MAY-JUNE
VOL. 53 NO. 5

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published ten times a year by the American Telephone and Telegraph Company, J. D. deButts, Chairman and Chief Executive Officer, R. D. Lilley, President, J. J. Scanlon, Executive Vice President and Chief Financial Officer, F. A. Hutson, Jr., Secretary. Checks for subscriptions should be made payable to American Telephone and Telegraph Company and should be addressed to the Treasury Department, Room 1038, 195 Broadway, New York, N. Y. 10007. Subscriptions $11.00 per year; single copies $1.50 each. Foreign postage $1.00 per year; 15 cents per copy. Printed in U.S.A.

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 53 May–June 1974 Number 5

Capacity of the Gaussian Channel With Memory: The Multivariate Case

By L. H. BRANDENBURG and A. D. WYNER

(Manuscript received November 13, 1973)

A formula is derived for the capacity of a multi-input, multi-output linear channel with memory, and with additive Gaussian noise. The formula is justified by a coding theorem and converse. The channel model under consideration can represent multipair telephone cable including the effect of far-end crosstalk. For such cable under large signal-to-noise conditions, we show that channel capacity and cable length are linearly related; for small signal-to-noise ratio, capacity and length are logarithmically related. Crosstalk tends to reduce the dependence of capacity on cable length. Moreover, for any channel to which our capacity formula applies, and for large signal-to-noise ratio, there is an asymptotic linear relation between capacity and signal-to-noise ratio with slope independent of the channel transfer function. For small signal-to-noise ratio, capacity and signal-to-noise ratio are logarithmically related. Also provided is a numerical evaluation of the channel capacity formula, using measured parameters obtained from an experimental cable.

I. INTRODUCTION AND STATEMENT OF RESULTS

Our problem is to calculate the capacity of a multi-input, multi-output linear channel with additive Gaussian noise, and to justify the formula by a coding theorem and converse. Specifically, we consider the following channel. The channel input and output are sequences

$\{x(n)\}_{-\infty}^{\infty}$, $\{y(n)\}_{-\infty}^{\infty}$ of real s-vectors* related by

$$y(n) = \sum_{k=-\infty}^{\infty} h(n-k)x(k) + z(n), \tag{1}$$

where $\{h(m)\}_{m=-\infty}^{\infty}$ is a fixed sequence of real $s \times s$ matrices (the indicated operations being ordinary matrix arithmetic), and $\{z(n)\}_{-\infty}^{\infty}$ is a sequence of Gaussian random s-vectors for which $Ez(n) = 0$, and

$$Ez(n)[z(n-m)]^t = r(m), \quad -\infty < n, \quad m < \infty, \tag{2}$$

where $r(m)$ is an $s \times s$ matrix. The motivation for this problem is that it is a model for a multipair telephone cable.

The first sections of this paper are highly theoretical; the formula for channel capacity is carefully and precisely established by means of several rather technical theorems. In the final section, Section IV, we discuss some engineering implications of our formula in terms of its asymptotic behavior, and evaluate the capacity numerically with measured parameters obtained from an experimental multipair telephone cable.

A *code* for this channel with parameters (M, N, S, λ) is a set of M pairs $\{(\mathbf{x}_i, B_i)\}_{i=1}^{M}$, where $\mathbf{x}_i = (\cdots, x_i^t(-2), x_i^t(-1), x_i^t(0), x_i^t(1), \cdots)^t$ is a sequence of s-vectors that satisfy the following:

$$x_i(n) = 0, \quad \text{for} \quad n < 0, \quad \text{and} \quad n \geqq N, \tag{3a}$$

$$\frac{1}{N}\sum_{n=0}^{N-1} \|x_i(n)\|^2 \leqq S, \tag{3b}$$

(where $\|\cdot\|$ denotes Euclidean norm) and the B_i are (measurable) subsets of $\mathbb{R}^{sN}$ with the following property. Let $\mathbf{y}_i = (\cdots, y_i^t(-1), y_i^t(0), y_i^t(1), \cdots)^t$ be the channel output vector which results when the channel input is $\mathbf{x}_i$, i.e.,

$$y_i(n) = \sum_{k=-\infty}^{\infty} h(n-k)x_i(k) + z(n)$$

$$= \sum_{k=0}^{N-1} h(n-k)x_i(k) + z(n).$$

Let $\mathbf{y}_i^{(N)} = [y_i^t(0), \cdots, y_i^t(N-1)]^t \in \mathbb{R}^{sN}$. Then $B_i (1 \leqq i \leqq M)$ must satisfy

$$P_{ei} \triangleq Pr\{\mathbf{y}_i^{(N)} \in B_i\} \leqq \lambda. \tag{4}$$

* Vectors will be taken to be column matrices unless otherwise indicated.

Thus the $\mathbf{x}_i$ are code words and B_i is the set of output N-vectors which are decoded at the channel output as $\mathbf{x}_i$. Inequality (4) expresses the requirement that the error probability, given that $\mathbf{x}_i$ is transmitted, does not exceed λ.

A number $\rho \geqq 0$ is said to be "S-admissible" ($S \geqq 0$) if for all $\lambda > 0$ there is an N such that there exists a code with parameters $([2^{N\rho}], N, S, \lambda)$. The *channel capacity* C_S is defined as the supremum of S-admissible rates. Our problem is the calculation of C_S, which we shall solve provided the channel satisfies the following conditions:

(*i*) *The filter* $\{h(m)\}$: We assume that the filter is causal, i.e., $h(m) = 0$, $m < 0$. We also assume that

$$\sum_{m=0}^{\infty} \|h(m)\| < \infty, \tag{5a}$$

and that there exists a $B > 0$ such that for $m > 0$,

$$\|h(m)\| \leqq Bm^{-1}, \tag{5b}$$

where the Euclidean norm "$\|\cdot\|$" of a matrix is the square root of the sum of the squares of its entries. From (5), the (discrete) transfer function,

$$H(\theta) = \sum_{n=0}^{\infty} h(n)e^{in\theta}, \quad -\pi \leqq \theta \leqq \pi, \tag{6}$$

exists and is continuous. $H(\theta)$ is an $s \times s$ matrix. We assume that for $-\pi \leqq \theta \leqq \pi$, $\det H(\theta) \neq 0$.

(*ii*) *The noise covariance*: We assume that the covariance sequence $r(\cdot)$ satisfies

$$\sum_{n=-\infty}^{\infty} \|r(n)\| < \infty, \tag{7}$$

so that the (discrete) power spectral density

$$R(\theta) = \sum_{n=-\infty}^{\infty} r(n)e^{in\theta}, \quad -\pi \leqq \theta \leqq \pi \tag{8}$$

exists and is continuous. $R(\theta)$ is an $s \times s$ matrix. We also assume that for $-\pi \leqq \theta \leqq \pi$, $\det R(\theta) \neq 0$.

We can now give the capacity formula. Let the $s \times s$ matrix[†] $\Gamma(\theta) = H(\theta)^{-1}R(\theta)H(\theta)^{-*}$, and let $\lambda_1(\theta), \lambda_2(\theta), \cdots, \lambda_s(\theta)$ be the eigen-

[†] For any nonsingular complex matrix A, A^{-*} is the transpose conjugate of A^{-1}.

values of $\Gamma(\theta)$, $-\pi \leqq \theta \leqq \pi$. Then $\lambda_j(\theta) > 0$, $1 \leqq j \leqq s$, $-\pi \leqq \theta \leqq \pi$. Let $S \geqq 0$ be given, and let K_S be the (unique) positive number such that

$$\frac{1}{2\pi} \sum_{j=1}^{s} \int_{-\pi}^{\pi} d\theta \max[0, K_S - \lambda_j(\theta)] = S. \tag{9a}$$

Then

$$C_S = \frac{1}{4\pi} \sum_{j=1}^{s} \int_{-\pi}^{\pi} d\theta \max\left(0, \log_2 \frac{K_S}{\lambda_j(\theta)}\right). \tag{9b}$$

Our main result is the following. Consider the channel defined by (1) with $H = H_1(\theta)$ and $R = R_1(\theta)$. Then C_S is calculated from (9) with $\Gamma(\theta) = H_1(\theta)^{-1}R_1(\theta)H_1(\theta)^{-*}$.

Theorem 1a (Converse): Let $\rho \geqq 0$ be an S-admissible rate for this channel. Then $\rho \leqq C_S$.

Theorem 1b (Direct-Half): Let $S \geqq 0$, $\epsilon > 0$ and $\rho(0 \leqq \rho < C_S)$ be arbitrary. Then for N sufficiently large, there exists a code with parameters (M, N, S, λ) where

$$M \geqq e^{\rho N} \quad \text{and} \quad \lambda \leqq \epsilon.$$

Sections II and III of this paper are concerned with the proof of Theorem 1. Section IV is concerned with the asymptotic behavior and numerical evaluation of the channel capacity formula (9), with specific attention to multipair telephone cable.

Theorem 1 is very similar to the results on continuous-time Gaussian channels due to Holsinger and Gallager.[1] In fact, for the special case in which $H(\theta)$ is the $s \times s$ identity, the theorem follows immediately from the analysis in Ref. 1. We suspect that it might be possible to obtain all of Theorem 1 by paralleling Gallager's techniques for this discrete case, although such an approach is somewhat more cumbersome than the approach followed here. Furthermore, the present approach lends itself immediately to broadening the model to consider the effects of intersymbol interference from previous channel uses, as Gallager's approach does not.[2] In fact, to establish Theorem 1b for the intersymbol interference channel we require only to add in one of our lemmas a term "γ_3" (representing the effect of previous channel uses), and to show that its norm $|||\gamma_3||| = o(N^{\frac{1}{2}})$.

Careful analysis of the proof of Theorem 1 will indicate that the conditions on the filter $\{h(m)\}$ given in Section I can be replaced by simply requiring that the filter have a causal inverse $\{g(m)\}$, such that $g(m) = 0$, $m \geqq m_0$ (i.e., finite memory). Thus, our results contain a generalization of those given by Toms and Berger.[3]

II. NOTATION AND MATHEMATICAL PRELIMINARIES

Let $l_2^{(s)}(a, b)$, $s = 1, 2, \cdots$, $-\infty \leqq a < b \leqq \infty$, where a, b are integers, be the set of sequences $\{x(n)\}_{n=a}^{b}$ where $x(n)$ is a real s-vector. Such sequences will often be written as column matrices $\mathbf{x} = [x^t(a), x^t(a+1), \cdots, x^t(b)]^t$. Let $\|\cdot\|$ denote ordinary Euclidean norm in s-space, and for $s \times s$ matrices A, let $\|A\|$ be the Euclidean norm (i.e., the square root of the sum of the squares of the components of A). For sequences $\mathbf{x} \in l_2^{(s)}(a, b)$, the (Euclidean) norm is

$$|||\mathbf{x}||| = \left[\sum_{n=a}^{b} \|x(n)\|^2 \right]^{\frac{1}{2}}. \tag{10}$$

The space $l_2^{(s)}(a, b)$ is a Hilbert space with the obvious inner product, written $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{n=a}^{b} x^t(n)y(n)$, $\mathbf{x},\mathbf{y} \in l_2^{(s)}(a, b)$. For $\mathbf{x} \in l_2^{(s)}(-\infty, \infty)$, denote by $\mathbf{x}^{(N)} \in l_2^{(s)}(0, N-1)$ the column matrix

$$\mathbf{x}^{(N)} = [x^t(0), x^t(1), \cdots, x^t(N-1)]^t. \tag{11}$$

We denote operators on $l_2^{(s)}(a, b)$ by script letters, e.g., $\mathcal{F}$. We define the norm of $\mathcal{F}$ by

$$|\mathcal{F}| = \sup_{\mathbf{x} \in l_2^{(s)}(a,b)} \frac{|||\mathcal{F}\mathbf{x}|||}{|||\mathbf{x}|||}. \tag{12}$$

If $|\mathcal{F}| < \infty$, we say that $\mathcal{F}$ is bounded. An operator $\mathcal{F}$ is said to be of a convolution type if, for $\mathbf{x} \in l_2^{(s)}(a, b)$,

$$(\mathcal{F}\mathbf{x})(n) = \sum_{k=a}^{b} f(n-k)x(k), \quad n = a, \cdots, b, \tag{13}$$

where $\{f(n)\}_{a-b}^{b-a}$ is a fixed sequence of $s \times s$ matrices and the indicated operations are ordinary matrix arithmetic. Let L be the set of convolution-type operators on $l_2^{(s)}(-\infty, \infty)$ for which

$$\sum_{n=-\infty}^{\infty} \|f(n)\| < \infty. \tag{14}$$

For operators $\mathcal{F}$ in L the transfer matrix

$$F(\theta) = \sum_{n=-\infty}^{\infty} f(n)e^{in\theta}, \quad -\pi \leqq \theta \leqq \pi \tag{15}$$

is well defined. $F(\theta)$ is an $s \times s$ matrix and is continuous (in Euclidean norm) for $-\pi \leqq \theta \leqq \pi$. Concatenation of operators $\mathcal{F}_1$, $\mathcal{F}_2 \in L$, defined by sequences $\{f_1(n)\}$ and $\{f_2(n)\}$, results in a convolution-type operator $\mathcal{F}_3 = \mathcal{F}_2 \cdot \mathcal{F}_1$ in L defined by the sequence $\{f_3(n)\}$ where

$f_3(n) = \sum_{k=-\infty}^{\infty} f_2(k)f_1(n-k)$. Further, the corresponding transfer functions satisfy $F_3(\theta) = F_2(\theta)F_1(\theta)$. Other relevant properties of the class L are given in the theorem below, the assertions of which are generalizations of well-known scalar results. The proof is given in the appendix (Section A.1).

Theorem 2: Given $\mathfrak{F} \in L$, defined by $\{f(n)\}$ or *$F(\theta)$, then*

a. $|\mathfrak{F}| \leqq \sum_{n=-\infty}^{\infty} \|f(n)\| < \infty$, *so that $\mathfrak{F}$ is bounded on $l_2^{(s)}(-\infty, \infty)$.*

b. $|\mathfrak{F}| \leqq \max\limits_{-\pi \leqq \theta \leqq \pi} \|F(\theta)\|$.

c. If $\mathfrak{F}$ is self-adjoint, then for all $\mathbf{x} \in l_2^{(s)}(-\infty, \infty)$,

$$\left|\sum_{n,m} x^t(n)f(n-m)x(m)\right| \leqq \left(\max_{-\pi \leqq \theta \leqq \pi} \|F(\theta)\|\right)|||\mathbf{x}|||^2.$$

d. $\mathfrak{F}$ has a bounded inverse denoted $\mathfrak{F}^{-1}$ if and only if det $F(\theta) \neq 0$, $-\pi \leqq \theta \leqq \pi$, in which case the transfer matrix corresponding to $\mathfrak{F}^{-1}$ is $[F(\theta)]^{-1}$, and $\mathfrak{F}^{-1} \in L$.

Let $\mathbf{z} = [\cdots, z^t(-1), z^t(0), z^t(1), \cdots,]^t$ be a sequence of random s-vectors with covariance

$$Ez(n)z(n-m)^t = r(m),$$

where $r(m)$ is an $s \times s$ matrix. Under the assumption that

$$\sum_{m=-\infty}^{\infty} \|r(m)\| < \infty, \tag{16}$$

the power spectrum

$$R(\theta) = \sum_{m=-\infty}^{\infty} r(m)e^{im\theta}, \quad -\pi \leqq \theta \leqq \pi, \tag{17}$$

is well defined. Let $\mathfrak{F}$ be an operator in L corresponding to the transfer matrix $F(\theta)$. Then $\hat{\mathbf{z}} = \mathfrak{F}\mathbf{z}$ is a sequence of random s-vectors with covariance $\{\hat{r}(m)\}$ and corresponding power spectrum $\hat{R}(\theta) = F(\theta)\, R(\theta)F(\theta)^*$.

Let $\mathbf{z}$ be a sequence of zero mean Gaussian n-vectors with covariance $r(m)$ satisfying (16) and power spectrum $R(\theta)$. Let $\lambda_{\min}(\theta)$ be the minimum eigenvalue of the matrix $R(\theta)$. Let $\mathbf{z}^{(N)} = [z^t(0), z^t(1), \cdots, z^t\,(N-1)]^t$ be a segment of $\mathbf{z}$ of duration N. Then there exists an $(N \cdot s) \times (N \cdot s)$ matrix T_N such that

$$\mathbf{w} = T_N\mathbf{z}^{(N)}, \tag{18}$$

is "white," i.e., $E\mathbf{w}\mathbf{w}^t = I_{N \cdot s}$. The indicated operation in (18) is matrix multiplication. The only property of T_N which we need here is that for any $N \cdot s$-vector $\mathbf{u}$

$$|||T_N\mathbf{u}||| \leqq \left[\frac{1}{\min\limits_{-\pi \leqq \theta \leqq \pi} \lambda_{\min}(\theta)} \right]^{\frac{1}{2}} |||\mathbf{u}|||. \tag{19}$$

Finally, let $\mathcal{H}$ be the convolution-type operator on $l_2^{(s)}(-\infty, \infty)$ defined by $\{h(m)\}$ (Section I). Note that by (5) $\mathcal{H} \in L$, so that by the assumption following (6) and by Theorem 2d it has an inverse, say $\mathcal{G} = \mathcal{H}^{-1}$, of the convolution type in L. Let $\{g(n)\}_{-\infty}^{\infty}$ be the sequence which defines $\mathcal{G}$. Let the operator $\mathcal{G}_N(N = 0, 1, 2, \cdots)$ be defined by

$$(\mathcal{G}_N\mathbf{x})(n) = \sum_{k=0}^{N-1} g(n-k)x(k), \quad -\infty < n < \infty. \tag{20}$$

We conclude this section by stating as lemmas two known results. We explain in the appendix (Section A.2) how to obtain these results from published material.

Lemma 3: For the special case when $H(\theta) = I_s$ (the $s \times s$ identity) and $R(\theta) = R_2(\theta)$ [so that $\Gamma = H^{-1}RH^{-*} = R_2(\theta)$], say that $\mathbf{x}$ is a random channel input sequence for which $x(n) = 0$, $n < 0$, $n \geqq N$, and $E|||\mathbf{x}|||^2 \leqq NS(S > 0$, $N = 0, 1, 2, \cdots)$. Let $\mathbf{y}$ be the corresponding channel output sequence. Then, the mutual information

$$I\{\mathbf{x}, \mathbf{y}\} \leqq NC_S$$

(where C_S is calculated with $\Gamma(\theta) = R_2(\theta)$).

Lemma 4: For the special case where $H(\theta) = I_s$ and $R(\theta) = R_2(\theta)$, then we have a stronger version of Theorem 1b: Let $S \geqq 0$ and ρ $(0 \leqq \rho < C_S)$ be arbitrary, where C_S is calculated with $\Gamma(\theta) = R_2(\theta)$. Then, for $N = 1, 2, \cdots$, there exists a code with parameters (M, N, S, λ) where

$$M \geqq 2^{\rho N} \quad \textit{and} \quad \lambda \leqq Ae^{-BN}, \quad A, B > 0.$$

III. PROOF OF THEOREM 1

3.1 Converse

Let $\{(\mathbf{x}_i, B_i)\}_1^M$ be a code with parameters (M, N, S, λ) for the channel of (1) with $H = H_1(\theta)$ and $R = R_1(\theta)$. Let $\mathbf{x}$ be the random sequence which results when $\mathbf{x} = \mathbf{x}_i$ with probability $1/M$ $(1 \leqq i \leqq M)$. Let $\mathbf{y} = \mathcal{H}\mathbf{x} + \mathbf{z}$ be the corresponding output sequence, and $\mathbf{y}^{(N)} = (y(0)^t, \cdots, y(N-1)^t)^t$. The theorem will follow in the standard way from the Fano inequality (see, for example, Ref. 1) if we can show that

$$I\{\mathbf{x}, \mathbf{y}^{(N)}\} \leqq NC_S. \tag{21}$$

But

$$I\{\mathbf{x}, \mathbf{y}^{(N)}\} \leqq I\{\mathbf{x}, \mathbf{y}\} = I\{\mathbf{x}, \mathcal{H}^{-1}\mathbf{y}\} = I\{\mathbf{x}, \mathbf{x} + \hat{\mathbf{z}}\}, \tag{22}$$

where $\hat{\mathbf{z}} = \mathcal{H}^{-1}\mathbf{z}$ is a stationary Gaussian random process with power spectrum $\Gamma(\theta) = H_1^{-1}(\theta)R_1(\theta)H_1^{-*}(\theta)$. Thus, we can apply Lemma 3 (since $\mathbf{x}$ satisfies the required hypotheses) with $R_2(\theta) = \Gamma(\theta)$ to obtain (21). Hence, the theorem follows.

3.2 Direct-half

Consider again the special case of our channel where $H(\theta) = I_s$ and $R(\theta) = R_2(\theta)$. The idea behind the proof is to construct codes for the general case (H, R arbitrary) by modifying codes (whose existence is guaranteed by Lemma 4) which are known to be good for this special case. We proceed as follows. Let $\{(\mathbf{x}_i, B_i)\}_1^M$ be a code for this channel with parameters $N = N_1$ and $S = S_1$. Then, for $1 \leqq i \leqq M$, we have

$$\mathbf{y}_i^{(N)} = \mathbf{x}_i^{(N)} + \mathbf{z}^{(N)},$$

where the superscript operation is defined by (11). Let T_N be the whitening filter (discussed in Section II) for which $T_N\mathbf{z}^{(N)} = \mathbf{w}$ and $E\mathbf{w}\mathbf{w}^t = I_{Ns}$. Letting $\mathbf{v}_i = T_N\mathbf{y}_i^{(N)}$ and $\mathbf{u}_i = T_N\mathbf{x}_i^{(N)}$, we have

$$\mathbf{v}_i = \mathbf{u}_i + \mathbf{w}. \tag{23}$$

Let us assume that the $\{B_i\}$ correspond to the minimum distance decoder, i.e., $\mathbf{y}^{(N)} \in B_i$ if $|||\mathbf{v} - \mathbf{u}_i||| < |||\mathbf{v} - \mathbf{u}_j|||$ for all $j \neq i$, where $\mathbf{v} = T_N\mathbf{y}^{(N)}$. Then

$$\begin{aligned} P_{ei} = Pr\{\mathbf{y}_i^{(N)} \notin B_i\} &= Pr \bigcup_{j\neq i} \{|||\mathbf{v}_i - \mathbf{u}_i|||| \geqq |||\mathbf{v}_i - \mathbf{u}_j||||\} \\ &= Pr \bigcup_{j\neq i} \{|||\mathbf{w}||| > |||\mathbf{w} - (\mathbf{u}_j - \mathbf{u}_i)||||\} \\ &= Pr \bigcup_{j\neq i} \{\langle \mathbf{w}, \mathbf{u}_j - \mathbf{u}_i\rangle \geqq \tfrac{1}{2}|||\mathbf{u}_j - \mathbf{u}_i|||^2\}. \end{aligned} \tag{24}$$

Thus, in particular, for all $j \neq i$,

$$\begin{aligned} P_{ei} &\geqq Pr\{\langle \mathbf{w}, \mathbf{u}_j - \mathbf{u}_i\rangle \geqq \tfrac{1}{2}|||\mathbf{u}_j - \mathbf{u}_i|||^2\} \\ &= \Phi_c(\tfrac{1}{2}|||\mathbf{u}_j - \mathbf{u}_i|||) = \Phi_c[\tfrac{1}{2}|||T_N(\mathbf{x}_j^{(N)} - \mathbf{x}_i^{(N)})|||], \end{aligned} \tag{25}$$

where

$$\Phi_c(u) = \frac{1}{\sqrt{2\pi}} \int_u^\infty e^{-u^2/2} du$$

is the complementary error function.

Let $H_1(\theta)$ and $R_1(\theta)$ be arbitrary, and suppose that we are given a code $\{(\mathbf{x}_i, B_i)\}_1^M$ with parameters (M, N, S, λ_1) for use on the special channel with $H = I_s$ and $R(\theta) = R_2 = \Gamma(\theta) = H_1^{-1}R_1H_1^{-*}$. Assume that the B_i corresponds to the minimum distance decoder so that P_{ei} is given by (24). We now construct a new code $\{(\mathbf{x}_i^*, B_i^*)\}_{i=1}^M$ with parameters $N = N_2 = (1 + \delta)N_1$ and $S = S_2 = \alpha^2 S_1/(1 + \delta)$ for use on the general channel with $H_1(\theta)$, $R_1(\theta)$ arbitrary. We set

$$x_i^*(n) = \begin{cases} \alpha x_i(n), & 0 \leqq n \leqq N_1 - 1, \\ 0, & \text{otherwise}, \end{cases} \tag{26}$$

where $\alpha > 1$ and $\delta > 0$ are arbitrary. Note that we have allowed a guard-band or dead-space or width δN_1 following the channel input signal. The decoding sets B_i^* $(1 \leqq i \leqq M)$ are described below.

