.(E.E.), 1966, Princeton University; Bell Telephone Laboratories, 1966—. Dr. Ullman has worked in research in computer science, principally language theory and switching theory. Member, Tau Beta Pi, Sigma Xi, Association for Computing Machinery, IEEE.

Peter Weiner, B.E.E., 1963, City College of New York; M.S., 1964, and Ph.D., 1967, Polytechnic Institute of Brooklyn. Since 1966 he has been an assistant professor of electrical engineering at Princeton University, Princeton, N.J., and has served as a consultant to Bell Laboratories since April 1968. Member, Association for Computing Machinery, Eta Kappa Nu, Sigma Xi, IEEE.

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.

Robert W. Wilson, B.A., (Physics), 1957, Rice University; Ph.D. (Physics), 1962, California Institute of Technology, Bell Telephone

Laboratories, 1963—. At Bell Laboratories Mr. Wilson has made radio astronomical and propagation measurements. In radio astronomy his work includes measurements of: absolute fluxes of radio sources, the cosmic background temperature, the disk component of the galaxy, and intergalactic hydrogen. His propagation measurements include measurements of 10μ and the short centimeter region. He continues to work in both fields. Member, American Astronomical Society, International Scientific Radio Union, Sigma Xi, Phi Beta Kappa.