The channel output is as in (1)

$$\mathbf{y} = \mathcal{H}\mathbf{x} + \mathbf{z},$$

where $\mathbf{x}$ is the channel input, $\mathbf{z}$ is the noise, and $\mathcal{H}$ is the operator corresponding to $H_1(\theta)$. Let $\mathcal{G}_{N_2}$ be the operator defined in (20), and let

$$\begin{aligned} \hat{\mathbf{y}} = \mathcal{G}_{N_2}\mathbf{y} &= \mathcal{G}_{N_2}\mathcal{H}\mathbf{x} + \mathcal{G}_{N_2}\mathbf{z} \\ &= \mathbf{x} + \hat{\mathbf{z}} + \boldsymbol{\xi}_1 + \boldsymbol{\xi}_2, \end{aligned}$$

where $\boldsymbol{\xi}_1 = \mathcal{G}_{N_2}\mathcal{H}\mathbf{x} - \mathbf{x}$, $\hat{\mathbf{z}} = \mathcal{G}\mathbf{z}$, and $\boldsymbol{\xi}_2 = \mathcal{G}_{N_2}\mathbf{z} - \hat{\mathbf{z}}$. Let us note that $\hat{\mathbf{y}}$ is calculable from $\mathbf{y}^{(N_2)} = (y^t(0), \cdots, y^t(N_2 - 1))^t$. Further, the noise $\hat{\mathbf{z}}$ has power spectrum $\Gamma(\theta) = H_1^{-1}(\theta)R_1(\theta)H_1^{-*}(\theta)$. (In fact, if $\boldsymbol{\xi}_1 = \boldsymbol{\xi}_2 = 0$, the channel would be equivalent to the special case, and the direct-half of the coding theorem would follow from Lemma 4. Although this is not the case, of course, we will show that $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are sufficiently small so that Lemma 4 can be applied anyway.) The decoding sets B_i^* are defined by: $\mathbf{y}^{(N_2)} \in B_i^*$ if $\hat{\mathbf{y}}^{(N_1)} \in B_i$, $1 \leqq i \leqq M$.

Letting $\mathbf{y}_i^* = \mathbf{x}_i^* + \hat{\mathbf{z}} + \boldsymbol{\xi}_1 + \boldsymbol{\xi}_2$ $(1 \leqq i \leqq M)$, and letting T_{N_1} be as above, we define

$$\begin{aligned} \mathbf{v}_i^* \triangleq T_{N_1}\mathbf{y}_i^{*(N_1)} &= T_{N_1}\mathbf{x}_i^{*(N_1)} + T_{N_1}\hat{\mathbf{z}}^{(N_1)} + T_{N_1}\boldsymbol{\xi}_1^{(N_1)} + T_{N_1}\boldsymbol{\xi}_2^{(N_1)} \\ &= \alpha\mathbf{u}_i + \mathbf{w} + \boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_2, \end{aligned} \tag{27}$$

where $\mathbf{u}_i$ and $\mathbf{w}$ are exactly as in (23) and $\boldsymbol{\gamma}_i = T_{N_1}\boldsymbol{\xi}_i^{(N_1)}$ $(i = 1, 2)$. The decoder for the derived code is the minimum distance decoder for the $\mathbf{v}^*$'s. Now, following the same steps as in (24), we have

$$\begin{aligned} P_{ei}^* \triangleq Pr\{\mathbf{y}_i^{(N_2)} \notin B_i^*\} &= Pr \bigcup_{j \neq i} \{|||\mathbf{v}_i^* - \alpha\mathbf{u}_i||| \geqq |||\mathbf{v}_i^* - \alpha\mathbf{u}_j|||\} \\ &= Pr \bigcup_{j \neq i} \{\langle \mathbf{w} + \boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_2, \mathbf{u}_j - \mathbf{u}_i \rangle \geqq \frac{\alpha}{2}|||\mathbf{u}_j - \mathbf{u}_i|||^2\}. \end{aligned} \tag{28}$$

Now, according to (9), the channel of Section I [with arbitrary $H_1(\theta)$ and $R_1(\theta)$] and the special channel with $H(\theta) = I_s$ and $R(\theta) = H_1(\theta)^{-1}R_1(\theta)H_1(\theta)^{-*} = \Gamma(\theta)$ have the same C_S. Let ϵ, $S > 0$ and ρ $(0 \leqq \rho < C_S)$ given, and let $\{(\mathbf{x}_i, B_i)\}_1^M$ be a code with parameters (M, N_1, S_1, λ_1) (as guaranteed by Lemma 4) such that

$$M \geqq 2^{\rho N_1} \quad \text{and} \quad \lambda_1 \leqq \epsilon.$$

We will show that with N_1 sufficiently large, the derived code has parameter $\lambda \leqq \lambda_1 + \epsilon$. Thus, we will have found a set of codes with parameters (M, N_2, S_2, λ) with

$$S_2 = \frac{\alpha^2 S_1}{(1+\delta)}, \quad M \geqq \exp_2\left\{\frac{\rho N_2}{1+\delta}\right\}$$

and

$$\lambda \leqq 2\epsilon.$$

Since C_S is continuous in its arguments, and α may be chosen arbitrarily close to 1, and δ arbitrarily close to 0, the direct-half of the coding theorem (Theorem 1b) will have been established.

To show that λ for the derived code $\leqq \lambda_1 + \epsilon$, we must show that for each $i = 1, 2, \cdots$,

$$Pr\{\mathbf{y}_i^{*(N_2)} \notin B_i^*\} \leqq Pr\{\mathbf{y}_i^{(N_1)} \notin B_i\} + \epsilon. \tag{29}$$

Inequality (29) will follow directly from the following lemmas, the proofs of which are given at the end of this section.

Lemma 5: Inequality (29) *is satisfied if*

$$Pr\{|||\boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_2||| \leqq \frac{(\alpha - 1)}{2} \min_{i \neq j} |||\mathbf{u}_i - \mathbf{u}_j|||\} \geqq 1 - \epsilon. \tag{30}$$

Lemma 6: For the codes $\{(\mathbf{x}_i, \mathbf{B}_i)\}_1^M$, *as* $N \to \infty$,

$$\min_{i \neq j} |||\mathbf{u}_i - \mathbf{u}_j|||^2 \geqq 0(N_1).$$

Lemma 7: For arbitrary $a > 0$,

$$Pr\{|||\boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_2|||^2 \leqq aN_1\} \to 1, \quad \textit{as} \quad N_1 \to \infty.$$

Now, from Lemmas 6 and 7, condition (30) in Lemma 4 will be satisfied for N_1 sufficiently large. This establishes Theorem 1b.

Proof of Lemma 5: Let

$$S = \left\{|||\boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_2||| \leqq \frac{(\alpha - 1)}{2} \min_{i \neq j} |||\mathbf{u}_i - \mathbf{u}_j|||\right\} \cdot \tag{31}$$

By hypothesis, $Pr\{S\} \geqq 1 - \epsilon$. Since B_i and B_i^* correspond to the minimum distance decoder, we have from (28)

$$Pr\{\mathbf{y}_i^{*(N_2)} \notin B_i^*\} \leqq Pr\{S \cap \{\mathbf{y}_i^{*(N_2)} \notin B_i^*\}\} + Pr\{S^c\}$$

$$\leqq Pr \bigcup_{j \neq i} S \cap \left\{\langle \mathbf{w} + \boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_2, \mathbf{u}_j - \mathbf{u}_i \rangle \geqq \frac{\alpha}{2} |||\mathbf{u}_j - \mathbf{u}_i|||^2\right\} + \epsilon. \quad (32)$$

Now if S occurs,

$$|\langle \boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_2, \mathbf{u}_j - \mathbf{u}_i \rangle| \leqq |||\boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_2||| \cdot |||\mathbf{u}_j - \mathbf{u}_i|||$$
$$\leqq \left(\frac{\alpha - 1}{2}\right) |||\mathbf{u}_j - \mathbf{u}_i|||^2.$$

Thus, the event in the right member of (32) satisfies

$$S \cap \left\{\langle \mathbf{w} + \boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_2, \mathbf{u}_j - \mathbf{u}_i \rangle \geqq \frac{\alpha}{2} |||\mathbf{u}_i - \mathbf{u}_i|||^2\right\}$$
$$\subseteq \left\{\langle \mathbf{w}, \mathbf{u}_j - \mathbf{u}_i \rangle \geqq \frac{\alpha}{2} |||\mathbf{u}_j - \mathbf{u}_i|||^2 - \left(\frac{\alpha - 1}{2}\right) |||\mathbf{u}_j - \mathbf{u}_i|||^2\right\}$$
$$\subseteq \{\langle \mathbf{w}, \mathbf{u}_j - \mathbf{u}_i \rangle \geqq \tfrac{1}{2} |||\mathbf{u}_j - \mathbf{u}_i|||^2\},$$

and (32) becomes

$$Pr\{\mathbf{y}_i^{*(N_2)} \notin B_i^*\} \leqq Pr \bigcup_{j \neq i} \{\langle \mathbf{w}, \mathbf{u}_j - \mathbf{u}_i \rangle \geqq \tfrac{1}{2} |||\mathbf{u}_j - \mathbf{u}_i|||^2\} + \epsilon$$
$$= Pr\{\mathbf{y}_i^{(N_1)} \notin B_i\} + \epsilon,$$

where the last equality follows from (24). This is (29) so that we have proved Lemma 5.

Proof of Lemma 6: For the codes $\{(\mathbf{x}_i, B_i)\}_1^M$, and $N_1 \to \infty$,

$$P_{ei} \leqq Ae^{-BN_1},$$

so that from (25),

$$\Phi_c(\tfrac{1}{2}|||\mathbf{u}_j - \mathbf{u}_i|||) \leqq Ae^{-BN_1}.$$

Since, as $\eta \to \infty$, $\Phi_c(\eta) = e^{-(\eta^2/2)[1+o(1)]}$, we have

$$|||\mathbf{u}_j - \mathbf{u}_i|||^2 \geqq 8BN_1[1 + o(1)],$$

which implies Lemma 6.

Proof of Lemma 7: First note that by (19)

$$|||\boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_2|||$$
$$= |||T_{N_1}(\boldsymbol{\xi}_1^{(N_1)} + \boldsymbol{\xi}_2^{(N_1)})||| \leqq \left[\frac{1}{\min \lambda_{\min}(\theta)}\right]^{\frac{1}{2}} |||\boldsymbol{\xi}_1^{(N_1)} + \boldsymbol{\xi}_2^{(N_1)}|||$$
$$\leqq \left[\frac{1}{\min \lambda_{\min}(\theta)}\right]^{\frac{1}{2}} [|||\boldsymbol{\xi}_1^{(N_1)}||| + |||\boldsymbol{\xi}_2^{(N_1)}|||].$$

Thus, it will suffice to show first, $|||\xi_1^{(N_1)}||| = o(N_1^{\frac{1}{2}})$ as $N_1 \to \infty$, and second, for arbitrary $a > 0$, $Pr\{|||\xi_2^{(N_1)}|||^2 \geqq aN_1\} \to 0$, as $N_1 \to \infty$.

1. Let $\mathbf{x}$ be one of the code vectors in $\{(\mathbf{x}_i^*, B_i^*)\}$. Then $\xi_1 = \mathcal{G}_{N_2}\mathcal{H}\mathbf{x} - \mathbf{x}$, so that for $-\infty < n < \infty$,

$$\xi_1(n) = \sum_{\substack{k<0\\k\geqq N_2}} g(n-k)(\mathcal{H}\mathbf{x})(k). \tag{33}$$

Now, since $\mathbf{x}$ is one of the code vectors $\{\mathbf{x}_i^*\}_1^M$, $x(n) = 0$ for $n \notin [0, N_1 - 1]$. Also since $\mathcal{H}$ is causal, we have $(\mathcal{H}\mathbf{x})(k) = 0$, $k < 0$, and

$$\begin{aligned}\xi_1(n) &= \sum_{k=N_2}^{\infty} g(n-k)(\mathcal{H}\mathbf{x})(k)\\ &= \sum_{k=N_2}^{\infty} g(n-k) \sum_{j=0}^{N_1-1} h(k-j)x(j).\end{aligned} \tag{34}$$

Next, define the sequence ψ by

$$\psi(k) = \begin{cases} \sum_{j=0}^{N_1-1} h(k-j)x(j), & k \geqq N_2. \\ 0, & k < N_2 \end{cases}$$

Then (34) is

$$\xi_1(n) = \sum_{k=-\infty}^{\infty} g(n-k)\psi(k),$$

i.e., $\xi_1 = \mathcal{G}\psi$, and

$$|||\xi_1||| \leqq |\mathcal{G}| \cdot |||\psi|||. \tag{35}$$

Now $|\mathcal{G}| \leqq \sum_{n=-\infty}^{\infty} ||g(n)|| < \infty$ from Theorem 2d, and

$$\begin{aligned}|||\psi|||^2 = \sum_{k=N_2}^{\infty} ||\psi(k)||^2 &= \sum_{k=N_2}^{\infty} \left\| \sum_{j=0}^{N_1-1} h(k-j)x(j) \right\|^2\\ &\leqq \sum_{k=N_2}^{\infty} \left(\sum_{j=0}^{N_1-1} ||h(k-j)||^2 \right)\left(\sum_{j=0}^{N_1-1} ||x(j)||^2 \right),\end{aligned}$$

where in the final step we have used the following form of the Schwarz inequality: if $\mathbf{a}$ is a sequence of $s \times s$ matrices, and $\mathbf{x}$ is a sequence of s-vectors, then

$$||\sum_n a(n)x(n)||^2 \leqq \sum_n ||a(n)||^2 \sum_n ||x(n)||^2.$$

But since $\sum_{j=0}^{N_1-1} \|x^2(j)\|^2 = |||\mathbf{x}|||^2 \leqq \alpha^2 S_1 N_1$, we have

$$|||\psi|||^2 \leqq \alpha^2 S_1 N_1 \sum_{k=N_2}^{\infty} \sum_{j=0}^{N_1-1} \|h(k-j)\|^2. \tag{36}$$

We will now show that, as $N_1 \rightarrow \infty$,

$$\sum_{k=N_2}^{\infty} \sum_{j=0}^{N_1-1} \|h(k-j)\|^2 = \sum_{k=N_2}^{\infty} \sum_{j=0}^{N_1-1} f^2(k-j) \rightarrow 0, \tag{37}$$

where $f(k) = \|h(k)\|$. Expressions (35), (36), and (37) together imply that

$$\|\xi_1{}^{(N_1)}\|^2 \leqq |||\xi_1|||^2 = o(N_1), \quad \text{as} \quad N_1 \rightarrow \infty,$$

which is what we set out to establish. It remains to establish (37).

Now, from (5), $\sum_0^\infty f(k) < \infty$, and $f(k) \leqq B/k$. Setting

$$F(k) = \sum_{j=k}^{\infty} f^2(j),$$

we have

$$\begin{aligned} Q &\triangleq \sum_{k=N_2}^{\infty} \sum_{j=0}^{N_1-1} f^2(k-j) \\ &= \sum_{k=N_2}^{\infty} \sum_{i=k-N_1+1}^{k} f^2(i) = \sum_{k=N_2}^{\infty} [F(k-N_1+1) - F(k)] \\ &= \sum_{k=N_2-N_1+1}^{N_2-1} F(k) \leqq \sum_{\delta N_1+1}^{N_2} F(k). \end{aligned}$$

Now*

$$\sum_{\delta N_1+1}^{N_2} F(k) = kF(k)\Big|_{\delta N_1}^{N_2} + \sum_{\delta N_1+1}^{N_2} (k-1) f^2(k-1).$$

But

$$kF(k)\Big|_{\delta N_1}^{N_2} \leqq N_2 F(\delta N_1) = N_2 \sum_{\delta N_1}^{\infty} f^2(k) \leqq \frac{(1+\delta)}{\delta} \sum_{\delta N_1}^{\infty} k f^2(k),$$

* We have made use of the formula (summation by parts)

$$\sum_a^b v(k) \Delta u(k) = v(k) u(k)\Big|_{a-1}^{b} - \sum_a^b u(k-1) \Delta v(k),$$

where $\Delta u(k) = u(k) - u(k-1)$. Here $v(k) = F(k)$ and $u(k) = k$.

and

$$\sum_{\delta N_1+1}^{N_2} (k-1)f^2(k-1) \leqq \sum_{\delta N_1}^{\infty} kf^2(k).$$

Thus,

$$Q \leqq \left(\frac{1+\delta}{\delta}+1\right)\sum_{\delta N_1}^{\infty} kf^2(k) \leqq \left(\frac{1+2\delta}{\delta}\right)B\sum_{\delta N_1}^{\infty} f(k) \rightarrow 0,$$

since $\sum_0^\infty f(k) < \infty$. Thus, (37) is established and we have finished the first part.

2. Since for any $a > 0$,

$$Pr\{|||\xi_2^{(N_1)}|||^2 \geqq aN_1\} \leqq \frac{E|||\xi_2^{(N_1)}|||^2}{aN_1},$$

and since $N_2 = (1+\delta)N_1$, it suffices to show that

$$E|||\xi_2^{(N_2)}|||^2 = o(N_2), \quad \text{as} \quad N_2 \rightarrow \infty.$$

Now $\xi_2 = \mathcal{G}_{N_2}\mathbf{z} - \hat{\mathbf{z}}$, where $\hat{\mathbf{z}} = \mathcal{H}^{-1}\mathbf{z}$. Hence,

$$E[\xi_2(n)\xi_2^t(n)] = \sum_{\substack{i,j<0\\ i,j\geqq N_2}} g(n-i)r(i-j)g^t(n-j).$$

Let β denote any fixed s-vector and define a sequence ψ of s-vectors by

$$\psi(n-i) = \begin{cases} g^t(n-i)\beta, & i<0 \quad \text{and} \quad i \geqq N_2. \\ 0, & 0 \leqq i < N_2. \end{cases}$$

Then

$$\beta^t\, E[\xi_2(n)\xi_2^t(n)]\,\beta = \sum_{i,j=\infty}^{\infty} \psi^t(n-i)r(i-j)\psi(n-j).$$

Since $r(\cdot)$ is a covariance, $r(k) = r^t(-k)$ for all k. Application of Theorem 2c shows that the double summation above is bounded by

$$\max_{-\pi\leqq\theta\leqq\pi} ||R(\theta)|| \sum_{i=-\infty}^{\infty} ||\psi(n-i)||^2.$$

Since

$$|||\xi_2^{(N_2)}|||^2 = \sum_{n=0}^{N_2-1} \xi_2^t(n)\xi_2(n) = \sum_{n=0}^{N_2-1}\sum_{\nu=1}^{s} e_\nu^t\xi_2(n)\xi_2^t(n)e_\nu,$$

where e_ν is the s-vector with jth entry $\delta_{\nu j}$ $(\nu,\, j = 1, \cdots, s)$, the desired result will follow if we show that

$$\frac{1}{N_2}\beta^t \sum_{n=0}^{N_2-1} E[\xi_2(n)\xi_2^t(n)]\beta = o(1)$$

for any β. Thus, it suffices to show that

$$\frac{1}{N_2}\sum_{n=0}^{N_2-1}\ \sum_{i=-\infty}^{\infty} \|\psi(n-i)\|^2 = o(1).$$

But from the definition of ψ, the last expression can be rewritten as

$$\frac{1}{N_2}\sum_{n=0}^{N_2-1}\ \sum_{k=n+1}^{\infty} [\|\psi(k)\|^2 + \|\psi(-k)\|^2].$$

Since $\mathcal{G}$ is in L, Theorem $2a$ implies ψ is in $l_2^{(s)}$ $(-\infty, \infty)$. In particular, as $n \longrightarrow \infty$

$$\sum_{k=n+1}^{\infty} [\|\psi(k)\|^2 + \|\psi(-k)\|^2] = o(1)$$

and the desired conclusion follows immediately.

IV. ASYMPTOTIC BEHAVIOR AND NUMERICAL EVALUATION OF THE CAPACITY FORMULA

In this final section, we discuss some implications of the channel capacity formula given in (9). As in the prior sections, the channel is assumed to be a multi-input, multi-output channel with memory, and with additive Gaussian noise. But in contrast to the previous case, instead of a discrete time channel, we consider an equivalent continuous time bandlimited channel.

Specifically, the channel inputs or code words (in a T second block coding interval) are vector-valued functions $x(\cdot)$ of dimensions s, bandlimited in frequency interval $[-W, W]$ such that the samples of $x(\cdot)$ satisfy

$$x\left(\frac{n}{2W}\right) = 0, \quad \text{for} \quad \frac{n}{2W} < 0 \quad \text{or} \quad \frac{n}{2W} \geqq T.$$

We also have the average power constraint

$$\int_{-\infty}^{\infty} \|x(t)\|^2 dt \leqq ST,$$

where $\|\cdot\|$ denotes Euclidean norm.

The channel has s inputs and s outputs and has transfer function matrix $H(f)$ for $f \in [-W, W]$. The additive noise is also vector-valued and has power spectral density matrix $R(f)$. Let $\{\lambda_i(f)\}$ denote the set of eigenvalues of $H^{-1}(f)R(f)[H^{-1}(f)]^*$.

The capacity of this channel is determined as follows. Let $S > 0$ be given and let K be the unique positive number satisfying

$$\sum_{i=1}^{s} \int_{-W}^{W} \max[0, K - \lambda_i(f)]df = S. \tag{38a}$$

Then

$$C = \frac{1}{2} \sum_{i=1}^{s} \int_{-W}^{W} \max\left(0, \log_2 \frac{K}{\lambda_i(f)}\right)df \tag{38b}$$

with C in bits per second.

Formula (38) can be obtained from the analogous formula (9) via application of the sampling theorem. The somewhat tedious derivation is carried out for the scalar case ($s = 1$) in the appendix (Section A.3).

We consider several implications of (38). Specifically, for large signal-to-noise ratio, C is linearly related to signal-to-noise ratio; a change in C is proportional to a change in signal-to-noise ratio (in dB). Furthermore, the constant of proportionality depends only on the product sW and is *independent of* any other characteristic of the channel. For small signal-to-noise ratio, C is logarithmically related to signal-to-noise ratio; a change in $\log_{10}C$ is proportional to a change in signal-to-noise ratio (in dB). The constant of proportionality is 0.1 for any channel. In the case in which the channel represents multipair telephone cable with small far-end crosstalk, we show that for large signal-to-noise ratio, C is linearly related to the length of the cable, and for small signal-to-noise ratio, C is logarithmically related to length. Furthermore, the effect of the crosstalk is to reduce the dependence of C on cable length. Finally, we present a numerical evaluation of (38) using realistic parameters obtained from an experimental cable consisting of two twisted pairs of wire.

4.1 Dependence of channel capacity on signal-to-noise ratio

Define a number N_o as

$$N_o = \frac{1}{sW} \int_{-W}^{W} \operatorname{trace} R(f)df.$$

Then N_o represents the noise power per hertz, per dimension, and sWN_o represents the total noise power. We define the following normalized quantities:

$$\begin{aligned} \lambda_i^*(f) &= 2\lambda_i(f)/N_o \\ K^* &= 2K/N_o \\ P &= S/sWN_o \\ P^* &= 10\,(\log_{10} P). \end{aligned}$$

Now P^* is a measure of signal-to-noise ratio in dB. By substituting the above quantities into (38), and using the fact that each λ_i^* is a symmetric function, we have

$$P = \frac{1}{sW} \sum_{i=1}^{s} \int_0^W \max(0, K^* - \lambda_i^*(f))df \tag{39a}$$

and

$$C = \sum_{i=1}^{s} \int_0^W \max\left(0, \log_2 \frac{K^*}{\lambda_i^*(f)}\right)df. \tag{39b}$$

We will determine the asymptotic behavior of C for both very large and very small P. For this purpose we define for every number K^* sets Δ_i, $i = 1, \cdots s$ as

$$\Delta_i = \{f: \lambda_i^*(f) \leqq K^*; f \geqq 0\}.$$

Let δ_i be the measure of Δ_i and define $\delta = (1/s) \sum \delta_i$. In addition, we require the definition of two average channel characteristics, $\bar{\lambda}$ and $\overline{\log \lambda}$. Let

$$\bar{\lambda} = \frac{1}{s\delta} \sum_{i=1}^{s} \int_{\Delta_i} \lambda_i^*(f)df,$$

and

$$\overline{\log \lambda} = \frac{1}{s\delta} \sum_{i=1}^{s} \int_{\Delta_i} \log_2 \lambda_i^*(f)df.$$

Note that δ, $\bar{\lambda}$ and $\overline{\log \lambda}$ are all functions of P. Let $\lambda_{\min} = \min\{\lambda_i^*(f): 0 \leqq f \leqq W; 1 \leqq i \leqq s\}$. Recall from Section I that $\lambda_{\min} > 0$.

Now, from (39),

$$\frac{WP}{\delta} = K^* - \bar{\lambda},$$

and

$$\frac{C}{s\delta} = \log_2 K^* - \overline{\log \lambda}.$$

These equations combine to yield

$$\frac{C}{s\delta} = \log_2\left(\frac{WP}{\delta\bar{\lambda}} + 1\right) + \log_2 \bar{\lambda} - \overline{\log \lambda}. \tag{40a}$$

We investigate (40a) for large P. Assume all the λ_i^*'s are bounded. (Actually, boundedness follows from the hypotheses in Section I.) From (39a), K^* is an increasing function of P. Let P be sufficiently large so that $\lambda_i(f) \leqq K^*$ for all $f \in [0, W]$ and $i = 1, \cdots s$. Then $\delta = W$ and (40a) yields

$$\frac{C}{sW} = \log_2\left(\frac{P}{\bar{\lambda}} + 1\right) + (\log_2 \bar{\lambda} - \overline{\log \lambda}). \tag{40b}$$

Note that for given $\bar{\lambda}$, s, W, and P, C is minimized when all the λ_i^*'s are equal and constant. Now, for $P \gg \bar{\lambda}$, we have from (40b),

$$C \approx sW(\log_2 P - \overline{\log \lambda}),$$

or

$$C \approx 0.3322\, sWP^* - sW\,\overline{\log \lambda}. \tag{41}$$

Now (41) represents a line in the $C - P^*$ plane with intercept $-sW\,\overline{\log \lambda}$ and slope 0.3322 sW, and the region of validity of (41) is $P \gg \bar{\lambda}$. Note that the slope is independent of the λ_i^*'s; the intercept and region of validity are determined by the average channel characteristics $\bar{\lambda}$ and $\overline{\log \lambda}$ evaluated over the whole interval $[0, W]$.

We now investigate (40a) for small P. Observe that $(WP/\delta\bar{\lambda}) + 1 = K^*/\bar{\lambda}$, and as P approaches zero, both K^* and $\bar{\lambda}$ approach $\lambda_{\min}$. Hence, $WP/\delta\bar{\lambda} \rightarrow 0$, as $P \rightarrow 0$. Then (40a) is approximately

$$\frac{C}{s\delta} \approx \frac{WP}{\delta\bar{\lambda}} \log_2 e + \log_2 \bar{\lambda} - \overline{\log \lambda},$$

which can be rewritten as

$$\frac{C}{sW} \approx \left[\log_2 e + \frac{\bar{\lambda}(\log_2 \bar{\lambda} - \overline{\log \lambda})}{K^* - \bar{\lambda}}\right] \frac{P}{\bar{\lambda}}. \tag{40c}$$

We show in appendix Section A.4 that for any channel characteristic with $\lambda_{\min} > 0$,

$$\lim_{P \rightarrow 0} \frac{\log_2 \bar{\lambda} - \overline{\log \lambda}}{K^* - \bar{\lambda}} = 0. \tag{42}$$

Hence, for small P,

$$C \approx \frac{(\log_2 e) sWP}{\lambda_{\min}},$$

or in logarithmic terms,

$$\log_{10} C \approx \frac{P^*}{10} + \log_{10}(sW \log_2 e) - \log_{10} \lambda_{\min}. \tag{43}$$

Now (43) represents a line in the $\log_{10} C - P^*$ plane with intercept $\log_{10}(sW \log_2 e) - \log_{10} \lambda_{\min}$, and slope 1/10. However, the region of validity of (43) is difficult to specify because the location of this region depends not just on $\lambda_{\min}$ but on the shape of the channel characteristic in a neighborhood of $\lambda_{\min}$.

4.2 Dependence of channel capacity on cable length

Suppose that the channel characteristic is a function of a length parameter l. Let l_1 and l_2 be two values of l and suppose that P^* is large and the channel capacity vs P^* characteristic is in the linear region for l_1 and l_2. Then, from (41),

$$C(l_2) - C(l_1) \approx sW[\overline{\log \lambda(l_1)} - \overline{\log \lambda(l_2)}],$$

or

$$C(l_2) - C(l_1) \approx \sum_{i=1}^{s} \int_0^W \log_2 \left(\frac{\lambda_i^*(l_1; f)}{\lambda_i^*(l_2; f)} \right) df, \tag{44}$$

where in these relations we have explicitly shown the dependence of λ_i^* on length. If P^* is very small so that (43) is valid, then

$$\log_{10} C(l_2) - \log_{10} C(l_1) \approx \log_{10} \left(\frac{\lambda_{\min}(l_1)}{\lambda_{\min}(l_2)} \right). \tag{45}$$

Now consider a multipair cable of length l, with s twisted pairs, small far-end crosstalk, and additive white noise. We assume that the crosstalk voltage on a single pair due to all disturbers is proportional to $l^{\frac{1}{2}} f$. Assume also that the attenuation on any pair is proportional to $l f^{\frac{1}{2}}$. If the crosstalk is very small, then a reasonable form for λ_i^* is

$$\lambda_i^*(l; f) = \frac{e^{b_i l f^{\frac{1}{2}}}}{1 + c_i l f^2},$$

where b_i and c_i are constants related to attenuation and crosstalk coupling.[4] Define the averages b and c as $b = \sum b_i/s$ and $c = \sum c_i/s$. Now λ_i^* can be expressed as

$$\lambda_i^* = \exp[b_i l f^{\frac{1}{2}} - \ln(1 + c_i f^2 l)],$$

and for small crosstalk, we have $c_i f^2 l \ll 1$ for all i and all f and l in a range of interest. Then we have approximately

$$\lambda_i^* \approx e^{l(b_i f^{\frac{1}{2}} - c_i f^2)},$$

and, from (44),

$$\frac{\Delta C}{\Delta l} \approx - \log_2(e) \sum_{i=1}^{s} \int_0^W (b_i f^{\frac{1}{2}} - c_i f^2) df,$$

which can be evaluated as

$$\frac{\Delta C}{\Delta l} \approx -\ 0.96\ bsW^{\frac{3}{2}}\left(1 - \frac{cW^{\frac{3}{2}}}{2b}\right) \tag{46}$$

(for large P). Thus, C and l are linearly related. Note that the effect of the crosstalk is to reduce $\Delta C/\Delta l$. If $cW^{\frac{3}{2}}/2b \approx 1$, then C is effectively independent of length. It is theoretically possible to have $cW^{\frac{3}{2}}/2b \approx 1$ and yet have $cW^2 l \ll 1$ as required by our analysis. These relations imply that $2W^{\frac{1}{2}}lb \ll 1$ is a necessary condition that very small crosstalk significantly reduce $\Delta C/\Delta l$. However, we expect that for realistic cable parameters, the reduction in $\Delta C/\Delta l$ due to small crosstalk will not be significant.

Now assume that the channel does not pass dc; i.e., the channel characteristic is that given above for $f \in [f_0, f_1]$, a band of strictly positive frequencies, and is infinite for frequencies outside this band. Let $b_k = \min_i b_i$. Then, for small crosstalk,

$$\lambda_{\min} \approx \exp l(b_k f_0^{\frac{1}{2}} - c_k f_0^2),$$

and

$$\log_{10} \lambda_{\min} \approx 0.434 b_k f_0^{\frac{1}{2}}\left(1 - \frac{c_k f_0^{\frac{3}{2}}}{b_k}\right)l.$$

Thus, for small P, we have from (45),

$$\frac{\Delta \log_{10} C}{\Delta l} \approx -\ 0.434 b_k f_0^{\frac{1}{2}}\left(1 - \frac{c_k f_0^{\frac{3}{2}}}{b_k}\right). \tag{47}$$

As in (46), the effect of the crosstalk is to reduce the dependence of channel capacity on cable length.

4.3 Numerical example

We consider a two-twisted-pair cable with white additive noise. The transfer function matrix $H(f)$ is given by

$$H(f) = e^{-\gamma l}\begin{bmatrix} 1 & i2\pi k l^{\frac{1}{2}} f \\ i2\pi k l^{\frac{1}{2}} f & 1 \end{bmatrix}$$

with l in feet and f in hertz and

$$\gamma = a\sqrt{2\pi} f^{\frac{1}{2}} + ib2\pi f.$$

The off-diagonal terms in the matrix $H(f)$ represent far-end crosstalk. This model is an approximate representation of an experimental

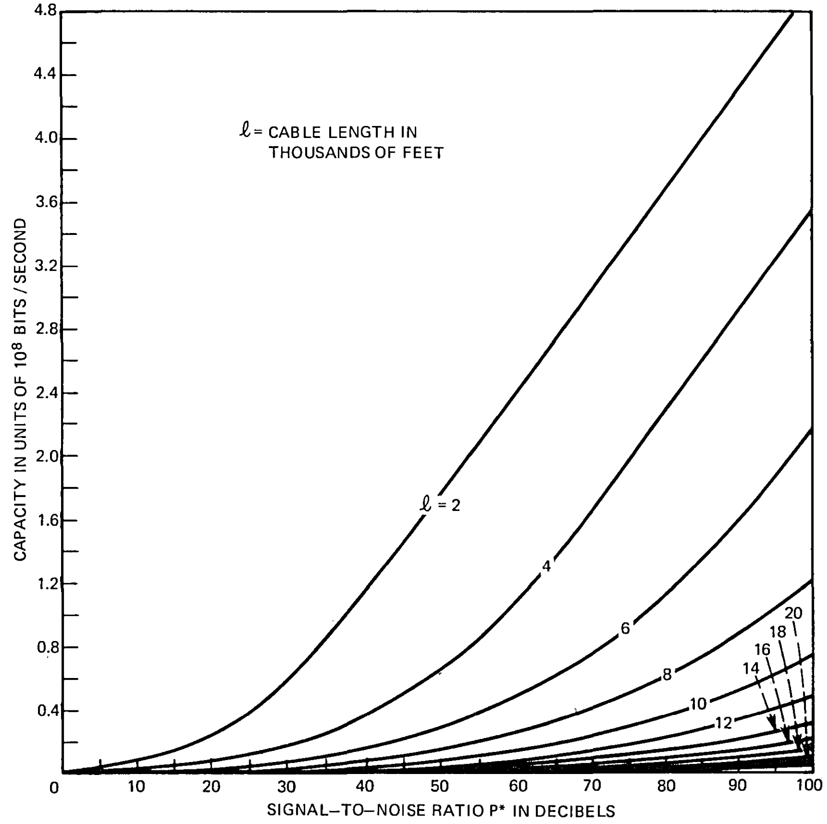


Fig. 1—Channel capacity for experimental cable. Capacity C in units of 10^8 bits per second is plotted as a function of signal-to-noise ratio P^* for various values of cable length l.

two-pair cable. Parameters obtained from measurement are

$$k = 1.26 \times 10^{-12},$$
$$a = 0.23 \times 10^{-6},$$
$$b = 1.48 \times 10^{-9}.$$

This model is valid in the range $10^3 \leqq l \leqq 50 \times 10^3$ feet, and $10^6/2 \leqq f \leqq 10^7$ Hz.

Since the noise is assumed white, the λ_i^*'s are the eigenvalues of $(H^*H)^{-1}$ and are given by

$$\lambda_1^* = \lambda_2^* = \lambda^* = \frac{\exp(2\sqrt{2\pi}\, a f^{\frac{1}{2}})}{1 + (2\pi)^2 f^2 k^2 l}.$$

The capacity equations (39) become

$$\frac{1}{W}\int_{f_0}^{f_1} \max[0, K^* - \lambda^*(f)]df = P \tag{48a}$$

and

$$C = 2\int_{f_0}^{f_1} \max\left(0, \log_2 \frac{K^*}{\lambda^*}\right)df, \tag{48b}$$

where $f_0 = 10^6/2$ and $f_1 = 10^7$.

Numerical evaluation of (48) for various values of P and l has been performed and the results are given in Figs. 1 and 2 and Table I. The figures show C vs P^* for various values of l. The C axis is linear in

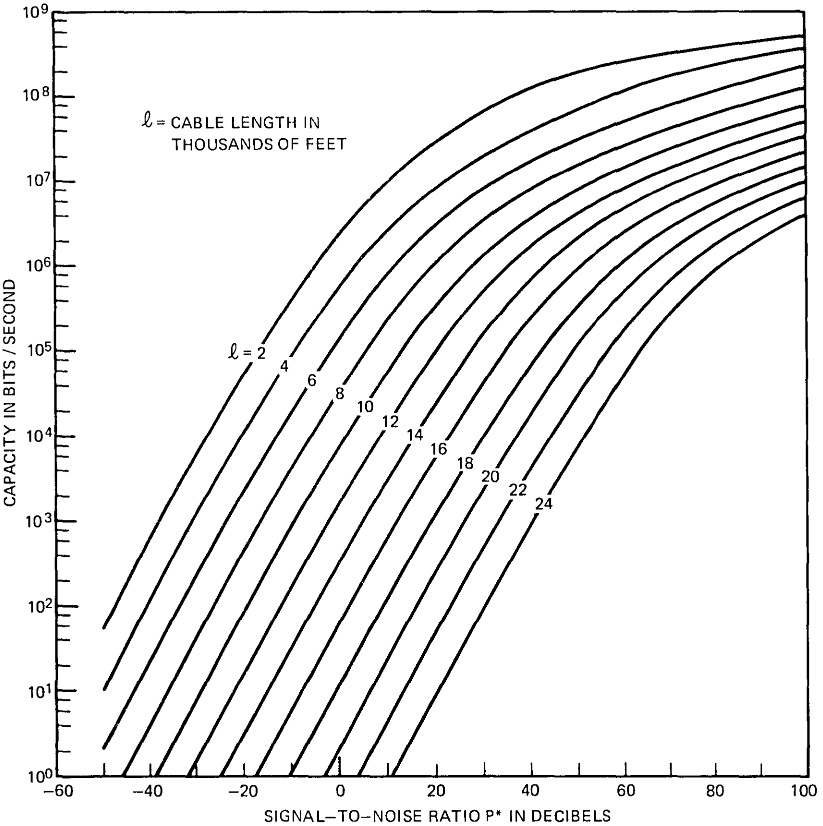


Fig. 2—Channel capacity for experimental cable. Capacity C in bits per second is plotted as a function of signal-to-noise ratio P^* for various values of cable length l. The C axis is logarithmic.

Table I — Values of channel capacity C in bits per second for various values of signal-to-noise ratio P^* and cable length. Exponential notation is employed for C ($aEb \equiv a \times 10^b$).

Signal-to-Noise Ratio P^* in dB	Cable Length l in Thousands of Feet											
	2	4	6	8	10	12	14	16	18	20	22	24
−46	1.33E2	2.60E1	5.07E0	—	—	—	—	—	—	—	—	—
−44	2.10E2	4.12E1	8.18E0	1.57E0	—	—	—	—	—	—	—	—
−42	3.32E2	6.53E1	1.28E1	2.53E0	—	—	—	—	—	—	—	—
−40	5.25E2	1.03E2	2.03E1	3.95E0	—	—	—	—	—	—	—	—
−38	8.29E2	1.63E2	3.21E1	6.29E0	1.21E0	—	—	—	—	—	—	—
−36	1.31E3	2.58E2	5.11E1	9.92E0	1.97E0	—	—	—	—	—	—	—
−34	2.06E3	4.07E2	8.06E1	1.58E1	3.04E0	—	—	—	—	—	—	—
−32	3.23E3	6.43E2	1.27E2	2.51E1	4.94E0	—	—	—	—	—	—	—
−30	5.07E3	1.01E3	2.01E2	3.96E1	7.86E0	1.52E0	—	—	—	—	—	—
−28	7.94E3	1.59E3	3.17E2	6.27E1	1.24E1	2.36E0	—	—	—	—	—	—
−26	1.24E4	2.50E3	4.99E2	9.92E1	1.95E1	3.79E0	—	—	—	—	—	—
−24	1.93E4	3.91E3	7.87E2	1.57E2	3.10E1	6.11E0	1.19E0	—	—	—	—	—
−22	2.98E4	6.09E3	1.26E3	2.47E2	4.91E1	9.66E0	1.89E0	—	—	—	—	—
−20	4.59E4	9.47E3	1.94E3	3.90E2	7.72E1	1.52E1	3.03E0	—	—	—	—	—
−18	7.02E4	1.47E4	3.03E3	6.14E2	1.22E2	2.41E1	4.77E0	—	—	—	—	—
−16	1.07E5	2.26E4	4.72E3	9.63E2	1.92E2	3.82E1	7.57E0	1.47E0	—	—	—	—
−14	1.60E5	3.45E4	7.33E3	1.51E3	3.03E2	6.00E1	1.18E1	2.29E0	—	—	—	—
−12	2.39E5	5.24E4	1.13E4	2.36E3	4.78E2	9.50E1	1.87E1	3.62E0	—	—	—	—
−10	3.53E5	7.88E4	1.74E4	3.68E3	7.50E2	1.51E2	2.98E1	5.87E0	1.11E0	—	—	—
−8	5.14E5	1.17E5	2.65E4	5.70E3	1.18E3	2.36E2	4.73E1	9.16E0	1.77E0	—	—	—
−6	7.40E5	1.73E5	4.01E4	8.80E3	1.84E3	3.72E2	7.42E1	1.48E1	2.88E0	—	—	—
−4	1.05E6	2.51E5	6.01E4	1.35E4	2.87E3	5.87E2	1.17E2	2.30E1	4.45E0	—	—	—
−2	1.47E6	3.60E5	8.91E4	2.06E4	4.45E3	9.18E2	1.84E2	3.67E1	7.32E0	1.35E0	—	—
0	2.03E6	5.10E5	1.31E5	3.11E4	6.87E3	1.44E3	2.91E2	5.97E1	1.15E1	2.25E0	—	—

Table I — continued

Signal-to-Noise Ratio P^* in dB	Cable Length l in Thousands of Feet											
	2	4	6	8	10	12	14	16	18	20	22	24
2	2.76E6	7.10E5	1.89E5	4.66E4	1.05E4	2.24E3	4.58E2	9.12E1	1.81E1	3.56E0	—	—
4	3.71E6	9.74E5	2.69E5	6.90E4	1.61E4	3.48E3	7.19E2	1.44E2	2.88E1	5.61E0	1.11E0	—
6	4.90E6	1.32E6	3.78E5	1.01E5	2.43E4	5.38E3	1.12E3	2.27E2	4.55E1	8.99E0	1.74E0	—
8	6.40E6	1.75E6	5.23E5	1.46E5	3.64E4	8.26E3	1.75E3	3.58E2	7.12E1	1.42E1	2.80E0	—
10	8.24E6	2.29E6	7.11E5	2.07E5	5.39E4	1.26E4	2.73E3	5.62E2	1.13E2	2.24E1	4.32E0	—
12	1.05E7	2.96E6	9.53E5	2.90E5	7.89E4	1.91E4	4.21E3	8.79E2	1.78E2	3.54E1	6.95E0	1.36E0
14	1.32E7	3.78E6	1.26E6	4.00E5	1.14E5	2.86E4	6.48E3	1.37E3	2.80E2	5.54E1	1.09E1	2.20E0
16	1.64E7	4.76E6	1.63E6	5.43E5	1.62E5	4.25E4	9.91E3	2.13E3	4.38E2	8.79E1	1.73E1	3.44E0
18	2.02E7	5.92E6	2.09E6	7.26E5	2.27E5	6.23E4	1.50E4	3.31E3	6.88E2	1.39E2	2.73E1	5.43E0
20	2.46E7	7.27E6	2.64E6	9.53E5	3.14E5	9.01E4	2.26E4	5.09E3	1.07E3	2.19E2	4.35E1	8.49E0
22	2.98E7	8.85E6	3.29E6	1.23E6	4.26E5	1.28E5	3.36E4	7.80E3	1.67E3	3.42E2	6.83E1	1.37E1
24	3.57E7	1.07E7	4.06E6	1.58E6	5.68E5	1.80E5	4.94E4	1.19E4	2.60E3	5.38E2	1.08E2	2.12E1
26	4.25E7	1.27E7	4.95E6	1.98E6	7.47E5	2.49E5	7.16E4	1.79E4	4.00E3	8.41E2	1.71E2	3.40E1
28	5.02E7	1.51E7	5.97E6	2.46E6	9.66E5	3.39E5	1.02E5	2.66E4	6.14E3	1.31E3	2.68E2	5.34E1
30	5.88E7	1.78E7	7.13E6	3.02E6	1.23E6	4.53E5	1.44E5	3.92E4	9.36E3	2.04E3	4.21E2	6.49E1
32	6.86E7	2.07E7	8.44E6	3.67E6	1.55E6	5.96E5	2.00E5	5.71E4	1.41E4	3.15E3	6.59E2	1.33E2
34	7.92E7	2.40E7	9.92E6	4.41E6	1.92E6	7.72E5	2.72E5	8.19E4	2.11E4	4.84E3	1.03E3	2.10E2
36	9.04E7	2.77E7	1.16E7	5.25E6	2.36E6	9.86E5	3.65E5	1.16E5	3.13E4	7.39E3	1.60E3	3.29E2
38	1.02E8	3.18E7	1.34E7	6.19E6	2.86E6	1.24E6	4.82E5	1.61E5	4.57E4	1.12E4	2.48E3	5.15E2
40	1.14E8	3.62E7	1.52E7	7.25E6	3.43E6	1.54E6	6.26E5	2.21E5	6.58E4	1.68E4	3.82E3	8.06E2
42	1.26E8	4.11E7	1.76E7	8.42E6	4.07E6	1.89E6	8.01E5	2.97E5	9.34E4	2.49E4	5.84E3	1.26E3
44	1.39E8	4.64E7	2.01E7	9.71E6	4.79E6	2.29E6	1.01E6	3.93E5	1.31E5	3.66E4	8.88E3	1.95E3
46	1.51E8	5.22E7	2.27E7	1.11E7	5.60E6	2.75E6	1.26E6	5.12E5	1.79E5	5.29E4	1.33E4	3.01E3
48	1.63E8	5.85E7	2.56E7	1.27E7	6.50E6	3.26E6	1.54E6	6.58E5	2.43E5	7.54E4	1.99E4	4.61E3

Table I — continued

Signal-to-Noise Ratio	Cable Length l in Thousands of Feet											
P^* in dB	2	4	6	8	10	12	14	16	18	20	22	24
50	1.76E8	6.52E7	2.88E7	1.44E7	7.48E6	3.84E6	1.87E6	8.32E5	3.23E5	1.06E5	2.93E4	7.02E3
52	1.88E8	7.25E7	3.21E7	1.62E7	8.56E6	4.49E6	2.25E6	1.04E6	4.22E5	1.46E5	4.26E4	1.06E4
54	2.01E8	8.04E7	3.58E7	1.82E7	9.74E6	5.20E6	2.68E6	1.28E6	5.44E5	1.99E5	6.10E4	1.59E4
56	2.14E8	8.88E7	3.97E7	2.04E7	1.10E7	5.99E6	3.15E6	1.55E6	6.90E5	2.66E5	8.61E4	2.35E4
58	2.26E8	9.79E7	4.39E7	2.27E7	1.24E7	6.85E6	3.68E6	1.87E6	8.64E5	3.50E5	1.20E5	3.43E4
60	2.39E8	1.08E8	4.84E7	2.52E7	1.39E7	7.78E6	4.27E6	2.23E6	1.07E6	4.52E5	1.63E5	4.93E4
62	2.51E8	1.18E8	5.31E7	2.79E7	1.55E7	8.80E6	4.91E6	2.63E6	1.30E6	5.76E5	2.20E5	7.00E4
64	2.64E8	1.29E8	5.82E7	3.07E7	1.73E7	9.90E6	5.62E6	3.07E6	1 57E6	7.24E5	2.90E5	9.78E4
66	2.76E8	1.40E8	6.37E7	3.37E7	1.91E7	1.10E7	6.35E6	3.56E6	1.87E6	8.98E5	3.77E5	1.34E5
68	2.89E8	1.52E8	6.94E7	3.70E7	2.11E7	1.24E7	7.22E6	4 11E6	2.21E6	1.10E6	4.83E5	1.82E5
70	3.01E8	1.64E8	7.55E7	4.04E7	2.32E7	1.37E7	8.12E6	4.70E6	2.59E6	1.33E6	6.10E5	2.41E5
72	3.14E8	1.76E8	8.19E7	4.40E7	2.55E7	1.52E7	9.09E6	5.35E6	3.02E6	1.59E6	7.59E5	3.17E5
74	3.26E8	1.89E8	8.88E7	4.78E7	2.78E7	1.67E7	1.01E7	6.05E6	3.48E6	1.88E6	9.32E5	4.06E5
76	3.39E8	2.01E8	9.59E7	5.19E7	3.04E7	1.84E7	1.13E7	6.80E6	3.98E6	2.21E6	1.13E6	5.15E5
78	3.52E8	2.14E8	1.04E8	5.62E7	3.30E7	2.02E7	1.24E7	7.62E6	4.54E6	2.58E6	1.36E6	6.44E5
80	3.64E8	2.26E8	1.11E8	6.07E7	3.58E7	2.20E7	1.37E7	8.49E6	5.13E6	2.98E6	1.61E6	7.94E5
82	3.77E8	2.39E8	1.20E8	6.54E7	3.88E7	2.40E7	1.51E7	9.42E6	5.78E6	3.41E6	1.90E6	9.67E5
84	3.89E8	2.51E8	1.29E8	7.04E7	4.19E7	2.61E7	1.65E7	1.04E7	6.48E6	3.89E6	2.22E6	1.17E6
86	4.02E8	2.64E8	1.39E8	7.56E7	4.52E7	2.83E7	1.80E7	1.15E7	7.22E6	4.41E6	2.57E6	1.39E6
88	4.14E8	2.76E8	1.47E8	8.11E7	4.87E7	3.06E7	1.96E7	1.26E7	8.02E6	4.97E6	2.95E6	1.64E6
90	4.27E8	2.89E8	1.58E8	8.68E7	5.23E7	3.30E7	2.13E7	1.38E7	8.87E6	5.57E6	3.37E6	1.92E6
92	4.39E8	3.01E8	1.68E8	9.28E7	5.61E7	3.56E7	2.31E7	1.51E7	9.78E6	6.22E6	3.82E6	2.23E6
94	4.52E8	3.14E8	1.79E8	9.91E7	6.01E7	3.82E7	2.49E7	1.64E7	1.07E7	6.91E6	4.31E6	2.56E6
96	4.64E8	3.27E8	1.91E8	1.06E8	6.42E7	4.10E7	2.69E7	1.78E7	1.18E7	7.65E6	4.84E6	2.93E6
98	4.77E8	3.39E8	2.02E8	1.13E8	6.86E7	4.40E7	2.90E7	1.93E7	1.28E7	8.43E6	5.41E6	3.33E6
100	4.90E8	3.52E8	2.14E8	1.20E8	7.31E7	4.70E7	3.11E7	2.09E7	1.40E7	9.27E6	6.01E6	3.77E6

Fig. 1, and is logarithmic in Fig. 2. The linear regions discussed above are evident in these figures. The asymptotic estimates for large P given in (41) and (46) are $\Delta C/\Delta P^* \approx 6.3 \times 10^6$ (b/s/dB), and $\Delta C/\Delta l \approx -70 \times 10^3$ (b/s/ft). For constant C, $\Delta P^*/\Delta l \approx 11 \times 10^{-3}$ (dB/ft) or about 58 dB/mile. This is the amount of increase of signal-to-noise ratio necessary to maintain a fixed level of C as length is increased. The asymptotic estimates for small P given (43) and (47) are $\Delta \log_{10} C/\Delta P^* \approx 1/10$, and $\Delta \log_{10} C/\Delta l \approx -0.353 \times 10^{-3}$. For constant C, $\Delta P^*/\Delta l \approx 3.53 \times 10^{-3}$(dB/ft), or about 19 dB/mile. These asymptotic estimates are borne out in the numerical evaluation.

V. ACKNOWLEDGMENT

The authors are indebted to M. I. Schwartz for several helpful discussions regarding Section IV of this work and for providing the data used in the numerical results, and to Ms. M. A. Gatto for programming assistance.

APPENDIX

A.1 *Proof of Theorem 2*

a. Let $y(n) = (\mathfrak{F}\mathbf{x})(n) = \sum_k f(n-k)x(k)$, and let $\sum_n \| f(n) = C < \infty$. From the triangle and Schwarz inequalities,

$$\|y(n)\|^2 \leqq \Big(\sum_k \|f(n-k)\| \cdot \|x(k)\|\Big)^2 \leqq C \sum_k \|f(n-k)\| \cdot \|x(k)\|^2.$$

Hence,

$$|||\mathbf{y}|||^2 = \sum_n \|y(n)\|^2 \leqq C \sum_n \sum_k \|f(n-k)\| \cdot \|x(k)\|^2 \leqq C^2 |||\mathbf{x}|||^2,$$

and $|\mathfrak{F}| \leqq C$.

b. From Parseval's theorem,

$$\frac{|||\mathfrak{F}\mathbf{x}|||^2}{|||\mathbf{x}|||^2} = \frac{\int_{-\pi}^{\pi} \|F(\theta)X(\theta)\|^2 d\theta}{\int_{-\pi}^{\pi} \|X(\theta)\|^2 d\theta} \leqq \max_\theta \|F(\theta)\|^2.$$

c. If $\mathfrak{F}$ is self-adjoint, then

$$|\langle \mathbf{x}, \mathfrak{F}\mathbf{x}\rangle| \leqq |\mathfrak{F}| \cdot |||\mathbf{x}|||^2 \leqq \max_{-\pi \leqq \theta \leqq \pi} \|F(\theta)\| \cdot |||\mathbf{x}|||^2.$$

d. The essence of this result is a matricized version of Wiener's well-known theorem on the reciprocal of an absolutely convergent Fourier series (see Ref. 5, p. 430).

Suppose $\det F(\theta) \neq 0$. We show that $\mathfrak{F}$ has a bounded inverse in L. Now $\det F(\theta)$ is a scalar function consisting of a sum of products of functions each possessing an absolutely convergent Fourier series.* Hence, $\det F(\theta)$ and, by Wiener's theorem, $[\det F(\theta)]^{-1}$ have absolutely convergent Fourier series. Each element of $F^{-1}(\theta)$ is the ratio of a minor determinant to $\det F(\theta)$ so each element has an absolutely convergent Fourier series. Hence, $F^{-1}(\theta)$ has a Fourier series $\sum_n g(n)e^{in\theta}$ with $\sum_n \|g(n)\| < \infty$. Consequently, $\mathfrak{F}$ has a bounded inverse $\mathfrak{F}^{-1}$ in L.

Conversely, let $\mathfrak{F}$ have a bounded inverse $\mathfrak{F}^{-1}$. Then there exists an $\alpha > 0$ such that for all $\mathbf{x}$ in $l_2^{(s)}$ $(-\infty, \infty)$, $|||\mathfrak{F}\mathbf{x}||| \geqq \alpha|||\mathbf{x}|||$. Let $X(\theta) = \sum_n x(n)e^{in\theta}$. From Parseval's theorem,

$$0 < \alpha^2 \leqq \inf_{X(\theta)} \frac{\int_{-\pi}^{\pi} \|F(\theta)X(\theta)\|^2 d\theta}{\int_{-\pi}^{\pi} \|X(\theta)\|^2 d\theta},$$

which implies, since $F(\theta)$ is continuous, that $F(\theta)$ is one-to-one; i.e., $\det F(\theta) \neq 0$ for $-\pi \leqq \theta \leqq \pi$. This completes the proof of Theorem 2.

There are other interesting properties of the class L, which are not directly relevant to the main results of this paper. For the sake of completeness, we mention two generalizations of Theorem 2d, that also have well-known scalar counterparts:

(*i*) Let $\mathfrak{F} \in L$ and let σ $(\mathfrak{F})$ denote the set of all eigenvalues of $F(\theta)$, $-\pi \leqq \theta \leqq \pi$. Let I be the identity on $l_2^{(s)}$. For λ any complex number, $\lambda I - \mathfrak{F}$ has a bounded inverse in L if and only if $\lambda \notin \sigma(\mathfrak{F})$.

(*ii*) Let $g(\cdot)$ be any function analytic in a neighborhood containing $\sigma(\mathfrak{F})$. Then there is an operator $g(\mathfrak{F})$ in L which has as its transfer matrix the function $g[F(\theta)]$.

A.2 *Lemmas 3 and 4*

These lemmas apply to the special case where $H(\theta) = I_s$, and $R(\theta) = R_2(\theta)$. Let $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ be the channel input, output, and noise sequences respectively, and let $\mathbf{x}^{(N)}$, $\mathbf{y}^{(N)}$, $\mathbf{z}^{(N)}$ be the corresponding finite sequences. Thus,

$$\mathbf{y}^{(N)} = \mathbf{x}^{(N)} + \mathbf{z}^{(N)}.$$

* Note that $\sum_n \|f(n)\| < \infty$ if and only if $\sum_n |f_{ij}(n)| < \infty$ for $1 \leqq i, j \leqq s$.

Letting T_N be the whitening matrix defined in (18), we have

$$\mathbf{v} \triangleq T_N \mathbf{y}^{(N)} = T_N \mathbf{x}^{(N)} + \mathbf{w},$$

where $E\mathbf{w}\mathbf{w}^t = I_{N\cdot s}$. This is precisely the discrete-time version of the problem treated in Chapter 8 of Gallager.[1] The results obtained there apply here exactly when we use, instead of his Lemma 8.5.2 (the Kac-Murdock-Szego theorem), the following discrete-time version:

Theorem. Let $\{c_i\}$ $i = 0, \pm 1, \cdots$ *be a sequence of* $s \times s$ *matrices such that the* $Ns \times Ns$ *matrix* $C_N = \{c_{i-j}\}$, $i, j = 0, \cdots, N-1$, *is Hermitian, and* $\sum_k \|c_k\| < \infty$. *Let* $v_1^{(N)}, v_2^{(N)}, \cdots, v_{sN}^{(N)}$ *be the eigenvalues of* C_N *(each counted according to its multiplicity) and let* $\lambda_1(\theta), \lambda_2(\theta), \cdots, \lambda_s(\theta)$ *be the eigenvalues of* $C(\theta) \equiv \sum_k c_k e^{ik\theta}$. *Let* $g(\cdot)$ *be any continuous function defined on an interval containing the values* $\{\lambda_k(\theta)\colon -\pi \leqq \theta \leqq \pi, k = 1, 2, \cdots, s\}$. *Then*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{sN} g(v_k^{(N)}) = \frac{1}{2\pi} \sum_{k=1}^{s} \int_{-\pi}^{\pi} g[\lambda_k(\theta)] d\theta.$$

Furthermore, let $D_N(x) = 1/N$ *(number of eigenvalues* $v_k^{(N)} \leqq x$*). Then*

$$\lim_{N\to\infty} D_N(x) = \frac{1}{2\pi} \sum_{k=1}^{s} \int_{\lambda_k(\theta) \leqq x} d\theta.$$

In the scalar case, $s = 1$, this theorem represents well-known results,[6] a simple account of which can also be found in Ref. 7. The validity of the theorem for $s > 1$ follows on verifying that the arguments employed in Ref. 7 are valid in general for $s \geqq 1$.

A.3 *Derivation of (38)*

We show how to obtain the capacity formula given in (38). We will do this for the scalar case; the result for vectors follows similarly. The capacity formula justified in Theorem 1 can be stated as follows.

The channel input and output are sequences $\mathbf{x} = \{x_n\}_{-\infty}^{\infty}$ and $\mathbf{y} = \{y_n\}_{-\infty}^{\infty}$ respectively, related by

$$y_n = \sum_{k=-\infty}^{\infty} h_{n-k} x_k + z_n, \tag{49}$$

where $\mathbf{h} = \{h_n\}_{n=-\infty}^{\infty}$ is a fixed sequence and $\mathbf{z} = \{z_n\}_{-\infty}^{\infty}$ is a stationary sequence of Gaussian random variables for which $Ez_n = 0$,

$$Ez_m z_{m-n} = \hat{r}_n, \quad -\infty < n, m < \infty. \tag{50}$$

We write (49) symbolically as

$$\mathbf{y} = \mathbf{h}*\mathbf{x} + \mathbf{z}, \tag{51}$$

where "*" denotes vector convolution. The capacity of this channel is given as follows.

For codes of block length N, say that the code vectors $\mathbf{x}$ must satisfy

$$x_n = 0, \quad n \notin [0, N-1]. \tag{52a}$$

$$\frac{1}{N}\sum_{n=0}^{N-1} x_n^2 \leqq S_D. \tag{52b}$$

The capacity is then given by

$$C = \frac{1}{4W}\int_{-W}^{W} df \max\left(0, \log_2 \frac{K|H_D(f)|^2}{\hat{R}_D(f)}\right), \tag{53a}$$

where

$$H_D(f) = \sum_{n=-\infty}^{\infty} h_n e^{(i\pi/W)nf}, \quad |f| \leqq W, \tag{53b}$$

and

$$\hat{R}_D(f) = \sum_{n=-\infty}^{\infty} \hat{r}_n e^{(i\pi/W)nf}, \quad |f| \leqq W, \tag{53c}$$

are the discrete Fourier transforms of $\{h_n\}$ and $\{\hat{r}_n\}$ respectively, and K is the unique solution of

$$S_D = \frac{1}{2W}\int_{-W}^{W} df \max\left(0, K - \frac{\hat{R}_D(f)}{|H_D(f)|^2}\right). \tag{53d}$$

A.3.1 *Some facts about band-limited functions*

Before discussing the continuous-time channel, we digress to mention several facts about band-limited functions.

We denote time functions by lower-case letters, e.g., $u(t)$, and the corresponding Fourier transform by upper-case letters, e.g., $U(f)$. Thus

$$U(f) = \int_{-\infty}^{\infty} u(t)e^{i2\pi ft}dt, \tag{54a}$$

and

$$u(t) = \int_{-\infty}^{\infty} U(f)e^{-i2\pi ft}df. \tag{54b}$$

We shall assume that all functions are square-integrable, and all integrals and infinite sums are limits in the mean. We say that u is band-limited to W Hz if $U(f) = 0$, $|f| > W$. Let

$$g_n(t) = \frac{\sin 2\pi W\left(t - \dfrac{n}{2W}\right)}{\pi\left(t - \dfrac{n}{2W}\right)}, \quad n = 0, \pm 1, \pm 2, \cdots, \tag{55}$$

be the sampling functions. Note that for k, $n = 0, \pm 1, \cdots$,

$$g_n\left(\frac{k}{2W}\right) = \begin{cases} 0 & k \neq n, \\ 2W & k = n, \end{cases} \tag{56a}$$

and

$$G_n(f) = \int_{-\infty}^{\infty} g_n(t) e^{i2\pi f t} dt = \begin{cases} e^{(in\pi/W)f} & |f| < W \\ 0 & |f| > W \end{cases} \tag{56b}$$

(so that g_n is band-limited), and for k, $n = 0, \pm 1, \cdots$

$$\begin{aligned} \int_{-\infty}^{\infty} g_n(t) g_k(t) dt &= \int_{-\infty}^{\infty} G_n(f) G_k^*(f) df \\ &= \int_{-W}^{W} e^{(i\pi f/W)(n-k)} df = \begin{cases} 0, & n \neq k, \\ 2W, & n = k. \end{cases} \end{aligned} \tag{56c}$$

Further, the well-known Sampling Theorem implies that any band-limited function, $u(t)$, can be written as

$$u(t) = \sum_{n=-\infty}^{\infty} u_n g_n(t), \tag{57a}$$

where

$$u_n = \frac{1}{2W} u\left(\frac{n}{2W}\right). \tag{57b}$$

Further, from (57), and (56c),

$$\int_{-\infty}^{\infty} u^2(t) dt = 2W \sum_{n=-\infty}^{\infty} u_n^2. \tag{57c}$$

Let $u(t) = \sum u_n g_n(t)$ and $v(t) = \sum v_n g_n(t)$ be band-limited functions. Then their convolution is

$$w(t) = \int_{-\infty}^{\infty} u(t - \lambda) v(t) dt = \sum w_n g_n(t), \tag{58a}$$

where

$$w_n = \sum_{m=-\infty}^{\infty} u_{n-m}v_m, \quad -\infty < n < \infty, \tag{58b}$$

i.e., $\mathbf{w} = \mathbf{u}*\mathbf{v}$.

Finally, let $z(t)$ $(-\infty < t < \infty)$ be a random process with $Ez(t) = 0$, and covariance

$$Ez(t)z(t-\tau) = r(\tau) \quad -\infty < t, \tau < \infty.$$

Let

$$R(f) = \int_{-\infty}^{\infty} r(t)e^{i2\pi ft}dt$$

satisfy $R(f) = 0$, $|f| > W$, so that $r(t)$ is band-limited. Then, from (57),

$$r(t) = \sum_n \frac{1}{2W} r\left(\frac{n}{2W}\right) g_n(t),$$

and from (56b),

$$R(f) = \sum_n \frac{1}{2W} r\left(\frac{n}{2W}\right) G_n(f) = \frac{1}{2W}\sum_n r\left(\frac{n}{2W}\right) e^{(i\pi/W)nf}. \tag{59}$$

Further, the random process $z(t)$ is a band-limited function so that by (57) we can write

$$z(t) = \sum_n z_n g_n(t) = \sum_n \frac{1}{2W} z\left(\frac{n}{2W}\right) g_n(t).$$

Thus,

$$\hat{r}_n \triangleq Ez_m z_{m-n} = \frac{1}{(2W)^2} Ez\left(\frac{m}{2W}\right) z\left(\frac{m-n}{2W}\right) = \frac{1}{(2W)^2} r\left(\frac{n}{2W}\right).$$

Thus, the discrete Fourier transform of $\{\hat{r}_m\}$ is, using (59),

$$\hat{R}_D(f) = \sum_n \hat{r}_n e^{(i\pi/W)nf} = \frac{1}{(2W)^2}\sum_n r\left(\frac{n}{2W}\right) e^{(i\pi/W)nf} = \frac{1}{2W} R(f). \tag{60}$$

A.3.2 *The continuous-time channel*

The continuous-time channel is defined as follows. The channel input and output are functions $x(t)$ and $y(t)$, respectively, where

$$y(t) = \int_{-\infty}^{\infty} h(t-\lambda)x(\lambda)\, d\lambda + z(t), \quad -\infty < t < \infty, \tag{61}$$

where $h(t)$ is a fixed function and $z(t)$ is a Gaussian random process with covariance as described above. We assume that $x(t)$, $h(t)$, and, therefore, $y(t)$ are band-limited to W Hz. Let us expand x, h, z, and y into series in $g_n(t)$ as in (57). Using (58) we obtain

$$y_n = \sum_{m=-\infty}^{\infty} h_{n-m}x_m + z_m. \tag{62}$$

Since knowledge of the sequences of coefficients $\{x_n\}$, $\{y_n\}$, etc. is equivalent to knowledge of the time functions $x(t)$, $y(t)$, etc., the continuous-time channel is equivalent to the discrete-time channel discussed at the beginning of this appendix. It remains to find the corresponding parameters.

Now the code words (in a T-second block-coding interval) are taken to be band-limited functions $x(t)$ such that the samples $x(n/2W) = 0$, for $(n/2W) < 0$ or $(n/2W) \geqq T$, i.e., $x(n/2W) = 0$, $n \notin [0, N-1]$, where $N = 2WT$. Thus, $x_n = (1/2W)x(n/2W) = 0$, $n \notin [0, N-1]$. The condition

$$\int_{-\infty}^{\infty} x^2(t)dt \leqq ST$$

is, in the light of (57c),

$$\frac{1}{N}\sum_{n=0}^{N-1} x_n^2 \leqq \frac{S}{(2W)^2}. \tag{63}$$

The quantity

$$H_D(f) = \sum_{n=-\infty}^{\infty} h_n e^{(i\pi/W)nf} = \sum_n h_n G_n(f) = H(f),$$

and by (60), $\hat{R}_D(f) = (1/2W)R(f)$. Thus, the continuous-time channel is equivalent to a discrete-time channel with $S_D = S/2W$, $H_D(f) = H(f)$, and $\hat{R}_D(f) = (1/2W)R(f)$. Thus, from (53)

$$C = \frac{1}{4W}\int_{-W}^{W} df \max\left(0, \log_2 \frac{(2WK)|H(f)|^2}{R(f)}\right),$$

where K is the solution to

$$\frac{S}{(2W)^2} = \frac{1}{2W}\int_{-W}^{W} df \max\left(0, K - \frac{1}{2W}\frac{R(f)}{|H(f)|^2}\right).$$

Letting $K^* = 2WK$, we have

$$S = \int_{-W}^{W} df \max\left(0, K^* - \frac{R(f)}{|H(f)|^2}\right)$$

and

$$C = \frac{1}{4W} \int_{-W}^{W} df \max\left(0, \log_2 \frac{K^*|H(f)|^2}{R(f)}\right).$$

The expression for C is in bits per sampling time. To obtain C in bits per second, multiply by $2W$.

A.4 *Proof of equation (42)*

Equation (42) follows at once from the theorem below if the functions λ_i^*, $i = 1, 2, \cdots s$, are replaced by a single function f representing a concatenation of the functions λ_i^*; f will be bounded away from zero since the same is true of each λ_i^*.

Theorem: Let $f(\cdot)$ *be a measurable function on a finite interval and* ess inf $f(x) = f_0 > 0$. *For any* $K > f_0$ *define* $\Delta = \{x; f(x) \leqq K\}$ *and let* δ *be the measure of* Δ. *Define*

$$I_f(K) = \frac{\log \frac{1}{\delta} \int_\Delta f(x)dx - \frac{1}{\delta} \int_\Delta \log f(x)dx}{K - \frac{1}{\delta} \int_\Delta f(x)dx}.$$

Then

$$\lim_{K \to f_0} I_f(K) = 0.$$

Proof: Without loss of generality we can take the log to be the natural log and can assume $f_0 = 1$. Let $f = 1 + g$ and $K = 1 + k$. For each $n \geqq 1$ define

$$\overline{g^n} = \frac{1}{\delta} \int_\Delta g^n(x)dx.$$

For all $x \in \Delta$, $g(x) \leqq k$ and $\bar{g} \leqq k$. For $k < 1$, the log may be expanded in a power series. After some rearrangement of terms, we have

$$I_f(K) = \frac{\sum_{n=1}^{\infty} \left[\frac{1}{2n} (\overline{g^{2n}} - \bar{g}^{2n}) - \frac{1}{2n+1} (\overline{g^{2n+1}} - \bar{g}^{2n+1}) \right]}{k - \bar{g}}.$$

But $\bar{g} \leqq k$ and by Jensen's inequality $\bar{g}^n \leqq \overline{g^n}$ for all $n \geqq 1$. Then

$$I_f(K) \leqq \frac{\sum_{n=1}^{\infty} \frac{1}{2n} (k^{2n} - \bar{g}^{2n})}{k - \bar{g}}.$$

Now for all $n \geqq 1$,

$$k^n - \bar{g}^n = (k - \bar{g}) \sum_{i=0}^{n-1} k^{n-i-1} \bar{g}^i \leqq (k - \bar{g}) n k^{n-1},$$

and

$$I_f(K) \leqq \sum_{n=1}^{\infty} \frac{k^{n-1}}{2} (k^n + \bar{g}^n) \leqq \frac{1}{k} \sum_{n=1}^{\infty} k^{2n} = \frac{k}{1 - k^2},$$

or, since $K = 1 + k$,

$$I_f(K) \leqq \frac{K - 1}{1 - (K - 1)^2},$$

and the result is proved.

REFERENCES

1. R. G. Gallager, *Information Theory and Reliable Communication,* New York: John Wiley, 1968.
2. A. D. Wyner, "On the Intersymbol Interference Problem for the Gaussian Channel," B.S.T.J., *50,* No. 7 (September 1971), pp. 2355–2363.
3. W. Toms and T. Berger, "Capacity and Error Exponents of a Channel Modeled as a Linear Dynamical System," IEEE Trans. Info. Theory, *IT-19,* No. 1 (January 1973), pp. 124–126.
4. M. I. Schwartz, private communication.
5. F. Riesz and B. Sz.-Nagy, *Functional Analysis,* New York: Frederick Ungar, 1955.
6. U. Grenander and G. Szego, *Toeplitz Forms and their Applications,* Berkeley: University of California, 1958.
7. R. M. Gray, "On the Asymptotic Eigenvalue Distribution of Toeplitz Matrices," IEEE Trans. Info. Theory, *IT-18,* No. 6 (November 1972), pp. 725–730.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 5, May–June 1974
Printed in U.S.A.

Experimental Results on a Single-Material Optical Fiber

By R. D. STANDLEY and W. S. HOLDEN

(Manuscript received November 20, 1973)

Experiments confirm a number of theoretical predictions regarding the behavior of single-material optical fibers. In particular, the predicted modal velocity spread (10's of ns/km) and the numerical aperture are found to be in good agreement with theory.

I. INTRODUCTION

The concept of optical fibers made from a single material was the subject of a recent article;[1] the advantages of such fibers are their construction of low-loss fused silica and their freedom from the problems associated with glass interfaces. In this paper, we present experimental results on the dispersion observed in a *particular* single-material fiber. The significance of this work lies in the good agreement between the theory and the experiment; the results do not represent a careful evaluation of single-material fiber as a transmission medium.

The results of the experiments are as follows:

(*i*) The predicted modal velocity spread was confirmed.
(*ii*) The measured numerical aperture is directly proportional to wavelength, as predicted by theory.
(*iii*) There was very little mode coupling between lowest order modes for lengths up to 100 meters.
(*iv*) Penetration of energy into the support structure was small (the decay constant is about 20 dB/μm).

II. THEORY OF MODAL VELOCITY SPREAD

The theory of single-material fibers is summarized in Ref. 1. In the present discussion, we concentrate on the modal dispersion of such fibers. Given the structure shown in Fig. 1, we assumed that the electromagnetic fields vary either sinusoidally or exponentially along

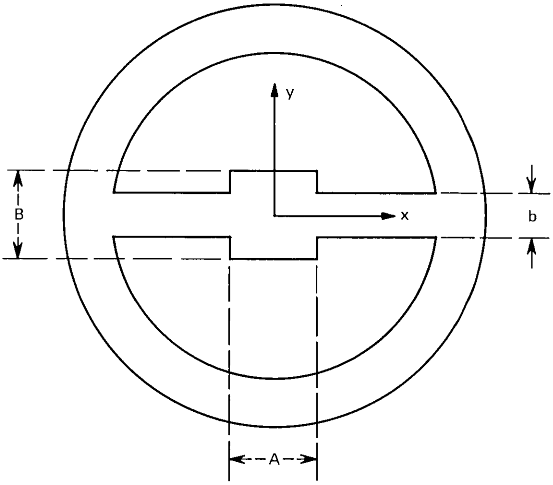


Fig. 1—Idealized fiber cross section.

x, y, and z. We further assume the modes of interest to be far above cutoff so that the transverse wave numbers can be considered to be constant; we approximate them as

$$k_x = \frac{\pi\mu}{A} \tag{1}$$

and

$$k_y = \frac{\pi\nu}{B}. \tag{2}$$

A and B are defined in Fig. 1 and μ and ν are the mode orders. It then can be shown that the group velocity of the mode of order $\mu\nu$ is approximately

$$V_{g_{\mu\nu}} = \frac{c}{n}\left\{1 - \frac{\lambda^2}{8n^2}\left[\left(\frac{\mu}{A}\right)^2 + \left(\frac{\nu}{B}\right)^2\right]\right\}. \tag{3}$$

This gives for the expected time spread between the fastest and the $\mu\nu$ pulses

$$\Delta\tau_{\mu\nu} \simeq \frac{L\lambda^2}{8cn}\left[\left(\frac{\mu}{A}\right)^2 + \left(\frac{\nu}{B}\right)^2\right], \tag{4}$$

where L is the fiber length. From Ref. 1, the maximum value of the bracketed term in eq. 4 is b^{-2}; b is defined in Fig. 1. Thus, the maximum time spread between pulses is

$$\Delta\tau_{\max} = \frac{L}{8cn}\left(\frac{\lambda}{b}\right)^2. \tag{5}$$

III. EXPERIMENTAL RESULTS

3.1 Description of the fibers

The fibers used for our experiments were made of fused silica and had the cross section shown in Fig. 2. Because of the multimode nature of this particular fiber, we anticipate that the theory developed in Section II for the rectangular fiber geometry will still approximately apply.

The following numerical result can serve as a guide in predicting the actual measurement. Assume a fiber with $A = B = 10$ μm, $b = 2$ μm, and $n = 1.46$, and let $\lambda = 1$ μm. Then,

$$\Delta\tau_{\mu\nu} = 2.85L(\mu^2 + \nu^2)ps. \tag{6}$$

For a 100-meter fiber length, the low-order pulses would have time separations of a few tenths of a nanosecond. This implies the following: if such a fiber were excited by a pulse whose width is small compared to $\Delta\tau_{\mu\nu}$, then, assuming little mode coupling, each mode should be

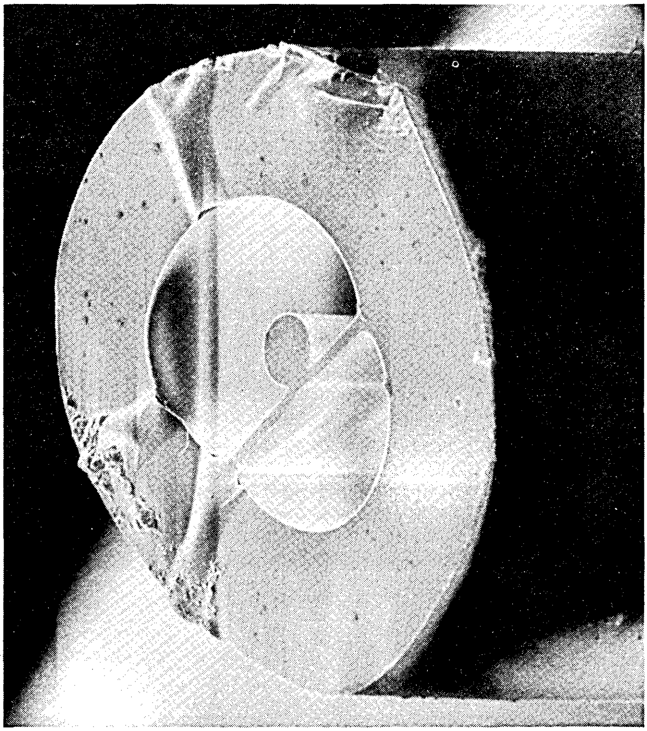


Fig. 2—Photograph of experimental fiber cross section.

observable at the fiber output, providing, of course, that the detection system is sufficiently broad band.

3.2 Dispersion

In the experiments, a mode-locked Nd:YAG laser was used as the source. Using a germanium photodetector, this system yields detector-limited output-pulse widths of less than 200 ps.[2] An early observation was that energy incident in a single input pulse appears at the fiber

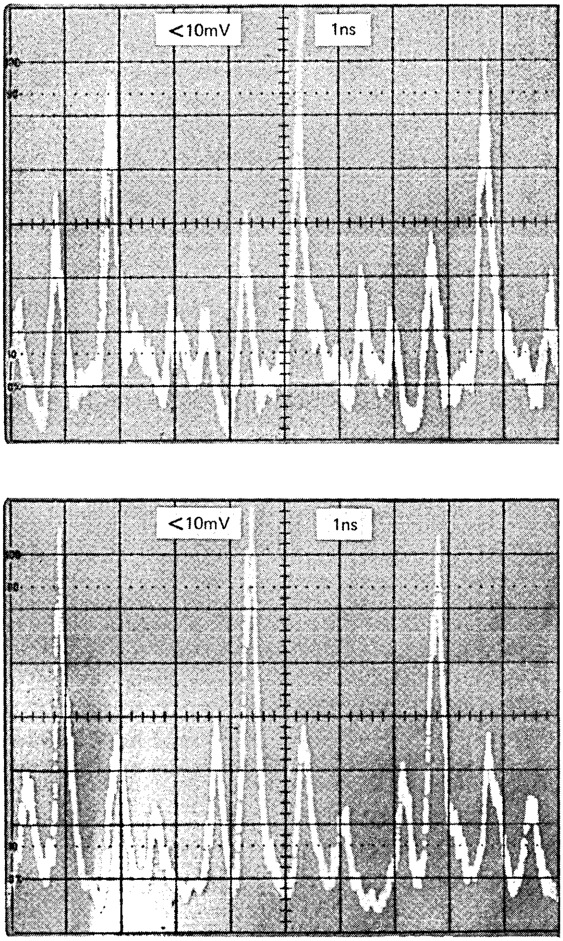


Fig. 3—Fiber output as a function of time and launching conditions.

output as several pulses each representing a mode group. (In this case the fiber was about 90 meters long.) The relative amplitudes of the output pulses could be changed by varying the input launching condition, which is demonstrated by the two photographs shown in Fig. 3.

Additional information was obtained by measuring the velocity difference between the output pulses as a function of fiber length; Fig. 4 shows the results for the first few modes. The results are in reasonable agreement with that predicted by eq. 4; the predicted time spread between the first four modes is 8 ps/m, 17 ps/m, and 21 ps/m, whereas the measured values were 7 ps/m, 16 ps/m, and 22 ps/m. Note that the time difference approaches zero for zero length, which suggests little mode coupling among the lower-order modes as does the previous observation that individual modes could be preferentially excited. Although not shown, the higher-order modes behave differently, having a time difference which appears to be related to the lower order modes. This suggests that the higher-order modes are possibly generated by mode coupling from the lower-order modes and that the higher-order modes suffer more loss.

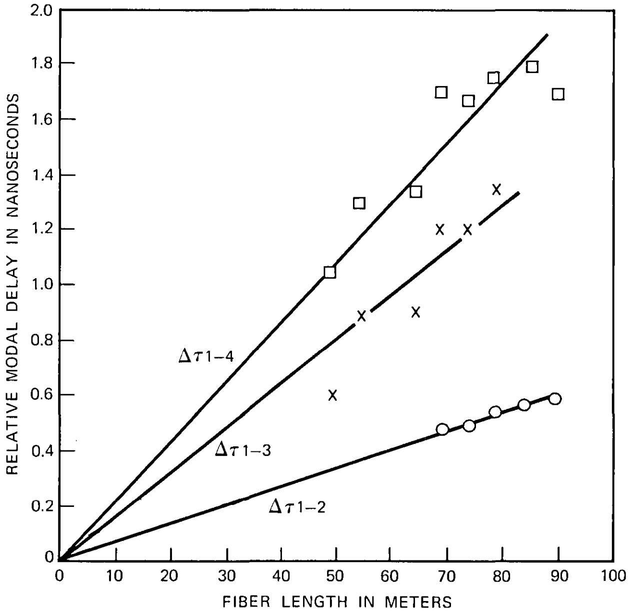


Fig. 4—Relative modal delay as a function of fiber length.

3.3 Numerical aperture

Theory predicts a numerical aperture (half-angle of the fiber radiation cone) of NA $= \lambda/2b$. For our fiber, b was 2 μm so that the theoretical NA at $\lambda = 1.06$ μm was 0.265, and at $\lambda = 0.6328$ μm the NA should be 0.1582. The predictions compare favorably with our measured results, which were 0.27 at $\lambda = 1.06$ μm and 0.16 at $\lambda = 0.6328$ μm as determined from the angular spread of the far field radiation.

IV. SUMMARY

The experimental investigation of the dispersion in a single-material optical fiber showed good agreement with theory. The linear dependence of numerical aperture on wavelength also was verified.

V. ACKNOWLEDGMENTS

The authors wish to thank P. Kaiser who supplied the fibers used in these studies.

REFERENCES

1. P. Kaiser, E. A. J. Marcatili, and S. E. Miller, "A New Optical Fiber," B.S.T.J., *52*, No. 2 (February 1973), pp. 265–269.
2. D. Gloge, E. L. Chinnock, R. D. Standley, and W. S. Holden, "Dispersion in a Low Loss Multimode Fiber Measured at Three Wavelengths," Electronics Letters, *8*, No. 21 (October 1972), pp. 527–529.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 5, May–June 1974
Printed in U.S.A.

Optimal Trade-Off of Mode-Mixing Optical Filtering and Index Difference in Digital Fiber Optic Communication Systems

By S. D. PERSONICK

(Manuscript received December 11, 1973)

In a digital fiber optical communication system, the optical power required at the receiver input to achieve a desired error rate depends upon the shape of the received pulses. In systems employing multimode fibers and/or broadband sources, we can experience pulse spreading in propagation because of the group velocity differences of different modes or because of dispersion. In an effort to control or compensate for pulse spreading, we can trade off coupling efficiency between the light source and the fiber (by varying the core-cladding index difference or bandlimiting the source), scattering loss in the fiber (by introducing mode coupling), and equalization in the receiver at baseband. This paper investigates the optimal trade-off for various fiber-source combinations.

I. INTRODUCTION AND REVIEW OF BACKGROUND MATERIAL

In digital fiber optic communication systems, as in other digital systems, the received power required at a repeater to achieve a desired error rate depends upon the shape of the received pulses. A previous paper[1] showed that the minimum average power requirement results from a pulse that is sufficiently narrow so that its energy spectrum is almost constant for all frequencies passed by the receiver (ideally, an impulse). For other received pulse shapes, we can define the additional power required, in decibels, as a "power penalty" for not having impulse-shaped pulses. Typical calculations of this power penalty for "on-off" signaling and a receiver employing avalanche gain with a high impedance front end[1] are shown for various families of received pulse shapes in Fig. 1. In that figure, the parameter σ/T is defined as follows:

$$\frac{\sigma^2}{T^2} = \frac{1}{T^2}\left\{\frac{1}{A}\int h_p(t)t^2dt - \left[\frac{1}{A}\int h_p(t)tdt\right]^2\right\}, \qquad (1)$$

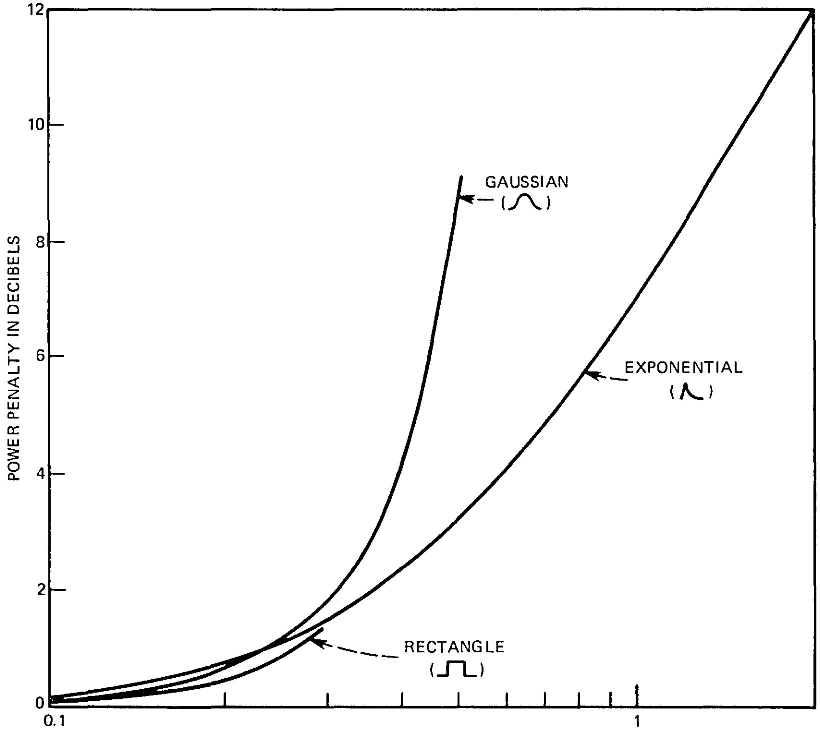


Fig. 1—Typical power penalty vs σ/T.

where

$$T = \text{spacing in time between binary digits}$$
$$h_p(t) = \text{received optical pulse shape}$$
$$A = \text{area under } h_p(t).$$

We shall refer to σ as the rms pulse width.*

It has been shown[2,3] that, in long fibers, the "power impulse response" of the fiber approaches a Gaussian shape. In the rest of this paper, we assume that the received optical pulse is Gaussian in shape and that it has an rms width determined by the fiber delay distortion. That is, we assume that the rms width of the fiber input pulse is sufficiently small so that the power penalty associated with the rms

* In a Gaussian-shaped pulse, the rms width is about 0.425 times the full width between half-amplitude points. In a rectangular pulse, the rms width is $1/\sqrt{12}$ the full width.

sum of that width and the rms fiber impulse response width is the same as the power penalty associated with the rms fiber impulse response width alone.

From various heuristic analyses (see the appendix), we can conclude that the rms width of the received pulse is approximately the rms sum of the delay distortion in the fiber resulting from material dispersion (because of the variation of group velocity with wavelength associated with the use of a broadband source) and the delay distortion associated with the spread in the group delays of various fiber modes (when a multimode fiber is used). That is,

$$\sigma = (\sigma^2_{\text{dispersion}} + \sigma^2_{\text{mode}})^{\frac{1}{2}}, \tag{2}$$

where $\sigma_{\text{dispersion}} \sim$ optical bandwidth·fiber length $= B \cdot L$ and where σ_{mode} is determined as follows[4]

Case 1. Conventional clad multimode fibers without mode coupling.

$$\sigma_{\text{mode}} = 0.289 \frac{\Delta n}{c} L,$$

where

n = index of refraction of the core
Δ = (index of refraction of the core − index of refraction of the cladding)/n
c = speed of light.

Case 2. Conventional clad fibers with complete mode coupling after a distance L_C.

$$\begin{aligned} \sigma_{\text{mode}} &= 0.289 \frac{\Delta n}{c} \sqrt{LL_C} \qquad \text{for} \quad L > L_C, \\ &= 0.289 \frac{\Delta n}{c} L \qquad\qquad \text{for} \quad L \leqq L_C. \end{aligned}$$

Case 3. Ideal graded index fiber without mode coupling.

$$\sigma_{\text{mode}} = 0.037 \frac{\Delta^2 n}{c} L.$$

Case 4. Other fibers[2] can be treated once the techniques outlined below are understood.

From the definitions of $\sigma_{\text{dispersion}}$ and σ_{mode} above, we see that, for a broadband (incoherent) source, the material dispersion contribution to σ can be controlled by limiting the optical bandwidth B being used. However, if the optical source bandwidth, B_o, must be reduced by filtering, then the average power into the guide will be reduced by the factor B/B_o. Similarly, we can control the mode delay spread by reducing the index difference Δ. If a multimode (incoherent) source is being used, the average power into the guide is proportional to Δ. Thus, we trade off input power against mode delay spread. Furthermore, even when we use a coherent source which in principle can be focused into any fiber, we must be careful when Δ becomes significantly smaller than 0.005, since the fiber loss at bends becomes large. (Exactly what value of Δ is too small to be practical is an open question.) For fibers with mode coupling, we can control σ_{mode} by decreasing L_C (increasing the mode coupling). However, this causes the coupling to radiating modes to increase, thus increasing the fiber loss.[5] Here again, there is a trade-off between the average power we receive at the fiber output and the received rms pulse width. For a *fixed shape* of the mechanical spectrum of fiber geometry perturbations, the radiation loss per unit length of the fiber resulting from mode coupling is inversely proportional to L_C, i.e.,

$$\text{radiation loss in nepers} = \alpha_o L/L_C, \tag{3}$$

where

α_o = constant depending upon the *shape* of the mechanical spectrum of the geometry perturbations causing coupling (and possibly upon the index difference Δ). L_C depends upon the *amplitude* of the mechanical perturbations.

In the following sections we derive the optimal trade-off between σ, B, Δ, and L_C for various combinations of sources and fibers to maximize the allowable fiber length L between the optical source and the repeater.

II. ANALYSIS

2.1 *Incoherent source, conventional clad fiber, no mode coupling*

Let the average power into the guide be P_s when the index difference Δ is at some maximum practical value Δ_o and when the full source optical bandwidth B_o is being used. Let the loss of the fiber be α nepers per kilometer. Let the power penalty from the nonzero value of the received rms pulse width, in nepers, be $f(\sigma/T)$. Let the required power at the receiver be P_r when $\sigma/T = 0$. If we use a value of $\Delta \leqq \Delta_o$ and filter the source output to have an optical bandwidth $B \leqq B_o$, then we must have

$$P_s \frac{\Delta}{\Delta_o} \frac{B}{B_o} e^{-\alpha L} \geqq P_r e^{f(\sigma/T)}, \tag{4}$$

where, from (2),

$$\sigma = \left\{\left(C_1 \frac{B}{B_o} L\right)^2 + \left(C_2 \frac{\Delta}{\Delta_o} L\right)^2\right\}^{\frac{1}{2}} = \{\sigma_d^2 + \sigma_m^2\}^{\frac{1}{2}}$$

and C_1, C_2 are constants.

We rewrite (4) as follows (using equality to maximize L),

$$\frac{P_s}{P_r} e^{-\alpha L} = \left\{\frac{\Delta}{\Delta_o} \frac{B}{B_o} e^{-f(\sigma/T)}\right\}^{-1}.$$

To maximize L, we must choose Δ/Δ_o and B/B_o to maximize the term in braces subject to the constraint that these ratios cannot exceed unity. We define -10 log (term in braces) as the "excess loss."

By equating appropriate partial derivatives of the excess loss to zero, we obtain the following equations for optimizing $\hat{\Delta}$ and $\hat{B}$ (Δ and B which minimize excess loss).

$$f'\left(\frac{\sigma}{T}\right)\frac{\sigma_m^2}{\sigma T} = 1 \qquad \text{provided } \hat{\Delta}/\Delta_o \leqq 1, \qquad \text{(5a)}$$

$$\text{otherwise } \hat{\Delta}/\Delta_o = 1,$$

$$f'\left(\frac{\sigma}{T}\right)\frac{\sigma_d^2}{\sigma T} = 1 \qquad \text{provided } \hat{B}/B_o < 1, \qquad \text{(5b)}$$

$$\text{otherwise } \hat{B}/B_o = 1,$$

where $f'(z) = d/dx[f(x)]|_{x=z}$ and σ_m and σ_d are defined in (4). For sufficiently long lengths L, where both $\hat{\Delta}/\Delta_o$ and $\hat{B}/B_o$ are less than unity, we obtain [by adding (5a) to 5(b)]

$$\sigma_m = \sigma_d = \frac{xT}{\sqrt{2}}, \qquad (6)$$

where x is the solution $f'(x)x/2 = 1$. More specifically, we obtain the following

$$C_1 \frac{\hat{B}L}{B_o} = \frac{xT}{\sqrt{2}}, \qquad C_2 \frac{\hat{\Delta}L}{\Delta_o} = \frac{xT}{\sqrt{2}} \qquad (7)$$

and therefore*

$$\text{excess loss} = -10 \log \left[\frac{x^2T^2}{2L^2C_1C_2} e^{-f(x)}\right]$$

for

$$\frac{L}{T} \geqq \text{maximum of} \left\{\frac{x}{\sqrt{2}C_1}, \frac{x}{\sqrt{2}C_2}\right\}.$$

* Throughout this paper we use the parameter L/T (guide length/time slot width) frequently. The larger the fiber length or the smaller the time slot width, the more excess loss must be incurred to control or compensate for pulse spreading.

From the Gaussian power penalty curve of Fig. 1, we obtain the value of x where $f'(x)x = 2$ to be 0.37. At that value of x,

$$-10 \log e^{-f(x)} = 3.3 \text{ dB}.$$

As L/T decreases, either $\hat{\Delta}/\Delta_o$ or $\hat{B}/B_o$ will eventually reach unity. When that happens, σ will approach σ_d or σ_m, respectively, for shorter lengths L. Then, from (5), σ/T will approach the solution of $f'(z)z = 1$. Furthermore, the excess loss will approach either

$$\text{excess loss} \rightarrow -10 \log \left[\frac{zT}{LC_1} e^{-f(z)} \right] \tag{8}$$

if $\hat{\Delta}/\Delta_o$ reaches unity first and $L/T > z/C_1$ or

$$\text{excess loss} \rightarrow -10 \log \left[\frac{zT}{LC_2} e^{-f(z)} \right] \tag{9}$$

if $\hat{B}/B_o$ reaches unity first and $L/T > z/C_2$.

From the Gaussian curve of Fig. 1, the solution of $f'(z)\, z = 1$ is $z = 0.3$. At that value of z, $-10 \log e^{-f(z)} = 1.8$ dB.

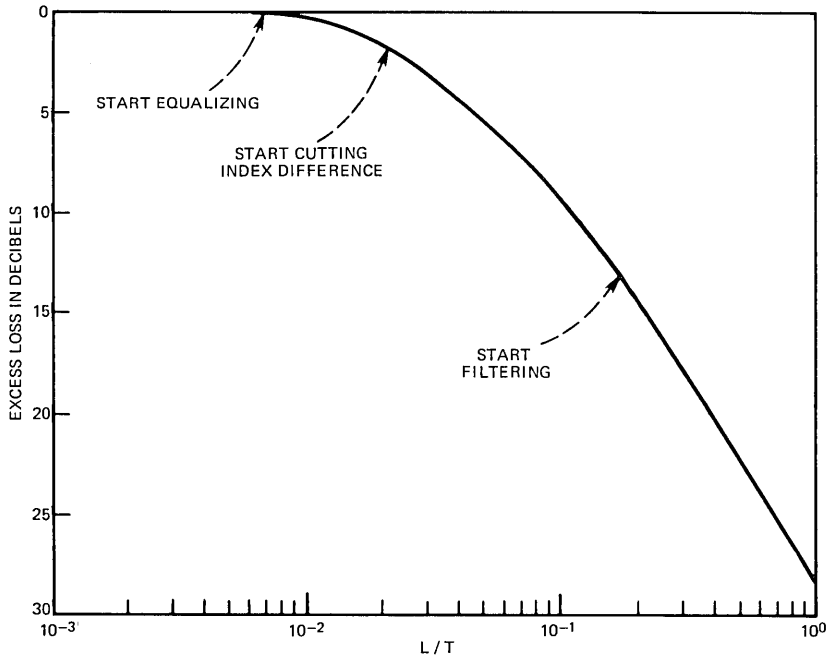


Fig. 2—Excess loss vs L/T for conventional fiber.

For L/T sufficiently small so that both $\hat{\Delta}/\Delta_o$ and $\hat{B}/B_o = 1$, the excess loss will approach zero as L/T decreases.

Example 1

In a fused silica fiber, the material dispersion has been measured at 9 ps/km per angstrom of wavelength difference. With a typical GaAs LED* (light-emitting diode), this results in a value of σ_d of 1.5 ns/km of pulse spreading at the full bandwidth B_o. Thus, C_1 can be set at 1.5 ns/km. For a fiber with a maximum index difference Δ of 0.01, C_2 would be given by 14.5 ns/km.

From (7) we obtain

$$\begin{aligned}\text{excess loss} &= -10 \log \frac{(0.37T)^2}{2L^2(1.5)(14.5)}\, e^{-f(0.37)} \\ &= -20 \log \frac{T}{L} + 28.3 \text{ dB}\end{aligned}$$

for $L/T > 0.174$, where the length L is in kilometers and the time slot width T is in nanoseconds.

From (8) we find that, for $0.0206 < L/T < 0.174$, the excess loss asymptotically approaches

$$\text{excess loss} \rightarrow -10 \log \left[\frac{zT}{LC_2} e^{-f(z)} \right] = -10 \log \frac{T}{L} + 18.6 \text{ dB}$$

for $0.0206 < L/T < 0.174$.

For $L/T < 0.0206$, the excess loss asymptotically approaches zero. Figure 2 is a plot of excess loss vs L/T in this example.

2.2 Incoherent source, ideal graded index fiber, no mode coupling

Following the same procedures as in 2.1, we can replace σ_m for a self-focusing fiber by $C_3(\Delta/\Delta_o)^2L$ (where Δ_o is the maximum allowable value of Δ). We then obtain the following set of equations which determine the values of Δ and B that minimize the excess loss

$$f'\left(\frac{\sigma}{T}\right)\frac{2\sigma_m^2}{\sigma T} = 1 \qquad \begin{array}{l}\text{provided } \hat{\Delta}/\Delta_o \leqq 1, \\ \text{otherwise } \hat{\Delta}/\Delta_o = 1,\end{array} \tag{9a}$$

$$f'\left(\frac{\sigma}{T}\right)\frac{\sigma_d^2}{\sigma T} = 1 \qquad \begin{array}{l}\text{provided } \hat{B}/B_c \leqq 1, \\ \text{otherwise } \hat{B}/B_o = 1.\end{array} \tag{9b}$$

* Assuming a Gaussian-shaped optical spectrum with bandwidth between the half-power points of about 400 Å.

For sufficiently long lengths L, where both $\hat{\Delta}/\Delta_o$ and $\hat{B}/B_o$ are less than unity, we obtain [by adding (9a) to twice (9b)]

$$\sigma_m = C_3\left(\frac{\hat{\Delta}}{\Delta_o}\right)^2 L = \frac{x'T}{\sqrt{3}} \tag{10}$$

$$\sigma_d = C_1\frac{\hat{B}}{B_o}L = x'T\sqrt{\frac{2}{3}}$$

$$\text{excess loss} = -10\log\left[\frac{(x')^{\frac{3}{2}}T^{\frac{3}{2}}\left(\frac{4}{27}\right)^{\frac{1}{4}}}{L^{\frac{3}{2}}C_1\sqrt{C_3}}\,e^{-f(x')}\right],$$

where x' is the solution of $f'(x')x' = 1.5$, and where we must have

$$\frac{L}{T} > \text{maximum of}\left\{\frac{x'}{C_1}\sqrt{\frac{2}{3}} \quad \text{and} \quad \frac{x'}{C_3}\sqrt{\frac{1}{3}}\right\}.$$

For the Gaussian power penalty curve shown in Fig. 1, we have $x' = 0.34$ and $-10\log e^{-f(x')} = 2.6$ dB.

As before, as L/T decreases, either $\hat{B}/B_o$ or $\hat{\Delta}/\Delta_o$ will reach unity. Thereafter, for smaller values of L/T we have the following: either

$$\sigma \longrightarrow \sigma_d;\ \text{excess loss} \longrightarrow -10\log\frac{zT}{LC_1}e^{-f(z)}, \tag{11}$$

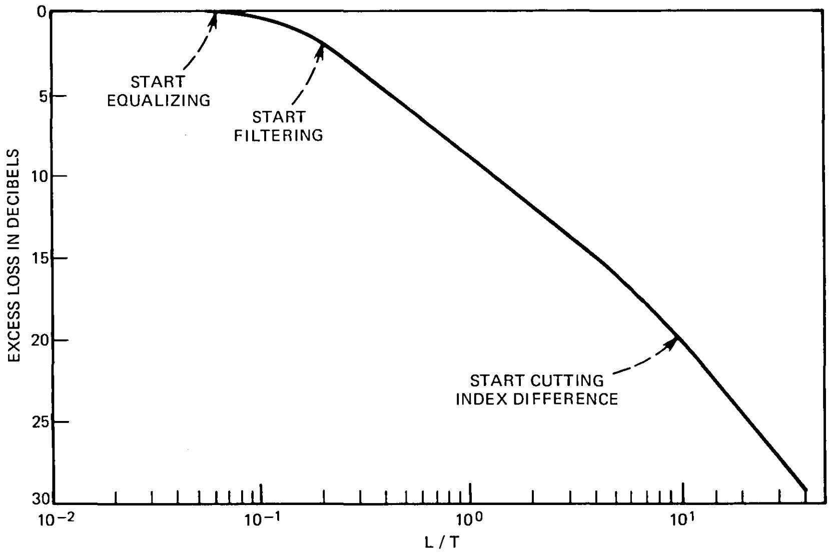


Fig. 3—Excess loss vs L/T for graded index fiber.

where $f'(z)z = 1$, provided $\hat{\Delta}/\Delta_o$ reaches unity first and $L/T > z/C_1$, or

$$\sigma \longrightarrow \sigma_m; \text{ excess loss} \longrightarrow -10 \log \left[\left(\frac{z'T}{LC_3} \right)^{\frac{1}{2}} e^{-f(z')} \right],$$

where $f'(z')2z' = 1$, provided $\hat{B}/B_o$ reaches unity first and $L/T > z'/C_3$. For the Gaussian power penalty curve of Fig. 1, we obtain $z = 0.3$, $-10 \log e^{-f(z)} = 1.8$ dB, $z' = 0.24$, $-10 \log e^{-f(z')} = 0.98$ dB.

When L/T is sufficiently small so that $\hat{\Delta}/\Delta_o$ and $\hat{B}/B_o$ are both equal to unity, then the excess loss asymptotically approaches zero for smaller L/T.

Example 2

From Example 1 we have C_1 typically 1.5 ns/km. From (2) we have C_3 typically 0.019 ns/km for an ideal graded index fiber with a maximum $\Delta = \Delta_o$ of 0.01.

From (10) we obtain

$$\text{excess loss} = -15 \log \left(\frac{T}{L} \right) + 4.85 \text{ dB}$$

for $L/T > 10.3$, where L is in kilometers and T is in nanoseconds.

From (11) we obtain

$$\text{excess loss} \longrightarrow -10 \log \left(\frac{T}{L} \right) + 8.78 \text{ dB}$$

for $0.2 < L/T < 10.3$.

For $L/T < 0.2$, the excess loss asymptotically approaches zero as L/T approaches zero.

Figure 3 shows a plot of excess loss vs L/T for this example.

2.3 Incoherent source, conventional fiber with adjustable mode coupling

Now consider a conventional fiber with adjustable mode coupling. Using the same notation as in 2.1, we have the following condition from (2), (3), and (4)*

$$P_s \frac{\Delta}{\Delta_o} \frac{B}{B_o} e^{-\alpha L} e^{-\alpha_o L/Lc} \geqq P_r e^{f(\sigma/T)}, \tag{12}$$

* As mentioned in Section I, the parameter α_o depends upon the shape of the mechanical coupling spectrum and possibly upon the index difference Δ. Since this dependence upon Δ is not known analytically, we assume α_o = constant independent of Δ. One could also assume $\alpha_o \propto \Delta^N$ for some N (probably negative) and still obtain simple results similar to those that follow using analogous techniques.

where

$$\sigma = \left\{\left(C_1 \frac{B}{B_o} L\right)^2 + \left(C_2 \frac{\Delta}{\Delta_o} \sqrt{LL_C}\right)^2\right\}^{\frac{1}{2}} = \{\sigma_d^2 + \sigma_m^2\}^{\frac{1}{2}},$$

provided $L > L_C$.

By setting appropriate partial derivatives to zero, we can minimize the excess loss given by

$$\text{excess loss} = -10 \log \left\{ \frac{\Delta}{\Delta_o} \frac{B}{B_o} e^{-\alpha_o L/L_C} e^{-f(\sigma/T)} \right\}. \tag{13}$$

The optimizing values of $\hat{\Delta}$, $\hat{B}$, and $\hat{L}_C$ satisfy the following equations (we are assuming α_o fixed by the shape of mechanical coupling spectrum).

$$f'\left(\frac{\sigma}{T}\right)\frac{\sigma_m^2}{\sigma T} = 1, \qquad \text{provided } \hat{\Delta}/\Delta_o \leqq 1, \quad \text{otherwise } \hat{\Delta}/\Delta_o = 1, \tag{14a}$$

$$f'\left(\frac{\sigma}{T}\right)\frac{\sigma_d^2}{\sigma T} = 1, \qquad \text{provided } \hat{B}/B_o \leqq 1, \quad \text{otherwise } \hat{B}/B_o = 1, \tag{14b}$$

$$\frac{\alpha_o L}{\hat{L}_C} = f'\left(\frac{\sigma}{T}\right)\frac{\sigma_m^2}{2\sigma T}, \qquad \text{provided } L > \hat{L}_C, \tag{14c}$$

$$\frac{\alpha_o L}{\hat{L}_C} = 0.5, \qquad \text{provided } \hat{\Delta}/\Delta_o \leqq 1 \text{ and } L > \hat{L}_C. \tag{14d}$$

For sufficiently long fibers and if $\alpha_o \leqq 0.5$, we will have $L > \hat{L}_C$, $\hat{\Delta}/\Delta_o < 1$, $\hat{B}/B_o < 1$, and therefore the following will hold:*

$$\text{excess loss} = -10 \log \left[\frac{x^2 T^2 e^{-0.5}}{2\sqrt{2}L^2 C_1 C_2 \sqrt{\alpha_o}} e^{-f(x)} \right], \tag{15}$$

where x is the solution of $f'(x)x/2 = 1$

$$\frac{\alpha_o L}{\hat{L}_C} = 0.5, \qquad \frac{\hat{\Delta}}{\Delta_o} = \frac{xT}{2C_2 L\sqrt{\alpha_o}}, \qquad \frac{\hat{B}}{B_o} = \frac{xT}{\sqrt{2}LC_1},$$

provided

$$\frac{L}{T} \geqq \frac{x}{2C_2\sqrt{\alpha_o}}; \qquad \frac{L}{T} \geqq \frac{x}{\sqrt{2}C_1}; \qquad \alpha_o \leqq 0.5.$$

It is convenient to consider L/T and $L/\hat{L}_C$ as separate parameters.

* It is interesting to note that, with optimal mode coupling, in the region where $\hat{\Delta} < \Delta_o$, the optimal value of Δ is increased by the factor $\sqrt{0.5/\alpha_o}$ relative to the no-mode coupling case [see formulas for $\hat{\Delta}/\Delta_o$ in (7) and (15) and also (8) and (16)]. Further in this region ($\hat{\Delta} < \Delta_o$), the excess radiation loss from mode coupling ($\alpha_o L/L_c$) is always 0.5 neper.

As L/T decreases, either $\hat{\Delta}/\Delta_o$ or $\hat{B}/B_o$ will reach unity.

If $\hat{B}/B_o$ reaches unity first, then the excess loss will asymptotically approach the following for smaller values of L/T.

$$\text{excess loss} \rightarrow -10 \log \left\{ \frac{zT}{C_2 L \sqrt{2\alpha_o}} e^{-f(z)} e^{-0.5} \right\} \tag{16}$$

$$\frac{\alpha_o L}{\hat{L}_c} = 0.5, \qquad \frac{\hat{\Delta}}{\Delta_o} = \frac{zT}{C_2 L \sqrt{2\alpha_o}},$$

where $f'(z)z = 1$, provided $\hat{B}/B_o$ reaches unity first and

$$L/T > z/(C_2 \sqrt{2\alpha_0}).$$

If $\hat{\Delta}/\Delta_o$ reaches unity first, then the excess loss will asymptotically approach the following for smaller values of L/T

$$\text{excess loss} \rightarrow -10 \log \left\{ \frac{zT}{LC_1} e^{-f(z)} \right\} \tag{17}$$

$$\frac{\hat{B}}{B_o} = \frac{zT}{LC_1},$$

provided $\hat{\Delta}/\Delta_o$ reaches unity first and $L/T \geqq z/C_1$.

For values L/T below that at which $\hat{\Delta}/\Delta_o$ and $\hat{B}/B_o$ both equal unity, the excess loss asymptotically approaches zero.

Example 3

Using the same parameter values as in Example 1 and assuming* $\alpha_o = 0.1$, we obtain the following

$$\text{excess loss} = -20 \log \frac{T}{L} + 27 \text{ dB}$$

for $L/T > 0.174$ km/ns,

$$\text{excess loss} \Rightarrow -10 \log \frac{T}{L} + 17.3 \text{ dB}$$

for $0.0462 < L/T < 0.174$, and

$$\text{excess loss} \Rightarrow 0 \text{ for } \frac{L}{T} < 0.0462.$$

Figure 4 shows a plot of excess loss vs L/T for this example.

* At this point, the achievable value of α_o in practical fibers is a subject of speculation. We choose $\alpha_o = 0.1$ arbitrarily.

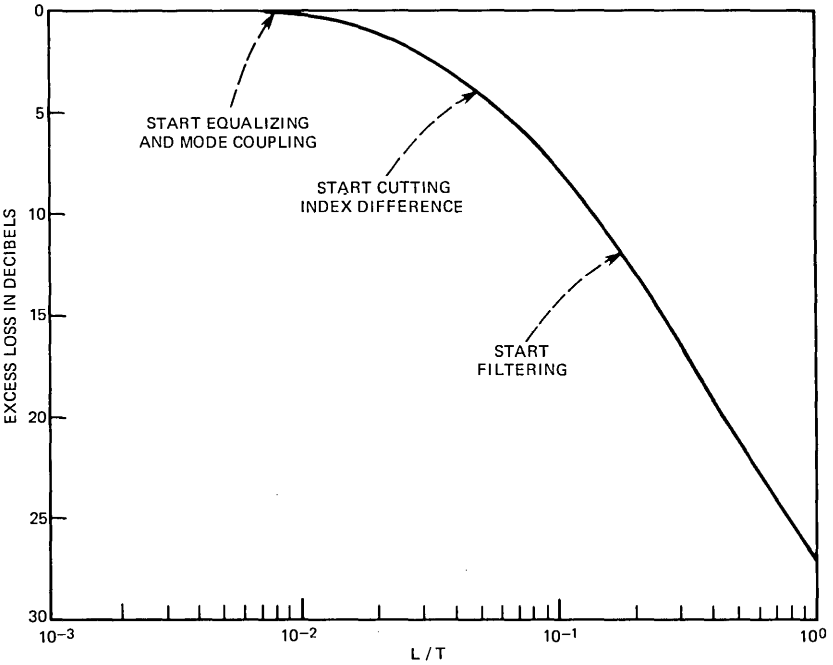


Fig. 4—Excess loss vs L/T for conventional fiber with coupling (α_o = 0.1).

2.4 Laser source, conventional fiber with adjustable mode coupling

Here we assume that, to avoid excessive loss at bends, the index difference in the fiber, Δ, is fixed at some minimum allowable value $\Delta_{\min}$. To control pulse spreading we can trade off mode-coupling radiation loss against equalization penalty. The condition we must satisfy is

$$P_s e^{-\alpha L} e^{-\alpha_0 L/L_C} \geqq P_r e^{f(\sigma/T)}.$$

We wish to choose $\hat{L}_C$ to minimize the excess loss given by

$$\text{excess loss} = -10 \log \{e^{-\alpha_0 L/L_C} e^{-f(\sigma/T)}\}, \tag{18}$$

where*

$$\sigma = C_4\sqrt{LL_C}, \qquad C_4 = 0.289\Delta_{\min} n/c. \tag{19}$$

We obtain the optimizing equation:

$$\alpha_0 L/\hat{L}_C = \frac{1}{2} f'\left(\frac{\sigma}{T}\right)\frac{\sigma}{T}. \tag{20}$$

* Material dispersion is assumed negligible for a coherent source. That is, we assume a single mode and a short-term bandwidth less than 1 Å.

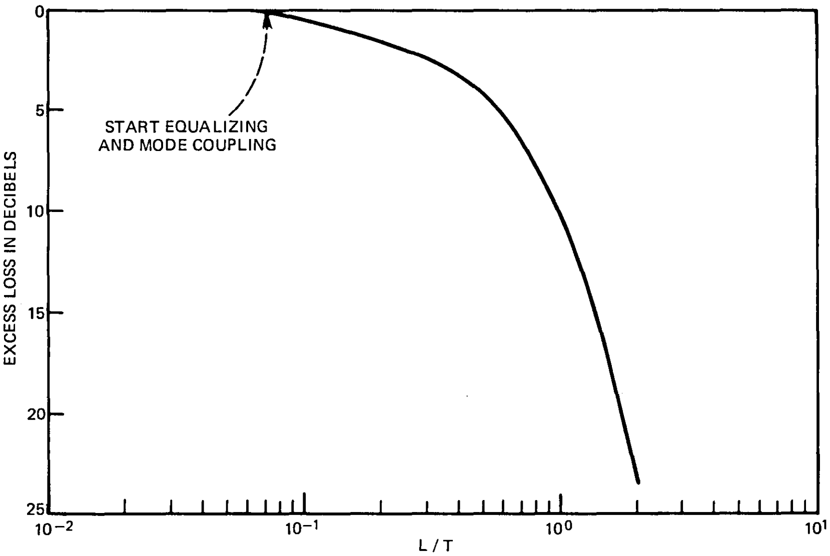


Fig. 5—Excess loss vs L/T—laser source, conventional fiber $\Delta_{\min} = 0.001$, $\alpha_o = 0.1$.

To solve (20) we can pick a value of σ/T and solve for $f'(\sigma/T)$ and $f(\sigma/T)$ graphically from Fig. 1. We then use those results in (20) to solve for $L/\hat{L}_C$. Then we substitute into (19) to find L/T and into (18) to find the total excess loss.

Example 4

Using $\Delta_{\min} = 0.001$ and $\alpha_o = 0.1$, Fig. 5 shows a plot of excess loss vs L/T for this example.

III. APPLICATIONS

If the optical power required at the receiver when the received pulses are very narrow is P_r and the transmitted power (at maximum bandwidth and index difference) is P_s and if the fiber loss in the absence of mode coupling loss is αL, then we must have

$$10 \log (P_s e^{-\alpha L}) - \text{excess loss } (L) = 10 \log (P_r)$$

or, equivalently,*

$$10 \log \frac{P_s}{P_r} e^{-\alpha L} = \text{“excess gain } (L)\text{”} \geqq \text{excess loss}(\, L).$$

* We define excess gain as number of decibels by which $P_s e^{-\alpha L}$ exceeds P_r. In effect, it is equal to the "allowable excess loss."

At a given bit rate, $1/T$, and given α, P_s, and P_r, we can plot excess loss (L) and "excess gain (L)" simultaneously. The intersection of the two curves gives the maximum allowable distance L between the transmitter and the receiver.

Example 5

Assume that, at a bit rate of 25 Mb/s ($T = 40$ ns), the required received power P_r is approximately -58 dBm. Assume that a conventional fiber with mode coupling ($\alpha_o = 0.1$) and loss $\alpha = 5$ dB/km is used. Assume that an incoherent source is being used with $P_s = -13$ dBm for $\Delta_{\max} = 0.01$. Assume that C_1 and C_2 are 1.5 and 14.5 ns/km so that Fig. 4 applies. Figure 6 shows a plot of excess loss and excess gain vs L. It is apparent that the maximum length L between the transmitter and the receiver is 6.8 km. At that distance, the excess loss $= 11.5$ dB. Further, we have $\hat{\Delta} \cong 0.0024$, $\hat{B}/B_o = 1$ and $\hat{L}_C = 1.36$ km.*

IV. COMMENTS AND CONCLUSIONS

The purpose of this paper has been to show how we can combine analytical results on fibers and repeaters to determine maximum repeater spacings by optimization of available parameters. Since the fiber art is still young, many assumptions above are subject to question. We can summarize a few possible criticisms here.

It is not known whether the assumption of the Gaussian pulse shape leads to overly conservative estimates of the equalization penalty. With time and experiments, the Gaussian pulse shape approximation will probably be improved upon.

It is not known yet how much control the designer will have over the mode coupling and the index difference. Future analyses will have to take into account the practical constraints on these parameters.

It is not known yet whether optical filters of the type assumed above can be built. Further, the above analysis neglects in-band insertion loss.

It is hoped that, although the above analysis is somewhat simplistic, it can serve as a guide to the fiber system designer by pointing out the concepts and trade-offs involved.

* It is interesting to note that, if mode coupling were not allowed, the excess loss curve would be about the same (calculated from Fig. 2) and therefore the maximum length L would be the same. However, at the maximum length $\hat{\Delta}$ would be 0.0011.

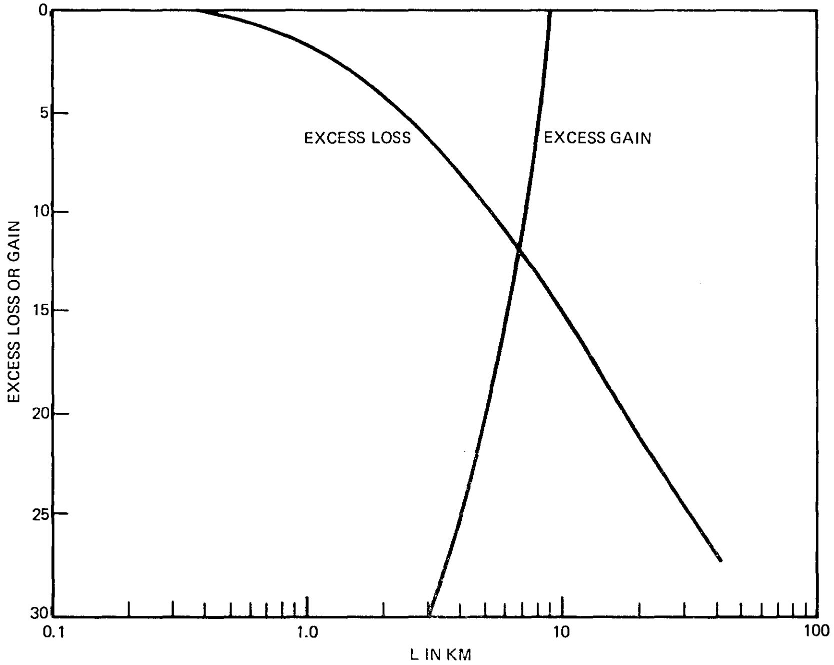


Fig. 6—Excess loss and gain vs L.

APPENDIX

We wish to show heuristically that the total rms width of the fiber impulse response is the rms sum of the contribution from dispersion and the contribution resulting from mode delay spread.

Suppose that, if the fiber is excited by a narrow-band source at wavelength λ, the resultant output response is $h_\lambda(t)$. Let the mean arrival time and rms width of $h_\lambda(t)$ be defined as

$$\tau_\lambda = \frac{1}{A_\lambda} \int t h_\lambda(t) dt \tag{21}$$

$$\sigma_\lambda = \left\{ \frac{1}{A_\lambda} \int t^2 h_\lambda(t) dt - \tau_\lambda^2 \right\}^{\frac{1}{2}},$$

where

$$A_\lambda = \int h_\lambda(t) dt = \text{area of } h_\lambda(t).$$

Now suppose the fiber is excited by a narrow pulse from a broadband source having its output distributed in wavelength according to the spectrum $S(\lambda)$. Intuitively,* we can write the fiber output response as follows:

$$h(t) = \int S(\lambda)h_\lambda(t)d\lambda = \int [S(\lambda)A_\lambda]\left(\frac{h_\lambda(t)}{A_\lambda}\right)d\lambda. \tag{22}$$

Define $\tilde{S}(\lambda)$ as $S(\lambda)A_\lambda$ and let

$$\tilde{S} = \int S(\lambda)A_\lambda d\lambda = \int h(t)dt.$$

Let $\hat{S}(\lambda) = [S(\lambda)A_\lambda/\tilde{S}]$. It follows from (22) that the rms width of $h(t)$ is given by

$$\begin{aligned}\sigma &= \left\{\frac{1}{\tilde{S}}\int t^2h(t) - \left(\frac{1}{\tilde{S}}\int th(t)\right)^2\right\}^{\frac{1}{2}} \\ &= \left\{\left[\int \sigma_\lambda^2\hat{S}(\lambda)d\lambda\right]\right. \\ &\qquad \left. + \left[\int \tau^2(\lambda)\hat{S}(\lambda)d\lambda - \left(\int \tau(\lambda)\hat{S}(\lambda)d\lambda\right)^2\right]\right\}^{\frac{1}{2}}.\end{aligned} \tag{23}$$

The first term in square brackets in (23) is a weighted average of the mean square width of the narrow-band pulse at different wavelengths. The second term is the mean square deviation of the narrow-band-mean-arrival time, i.e., the dispersion σ_d^2. If we next assume that $\sigma_\lambda \approx$ constant $= \sigma_m$ (i.e., that the rms width of the narrow-band impulse response is not dependent upon wavelength within the band of interest), then we obtain

$$\sigma = \{\sigma_m^2 + \sigma_d^2\}^{\frac{1}{2}},$$

which is the desired result.

REFERENCES

1. S. D. Personick, "Receiver Design for Digital Fiber Optic Communication Systems, Part I," B.S.T.J., *52*, No. 6 (July–August 1973), pp. 843–874.
2. S. D. Personick, "Time Dispersion In Dielectric Waveguides," B.S.T.J., *50*, No. 3 (March 1971), pp. 843–859.
3. H. E. Rowe and D. T. Young, "Transmission Distortion in Multimode Random Waveguides," IEEE Trans. on Microwave Theory and Technique, June 1972, pp. 349–356.
4. S. E. Miller, T. Li, and E. A. J. Marcatili, "Research Toward Optical Fiber Transmission Systems," Proc. IEEE, *61*, No. 12 (December 1973), pp. 1703–1751.
5. D. Marcuse, "Pulse Propagation in Multimode Dielectric Waveguides," B.S.T.J., *51*, No. 6 (July–August 1972), pp. 1199–1232.

* But not rigorously. This is where the argument becomes heuristic.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 5, May–June 1974
Printed in U.S.A.

A Map Technique for Identifying Variables of Symmetry

By L. M. GOODRICH

(Manuscript received July 2, 1973)

This paper presents a new map technique for identifying symmetrizable functions. The technique greatly reduces the work in ascertaining symmetricity, and it is unique in being also applicable to completely or incompletely specified functions which:

(*i*) *Contain imbedded symmetrizable function*(*s*).
(*ii*) *Are the complement of a function of type* (*i*).
(*iii*) *Contain an imbedded function of type* (*ii*).

Discussion of the technique and its extensions is included.

I. INTRODUCTION

Recognition of symmetry in circuit design often can drastically reduce the problem of finding the least expensive circuit configuration. Multi-output circuits frequently have a symmetric circuit as a common portion so no single error will result in a wrong output.

As a consequence, numerous papers and chapters of books have presented recognition of symmetry in a switching function.[1–34] However, none of these articles has presented a technique that is simple to apply and has natural extensions to accommodate both completely and incompletely specified functions that are almost symmetrizable.* This paper presents such a technique.

Caldwell[1,2] has demonstrated a technique using Karnaugh maps for recognizing symmetrizable functions (SF's) of three or four variables, and has also demonstrated a procedure for extending this to functions of more variables. The extension requires the use of a large number of maps and the use of an expansion theorem a multiplicity of times. The Caldwell technique requires mapping all possible submaps in four of the variables.

* See Section II for definitions.

Few authors have dealt with any subset of almost SF's (ASF's), and even fewer have tried to give practical examples of a technique for their solution.

Born and Scidmore[3] have dealt with the special subset of ASF's commonly called partial symmetric functions. They have solved a function in six variables for which, by their definition, a minimum solution exists and a resulting realization is shown in Fig. 1.* This same example has been solved by the technique presented in this paper and the resulting circuit, shown in Fig. 2, is significantly more economical.†

In general, other techniques attempt to ascertain symmetry and find the center of symmetry (COS) simultaneously. The technique presented here first uses a set of overlapping maps to ascertain what the COS must be if the function is an SF (or ASF) and then verifies whether the function is symmetric (or almost symmetric) about that COS.

II. METHOD

2.1 Theory

Shannon first stated the definition of a symmetric function as follows: "A function of n variables L_1, L_2, L_3, $\cdots$, L_n is said to be symmetric in these variables if any interchange of the variables leaves the function the same. . . . Since any permutation of variables may be obtained by successive interchange of two variables, a necessary and sufficient condition that a function be a symmetric is that any interchange of two variables leaves the function unaltered." [4]

The nomenclature for a symmetric function has been established by prior usage.[5]

A function f is called an SF if and only if f is equivalent to some function g where g is a symmetric function. Two functions are considered equivalent when one may be obtained from the other by complementing some variables.

When a function f is an SF, the variables of g (any symmetric function equivalent to f) and their complements are called COS's. Such a pair defines an axis of symmetry uniquely specified by either member of the pair. Although any SF has at least two COS's, the function is not symmetric in the same degree (the subscript in standard symmetric notation) about the two centers. In fact, if a function can be represented as m out of n about one COS, then it is $(n - m)$ out of n about the complemented COS.

* This circuit realization could have been as easily done with FET's in MOS technology. Each contact would be replaced with a single FET.

† The details of this are presented in the appendix.

RELAY	CONTACT LOAD
A	4T
B	5T
C	5T,1B
D	5T,1B
E	5T
F	4T
TOTAL	28T,2B

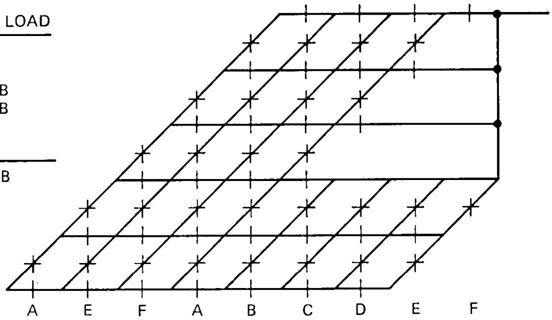


Fig. 1—Simplified circuit for Born and Scidmore problem by Born and Scidmore technique.

A function is an ASF if it

(*i*) Contains one or more imbedded SF's, i.e., the function can be expressed as the sum (oring) of two or more functions where one or more functions are SF's.

(*ii*) Is the complement of a function containing one or more imbedded SF's.

(*iii*) Can be expressed as the sum of two functions, one of which is the complement of a function containing one or more imbedded SF's.

Any function is an ASF if the limits are stretched far enough, but to benefit from symmetricity the number of imbedded SF's should be small and the terms not included in the SF's should be relatively few.

RELAY	LOAD
A	3T
B	2T
C	1T
D	2T,1B
E	2T
F	1T
TOTAL	11T,1B

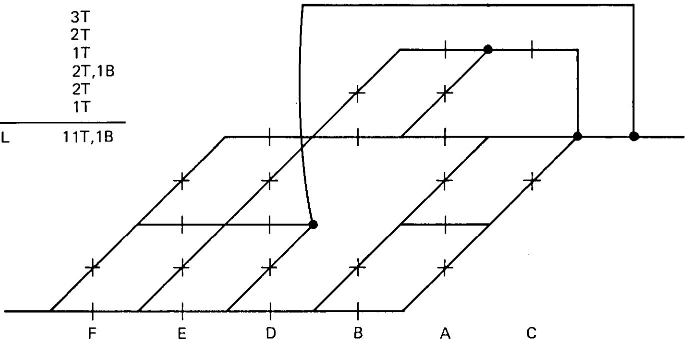


Fig. 2—Simplified circuit for Born and Scidmore problem by author's technique.

Any function fits into one of three categories:

(*i*) The function is not symmetrizable.
(*ii*) The function is symmetrizable about all points.
(*iii*) The function is symmetrizable about exactly two centers of symmetry.[6]

The first category is not the subject of this paper.

The second category consists of only four members, the "zero" function, the "one" function, and the odd and the even parity functions. The first two functions are trivial, while the parity functions are very specific functions that have been the subject of numerous papers.[*] These functions have exactly 2^{2-1} minterms for a function of n variables, and no minterm differs from any other minterm in the state of an odd number of variables.

The third category is of practical interest, being the class of SF's that are not just trivially symmetric. It is also the large majority of SF's with more than three variables ($n > 3$), since this category has $2^{n+1} - 4$ members.

A necessary and sufficient condition for a function to be symmetric about a specific COS may be expressed as follows: If one minterm matches the COS in exactly m out of the total of n variables, then exactly ${}_mC_n$[†] minterms must match that COS in exactly m out of n variables. This is a direct result of the definition of an SF.

Thus, a simple test exists for ascertaining whether a given function is an SF about a specific COS.[‡] The problem then is determining what a COS must be if the function is an SF. Consequently, rather than exhaustively testing a function for symmetry, our technique first finds what the COS must be if the function is an SF and then verifies the actual symmetry about that point.

A direct result of the theorem[11]

$$S_a^n(L_1, L_2, L_3, \cdots, L_n) = \sum_{j=0}^{m} S_j^m(L_1, L_2, L_3, \cdots, L_m) \cdot S_{a-j}^{n-m}(L_{m+1}, L_{m+2}, \cdots, L_n)\,[\S]$$

is that a function f is an SF in $L_1, L_2, L_3, \cdots, L_n$ only if a specific subset of the minterms is an SF in $L_1, L_2, L_3, \cdots, L_m$ for $m \leqq n$. Thus,

[*] Garner (Ref. 7) is one of many who have studied this function.

[†] ${}_mC_n$ is the number of combinations of n things taken m at a time and equals $n!/[(n-m)!\,n!]$.

[‡] In essence, this is the same test that McCluskey (Refs. 8 and 9), Marcus (Refs. 5 and 10), and others have used.

[§] $S_a^n(L_1, L_2, L_3, \cdots, L_n)$ is standard symmetric notation for "symmetric a-out-of-n function of variables $L_1, L_2, \cdots, L_n$."

for any SF, a COS in n variables when all other variables are held constant is a subset of the total COS of the function.

We have shown that a necessary condition for an SF is that the function be symmetrizable in subsets of the variables. This, however, is not a sufficient condition even if the subsets encompass all variables, as we can show by an example. The eight-variable function $f = L_1L_2L_3L_4S_3^4(L_5, L_6, L_7, L_8)$ is symmetrizable in both the four most and the four least significant variables, but is not an SF in all eight variables. Even the fact that the function is symmetrizable in overlapping subsets of variables is not a sufficient condition, as we can demonstrate by the seven-variable function $f = S_6^7(L_1, L_2, \cdots, L_7) + (L_1L_2L_3L_4'L_5'L_6'L_7')$. Thus, the possibility exists that a function is not an SF, even though a composite of the COS's of subsets can be found. However, if COS's are found for subsets of variables where the group of subsets includes all variables in the function, then a COS of the function (if it exists) must be the composite of the COS's of the subsets.

Since each COS has a mate (the point where all variables are the complements of the variables of the first COS), a COS of the function must be a composite involving one or the other COS. The ambiguity as to which COS of each pair to select can be resolved by having each set of variables overlap another set in at least one variable. Thus, if there are n variables and each subset selected has k variables, then at least $|\overline{(n - 1)/(k - 1)}|$ subsets are required to find the COS.*

S. H. Caldwell[1,2] has demonstrated a technique using Karnaugh maps for recognizing symmetry in functions of three or four variables. We review it here.

Any single square on a Karnaugh map represents an SF of n out of n, where n equals the number of variables. Thus, on a three-variable map, each square is an SF of three out of three [written $S_3^3(L_1, L_2, L_3)$] of some set of variables L_1, L_2, and L_3. As an example, the square α in Fig. 3 is $A'BC'$ and is the SF $S_3^3(A'BC')$.

On a Karnaugh map, only one variable changes state in going from one square to an adjacent square. Thus, the four adjacent squares match the center square in two out of three variables.† Similarly, any squares that are two squares from the given square match it in one variable, etc. Thus, the squares labeled 2, 1, and 0 are S_2^3, S_1^3, and S_0^3 of

* $|\overline{\quad\quad}|$ Notation for the least integer equal to or greater than the argument.

† To see the pattern, it is necessary to extend the map-repeating columns and rows, as shown in Fig. 3.

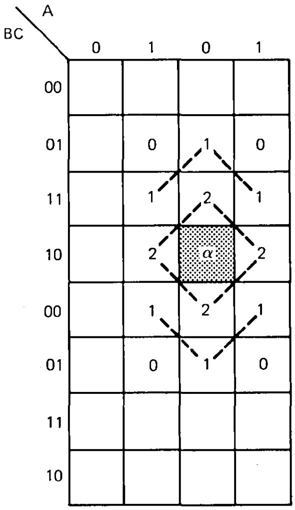


Fig. 3—Symmetry in a three-variable Karnaugh map.

square α, respectively. Note that the coordinates of square α are the variables of symmetry L_1, L_2, L_3, etc. (i.e., the COS).

Expanding this to four variables represents the situation shown in Fig. 4, where the 3's represent S_3^4 of square β, the 2's represent S_2^4 of square β, etc.

Since each SF of three or four variables yields a distinctive pattern, pattern recognition permits recognizing any SF of three or four vari-

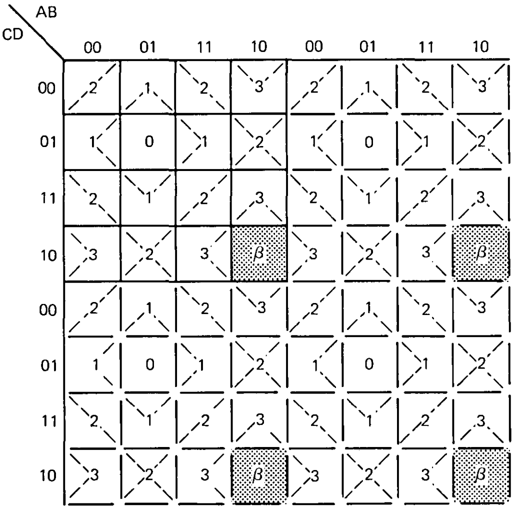
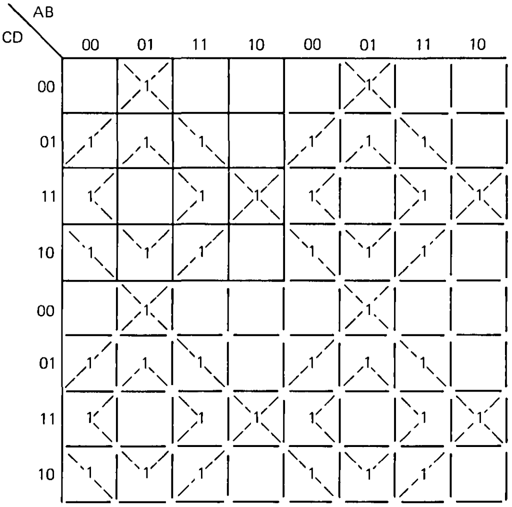


Fig. 4—Symmetry in a four-variable Karnaugh map.



$S^4_{2,3}$ (A'BCD) = S^4_2 (A'BCD) + S^4_3 (A'BCD)

Fig. 5—Sum of two fundamental symmetrics.

ables, and the COS identifies the variables of symmetry. Note that $S^4_0(L_1, L_2, L_3, L_4) = S^4_4(L'_1, L'_2, L'_3, L'_4)$, etc.

The identification of an SF which is the sum of more than one fundamental SF[12] in the same variables is depicted in Fig. 5. This method can be extended to more than four variables, as explained below.

At this point, we depart from Caldwell's technique. Note that Caldwell has presented a means of expanding his method to more than four variables (in theory, to any number). However, for n variables, his technique requires the use of 2^{n-4} Karnaugh maps and $2^{n-4} - 1$ applications of an expansion theorem.[13]

Since a technique exists for handling subsets of four or less variables, four can be substituted for k in the expression $|\overline{(n - 1)/(k - 1)}|$, yielding $|\overline{(n - 1)/3}|$ as the minimum number of subsets required to find the COS of a function of n variables, if it exists. By appropriate selection of subsets, it is possible to require only the use of the minimum number of subsets.

2.2 Technique

First, we list the terms in decimal notation* where unprimed and primed variables are represented by binary ones and zeros, respectively.

* The use of decimal notation is not essential, but it speeds up the mapping and selection of terms which differ only in a specific set of successive variables.

Then, some four-variable Karnaugh maps are drawn. In general, one such map covering the lowest decimal numbers in the function forms a good starting point. Figure 6 presents a review of the decimal notation for Karnaugh maps with six variables.

Once the Karnaugh map has been plotted, we can try to identify an SF on it. If the submap is quite full of ones, checking the remaining zeros for symmetry may be easier. The SF of $S^4_{0,1,2,4}(A, B, C, D')$ shown in Fig. 7 is easier to identify as the complement of $S^4_3(A, B, C, D')$.

If no COS is found on a map, the function is not an SF. If a COS is found, then at least one of the four variables involved plus not more than three new variables is plotted in a similar fashion.* The resulting terms are plotted on a new four-variable map, and the COS (if it exists) is found.

This technique is repeated as many times as is necessary to account for all variables. Since the minimum number of maps is $\lceil (n - 1)/3 \rceil$, the minimum (and usual) number of maps for nine variables is three. Figure 8 tabulates the minimum number of extended four-variable Karnaugh maps required by this technique as compared with the Caldwell-Grea technique. Once the various COS's are found, they are combined to form the COS of the whole function, complementing all the variables in any COS required to make the overlap variables match. This possibility exists since either of two COS's could be selected—i.e., $S^4_3(ABCD) = S^4_1(A'B'C'D')$.

Once the potential COS is found, each term of the function is compared with it to see if a complete SF is represented.

The only restriction on the selection of a subset of minterms for which all variables are fixed except a specific four is that the subset must have at least one member and may not have either 8 or all 16 members.†

The reason for the eight-term restriction is that eight terms in four variables represent either the odd or even parity function, which are

* A suggested technique for this is to divide the decimal value of each term by 2 to include one new variable, by 4 to include two new variables, and by 8 to include three new variables. If insufficient terms are found by this technique, any number less than the divisor can be subtracted from each term's value prior to the division. Division by 8 selects those terms ending in 000 and discards the last three terms. Subtraction of 2 followed by division by 8 selects the leftmost $n - 3$ bits of those terms ending in 010, etc.

† $S^{10}_{0,1,2,3,4}$ $(ABCDEFGHIJ)$ has one full subset for any four variables (i.e., the subset where all fixed variables are zero although all other subsets in those variables are not full). There are no terms greater than 640 in the above function, since all such terms require at least five ones.

AB = 00

EF \ CD	00	01	11	10
00	0	4	12	8
01	1	5	13	9
11	3	7	15	11
10	2	6	14	10

AB = 01

EF \ CD	00	01	11	10
00	16	20	28	24
01	17	21	29	25
11	19	23	31	27
10	18	22	30	26

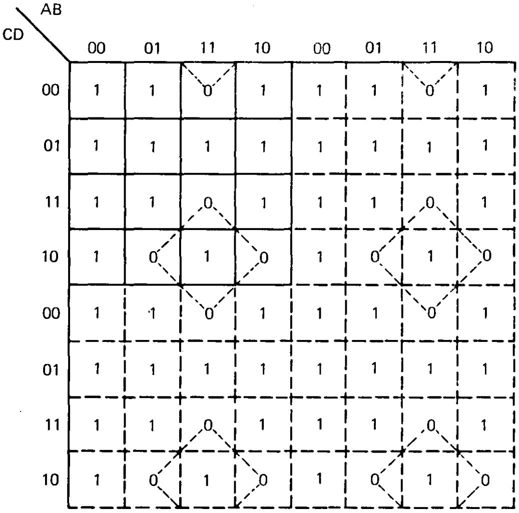
AB = 10

EF \ CD	00	01	11	10
00	32	36	44	40
01	33	37	45	41
11	35	39	47	43
10	34	38	46	42

AB = 11

EF \ CD	00	01	11	10
00	48	52	60	56
01	49	53	61	57
11	51	55	63	59
10	50	54	62	58

Fig. 6—Decimal notation for Karnaugh maps.



CD \ AB	00	01	11	10	00	01	11	10
00	1	1	0	1	1	1	0	1
01	1	1	1	1	1	1	1	1
11	1	1	0	1	1	1	0	1
10	1	0	1	0	1	0	1	0
00	1	1	0	1	1	1	0	1
01	1	1	1	1	1	1	1	1
11	1	1	0	1	1	1	0	1
10	1	0	1	0	1	0	1	0

THE O's REPRESENT S_3^4 (ABCD')

THEREFORE THE 1's REPRESENT $S^4_{0,1,2,4}$('ABCD')

Fig. 7—Identifying a large symmetric.

NUMBER OF VARIABLES	NUMBER OF EXTENDED FOUR VARIABLE KARNAUGH MAPS REQUIRED		NUMBER OF APPLICATIONS OF EXPANSION THEOREM	
	AUTHOR	CALDWELL–GREA	AUTHOR	CALDWELL–GREA
4	1	1	0	0
5	2	2	0	1
6	2	4	0	3
7	2	8	0	7
8	3	16	0	15
9	3	32	0	31
10	3	64	0	63
11	4	128	0	127
12	4	256	0	255
13	4	512	0	511
14	5	1024	0	1023
15	5	2048	0	2047
16	5	4096	0	4095
17	6	8192	0	8191
18	6	16,384	0	16,383
19	6	32,768	0	31,768
20	7	65,536	0	65,535

Fig. 8—Comparison of author's technique and Caldwell-Grea technique.

symmetrical about all possible minterms in four variables. In general, the selection of another set of values for the fixed variables results in a map without eight terms.*

2.3 *Illustrative example*

We test here a 10-variable function for symmetry. Trying all 2^{10} possible combinations of terms or plotting the 64 four-variable Karnaugh maps and mathematically combining them are unattractive. On the other hand, a 10-variable function is not unwieldy by this technique.

Let us consider the function in variables A through J where $F = \sum$ (13, 21, 25, 28, 31, 37, 41, 44, 47, 49, 52, 55, 56, 59, 62, 93, 109, 117, 121, 124, 127, 157, 173, 181, 185, 188, 191, 253, 285, 301, 309, 313,

* $S^{10}_{1,3}$ $(ABCDEFGHIJ)$ will exhibit apparent parity for only those subsets where all fixed variables have a value of zero.

316, 319, 381, 445, 541, 550, 557, 565, 569, 572, 575, 637, 701, 706, 829, 834, 898, 960, 963, 966, 970, 978, 994).

Figure 9 is a worksheet listing the terms of the function and the terms to be plotted on Karnaugh submaps. The appropriate Karnaugh submaps are shown in Fig. 10.

For the subset in which A through F are zero, the only term is 13 and thus it must be a COS of the four variables $GHIJ$ (1101). Next, we find terms which fit the pattern 000XXXX000. Division by 8 yields one such term, i.e., 7, and so the COS in the four variables $DEFG$ must be 0111.

Next, we find a submap of the variables $ABCD$ by selecting those terms ending in 111000 (i.e., subtracting 7 from the terms found by the first division and then dividing by 8). The only resulting term is 0000. Thus, the COS of the whole function (if it exists) must be a composite of $ABCD$ = 0000 and $DEFG$ = 0111, and $GHIJ$ = 1101, i.e., 0000111101.

All that remains is to verify whether or not the function is an SF about this COS. Consequently, the COS is filled in at the head of the "MATCH" column on the worksheet and each individual term is

n	A n/8	(A-7)/8	TERM	MATCH	n	A n/8	(A-7)/8	TERM	MATCH
13					301				
21					309				
25					313				
28					316				
31					319				
37					381				
41					445				
44					450				
47					541				
49					557				
52					565				
55					569				
56	7	0			572				
59					575				
62					637				
93					701				
109					706				
117					829				
121					834				
124					898				
127					960	120			
157					963				
173					966				
181					970				
185					978				
188					994				
191									
253									
285									

Fig. 9—Worksheet for sample problem.

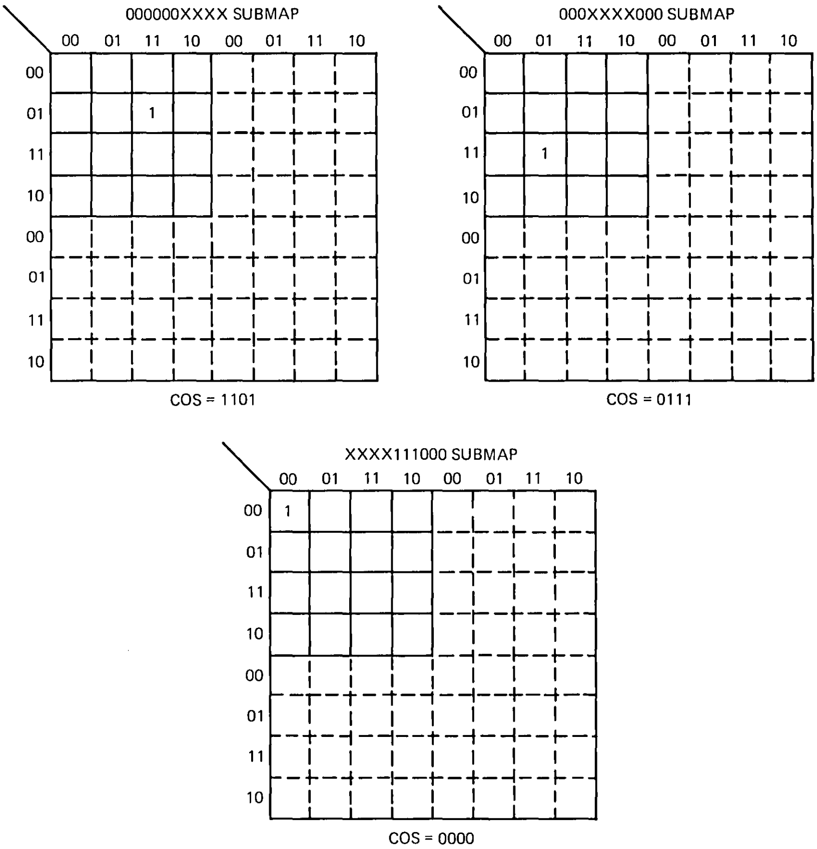


Fig. 10—Karnaugh submaps for sample problem.

compared with this pattern and the number of matching variables recorded. Figure 11 presents the completed worksheet. Next, the number of matches is compared with the number required for an SF, i.e., ${}_mC_n$. For $m = 8$ and $n = 10$, ${}_mC_n = 10!/(8! \times 2!) = 45$ and for $m = 1$ and $n = 10$, ${}_mC_n = 10!/(1!\,9!) = 10$.

Since 45 terms match on eight variables and 10 terms match on one variable, the function is an SF, i.e., $S^{10}_{1,8}(61)$.*

A technique has been presented that is simple to use manually and requires no extensive memorization of a routine such as required by the Marcus-McCluskey method.[10,8] However, the real power of the

* This is the decimal notation abbreviation for $S^{10}_{1,8}$ ($A'B'C'D'EFGHI'J$).

	A			MATCH		A			MATCH
n	n/8	(A-7)/8	TERM	0000111101	n	n/8	(A-7)/8	TERM	0000111101
13			0000001101	8	301			0100101101	8
21			0000010101	8	309			0100110101	8
25			0000011001	8	313			0100111001	8
28			0000011100	8	316			0100111100	8
31			0000011111	8	319			0100111111	8
37			0000100101	8	381			0101111101	8
41			0000101001	8	445			0110111101	8
44			0000101100	8	450			0111000010	1
47			0000101111	8	541			1000011101	8
49			0000110001	8	557			1000101101	8
52			0000110100	8	565			1000110101	8
55			0000110111	8	569			1000111001	8
56	7	0	0000111000	8	572			1000111100	8
59			0000111011	8	575			1000111111	8
62			0000111110	8	637			1001111101	8
93			0001011101	8	701			1010111101	8
109			0001101101	8	706			1011000010	1
117			0001110101	8	829			1100111101	8
121			0001111001	8	834			1101000010	1
124			0001111100	8	898			1110000010	1
127			0001111111	8	960	120		1111000000	1
157			0010011101	8	963			1111000011	1
173			0010101101	8	966			1111000110	1
181			0010110101	8	970			1111001010	1
185			0010111001	8	978			1111010010	1
188			0010111100	8	994			1111100010	1
191			0010111111	8					
253			0011111101	8					
285			0100011101	8					

Fig. 11—Completed worksheet for sample problem.

technique is that it can readily be extended to cases that are not pure SF's, while the other methods cannot be so extended. The reason the new technique can be extended is because it depends on pattern recognition at which humans are adept. Thus, patterns can be discerned in spite of extraneous data, i.e., "noise," while a technique which rigorously uses all data in a prescribed functional relationship cannot sort out the desired data. Obviously, an SF can be so obscured by "noise" as to be unrecognizable. However, the cases of the most value are those which have only limited obscuring terms.

In the next section, some extensions of this technique are presented.

III. EXTENSIONS OF THE TECHNIQUE

Since the technique may be extended to cover a variety of situations, we discuss some specific extensions here. The author has worked problems for each of the discussed extensions, and at least one of each has been presented in unpublished memoranda, while one example of a composite function is solved in the appendix.

For the purpose of this paper, a function is a multiple SF if it is the sum of two or more SF's whose COS's are not the same or complements of each other. Thus, the function $f = S_2^5(ABCDE') + S_3^5(A'BCDE)$ is said to be a multiple SF. Since any minterm of a function of n variables is the SF S_n^n (minterm), any function of m minterms can be represented as the sum of not more than m SF's. However, the useful cases are functions that are the sum of considerably less SF's than minterms, such as the above example.

An incompletely specified function is a function that contains at least one minterm for which transmission is neither required nor forbidden (don't-care terms).

A function is called an incomplete SF if all but a few minterms required for an SF are present.

An overly complete SF is a function composed of an SF plus additional minterms. Any function can be forced to fit this mold, but the cases of interest are those in which almost all the minterms are included in the SF.

Approximate SF's are functions that are not truly SF's but that differ only slightly from an SF. Incomplete SF's are included in approximate SF's, as are overly complete SF's. However, approximate SF's also cover functions that are the sum of incomplete SF's and some additional minterms.

Obviously, any function fits into the category of an approximate SF, but the useful cases are those whose deviation from an SF is slight. For example, the function $f = S_3^8(ABC'DE'FG'H) + AB'C'D'EF'GH' - ABC'D'EF'GH'$ would fit this category. Functions like this can often be simply constructed by modifying the SF.

A partial symmetric has been defined as a function which is an SF in some but not necessarily all variables.[14] Thus, SF's are a subset of partial symmetrics. However, usually a partial symmetric refers to one which is not an SF in all variables.

Both multiple SF's and overly complete SF's are examples of functions containing imbedded SF(s). An incomplete SF is a function whose complement contains an imbedded SF, while approximate SF's are the sum of two functions, one of which is the complement of a function containing an imbedded SF.

3.1 Overly complete SF's

In many cases, a function can be expressed as an SF plus certain additional terms. That is to say, the function has an imbedded SF. The cases of interest are those in which the number of additional terms is small, since a symmetric design can be used for most terms and then the additional terms can be treated individually. In some cases, part of the additional terms can be combined directly with the SF.

The technique is basically the same as it is for a pure SF. However, more maps may be required because of the masking effect of potential extraneous minterms. An extra minterm on a map could result in an ambiguity as to where the COS is. Worse than this, a single extraneous term could be the only term which appears on a specific submap. This could lead to the selection of a false COS for the entire function and an indication that the function did not contain an SF.

Because of this, it is best to select two submaps in the same four variables and get a concurrence as to the COS of the four variables. It is, of course, possible to find in some functions pairs of submaps in the same variables which do not concur on a single COS. Then a third submap is required in those same four variables. In the author's experience, this seldom occurs.

Theoretically, it is possible to have to continue making more submaps without actually determining what the COS of the four variables must be. However, the cases of the most value are those cases that require the least maps. In practice, if concurrence on a COS cannot be found by the use of four submaps in the same four variables, no very useful SF exists within the function. Also, if any one submap has more than three terms, it is usually possible to determine the COS of the four variables.

When dealing with an SF, we stated that a submap with eight terms represents one of the two parity functions. The parity function is recognized by the fact that, on a Karnaugh map, no two adjacent squares have the same value (zero or one). If eight terms appear on the submap and this relationship is not true, then the COS can usually be readily determined.

As before, once the various COS's are found, they are combined to form the point which must be the COS if the function is primarily composed of an SF. Then each term in the original function is compared with it to determine if all the terms of an SF are represented and which, if any, additional terms are included in the function.

3.2 Multiple SF's

It is usually possible, if there are two or more SF's imbedded in a function, to find each of them. The technique is based on the principle used in finding an SF. Again, submaps are selected for the sets of four variables, using more than one submap in the same four variables if necessary. In practice, there is often less confusion when the additional terms do form an SF. Usually it is possible to see more than one COS in a single submap unless the two SF's both have the same COS in the chosen variables. However, since it is possible to have two SF's which have the same COS in four variables although not in all the function variables, the appearance of only one COS in a submap does not rule out the possibility of having more than one SF. Also, since in some submaps no terms of one SF may appear, if a function is found to have some isolated terms, these terms should be investigated to see if they (plus perhaps some terms in the discovered SF) form another SF.

At any point in the analysis of the function that more than one SF appears to be found on the same submap, the remainder of the analysis should be done on the basis of a supposed multiple SF. When two (or more) COS's are found that are not identical or mates, each COS is used as a basis for determining the minterms to be used in a submap in the next selected set of variables. In general, selection of two new variables and two old variables works better when a multiple SF is suspected, since the new set of variables must contain enough information to make possible the selection of the appropriate COS. As long as the variables to be held fixed while new submaps are defined differ in at least one position, different submaps can be defined for finding the next COS, i.e., the submaps 00XXXX11 and 00XXXX01 are different and usually will give COS's that can be readily correlated with the COS's found in the lowest-order submaps (perhaps 1011 and 0001).

3.3 *Incomplete SF's*

An incomplete SF is a function that lacks a few specific terms of being an SF. Thus, it is the complement of a function with an imbedded SF. Such a function can be expressed as one of the three following functions.

(*i*) A modified symmetric circuit.
(*ii*) A symmetric circuit ANDed with another circuit which blocks those terms supplied in the symmetric circuit that are not part of the total function.
(*iii*) A combination of the above.

In the first case, such a modified symmetric circuit would be very efficient. In the second and third cases, the value of finding the SF decreases as the complexity of the blocking circuit increases.

The technique for finding an incomplete SF is similar to that used in previous examples. The major point of difference is that some terms needed to complete the SF are not in the original function, and hence they must be added to the function to form an SF and then blocked by circuit changes.

Finding the COS may require addition of certain terms to the function. Once the proposed COS is found, additional terms may be required to complete the SF.

All terms added to find the COS or to complete the SF must be recorded so the effect of these terms can be deleted in the realization of the function.

3.4 Incompletely specified functions

An incompletely specified function is one that has at least one minterm for which the function is undefined. The use of some terms for which the function is undefined in conjunction with those terms for which the function is defined often permits simpler circuit configurations than could be achieved if all these extra terms are ignored (tacitly forced to a definition of no transmission states). The literature abounds with references to use of these terms in conjunction with Karnaugh maps. Although the use of these terms to aid in completing an SF has been previously ignored by most authors, some work has been done by Arnold and Lawler.[6]

Since the method described here depends on pattern recognition, the use of some terms to complete map patterns is straightforward. Once some (or none) of the undefined minterms have been used to ascertain what the COS of the function must be, then all terms for which transmission is required plus all terms for which the function is undefined are compared with the COS. Those undefined terms which have the same order of symmetry as the required terms are then used to test for the symmetricity of the function and, if sufficient terms are found, these terms are used to transform the function into an SF.

3.5 Approximate SF's

An approximate SF is a function that differs only slightly from a true SF. The cases of greatest interest are those that have a few terms not in the SF and are missing some terms needed to complete the SF. Thus, we can think of this as a combination of an overly complete SF and an incomplete SF. The most interesting problem solutions in these functions occur where the extra and the missing terms can be paired to result in only a minor modification to the true SF.

Here, as before, the basic attack is to find a COS about which the majority of the terms must be symmetric if the function is to approximate an SF. Since both missing and extra terms may occur, it may be necessary to ignore some terms in attempting to find a COS and it may also be necessary to temporarily add to the function certain other terms to complete the symmetry in some subset of variables. Once a proposed COS is found, certain terms may not be symmetric about that COS, and there may not be enough terms symmetric about the proposed COS to complete the SF.

3.6 Composite function

A composite function is a function that may be an incompletely specified function and that may contain one or more imbedded, incomplete, or complete SF's. In other words, the function can be almost any function.

The procedure is basically the same as has already been used. However, because of the possible complexity of the function, some feature may be obscured and, even though a COS for part of the terms is found, it may be desirable to repeat the process, treating all terms in the first-found incomplete or complete SF as don't-care terms for subsequent analysis. This procedure can be repeated until the work entailed is not warranted by the number of remaining terms not identified with an SF. The Born-Scidmore[3] problem considered in the appendix is an example of a composite function.

IV. CONCLUSIONS

A technique has been presented here for isolating SF's (complete or incomplete) in the presence of other SF's or other terms, or in incompletely specified functions.

Since the method depends basically on pattern recognition, at which humans excel, as opposed to value manipulation, at which they do not excel, this technique is especially useful for manual use. However, since the number of patterns is quite restricted, such a technique could be implemented by a computer program.

Since symmetrics are powerful tools in implementing functions, the recognition of SF's within functions can greatly reduce the work in the synthesis of functions.

V. ACKNOWLEDGMENTS

The author wishes to acknowledge the helpful discussions with C. W. Hoffner during the review of this paper.

APPENDIX

In Reference 3, Born and Scidmore have transformed a partial symmetric in six variables into an SF (pure symmetric) in nine variables. Specifically, the function they examined was $F = \sum (1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 17, 18, 19, 20, 21, 22, 24, 25, 26, 28, 33, 34, 35, 36, 37, 38, 40, 41, 42, 44, 48, 49, 50, 52, 56, 57, 58, 60)$. The minimum SF they presented was $S^9_{2,3,4,5}$ $(A, B, C, D, D, E, E, F, F)$. This SF can be represented as shown in Fig. 12 and, as it is shown, requires 34 transfers plus 2 break contacts with one relay requiring 10 transfers and 2 breaks.

However, this circuit can be simplified by standard techniques. The resulting simplified circuit is shown in Fig. 13, which uses 28 transfers and 2 breaks with the largest single-contact load being 5 transfers and 1 break.

This problem can also be solved by the author's technique. Figure 14 is the worksheet and Fig. 15 the first set of Karnaugh submaps. The 00XXXX submap shows two axes of symmetry, one with one center at 0111 and the other with one center at 1111. Subtracting 3 and dividing by 4 yields four terms for the XXXX11 submap. These four terms almost identify a center of symmetry at 0000 (or 1111). Either one term is missing (0001) which would have come from a minterm 7 or the term 0000 is actually that term shifted out of place by a modification of the symmetric. Either case argues for the use of this 1111 as a center of symmetry. When this is combined with the centers of symmetry in 00XXXX, one match is found, namely, 111111. Thus, the terms are compared to this center of symmetry and 15 terms are found to match it in two variables. Since this is the value of ${}_2C_6$, the function contains $S^6_2(63)$. Next, the remaining terms are plotted on the

RELAY	CONTACT LOAD
A	1T
B	2T
C	3T
D	9T
E	10T,2B
F	9T
TOTAL	34T,2B

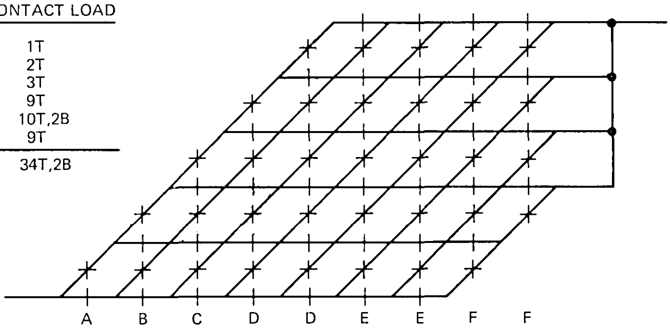


Fig. 12—Symmetric for $S^9_{2,3,4,5}$ $(A, B, C, D, D, E, E, F, F)$.

RELAY	CONTACT LOAD
A	4T
B	5T
C	5T,1B
D	5T,1B
E	5T
F	4T
TOTAL	28T,2B

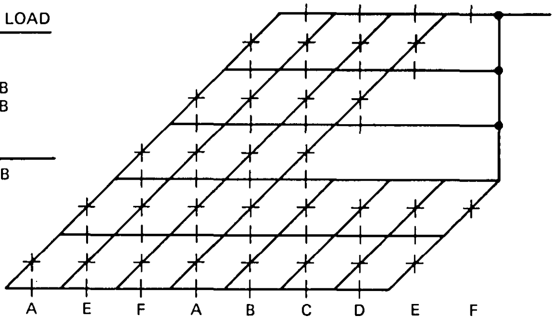


Fig. 13—Simplified symmetric for $S^9_{2,3,4,5}$ $(A, B, C, D, D, E, E, F, F)$.

N	(N-3)/4	TERM	MATCH 111111	MATCH 110111	TERMS MATCHING NEITHER IN TWO VARIABLES
1		000001	1	2	
2		000010	1	2	
3	0	000011	2	3	
4		000100	1	2	
5		000101	2	3	
6		000110	2	3	
9		001001	2	1	
10		001010	2	1	
11	2	001011	3	2	
12		001100	2	1	
13		001101	3	2	
14		001110	3	2	
17		010001	2	3	
18		010010	2	3	
19	4	010011	3	4	√
20		010100	2	3	
21		010101	3	4	√
22		010110	3	4	√
24		011000	2	1	
25		011001	3	2	
26		011010	3	2	
28		011100	3	2	
33		100001	2	3	
34		100010	2	3	
35	8	100011	3	4	√
36		100100	2	3	
37		100101	3	4	√
38		100110	3	4	√
40		101000	2	1	
41		101001	3	2	
42		101010	3	2	
44		101100	3	2	
48		110000	2	3	
49		110001	3	4	√
50		110010	3	4	√
52		110100	3	4	√
56		111000	3	2	
57		111001	4	3	√
58		111010	4	3	√
60		111100	4	3	√
	MISSING TERMS				
16		010000	1	2	
32		100000	1	2	

Fig. 14—Worksheet for Born and Scidmore problem.

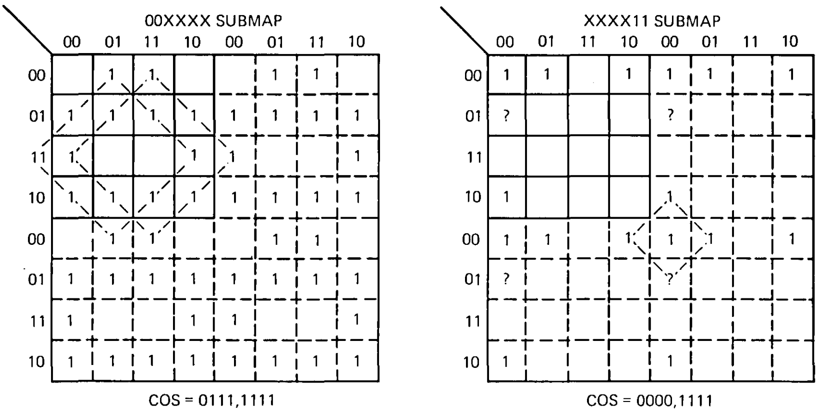


Fig. 15—Karnaugh submaps of Born and Scidmore problem.

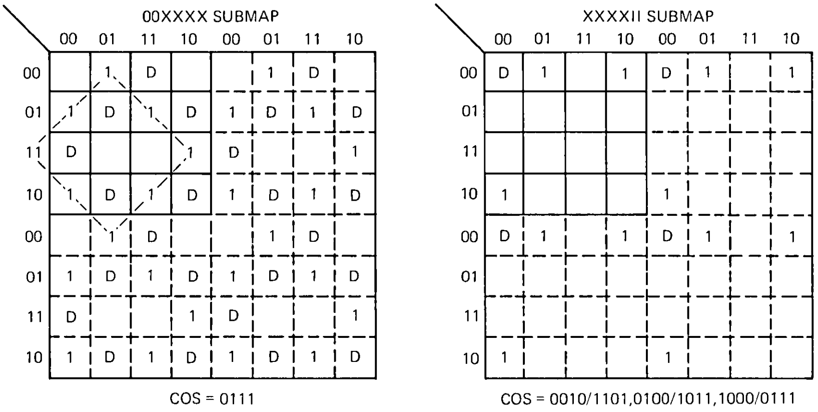
submap 00XXXX (Fig. 16), and a single axis of symmetry (center 0111) is found.* The same last three digits are here as before, so the same terms appear on the XXXX11 submap except for the terms which came from $S_2^6(63)$.

If the three terms on the XXXX11 submap define a center of symmetry with a missing term, then this center of symmetry must be the same as before. Part of the minterms of $S_3^6(63)$, $S_4^6(63)$, and $S_1^6(63)$ appear, but none in sufficient quantity. Therefore, we assume that two of the terms on the XXXX11 submap are from extra terms and try each of the three points as a center of symmetry. Only one of these (1101) can be combined with 0111 so that a proposed center of symmetry is 110111. The terms are compared with it and 13 terms are found to match 110111 on two variables. Since 15 terms are required, the comparison is made for the missing terms. With experience, this can be done very readily. However, for completeness, the terms of $S_2^6(55)$ are tabulated in Fig. 17. The missing terms are 16 and 32. Thus, the function contains the function $S_2^6(55)$-16-32.

Looking at the 12 remaining terms we note that three of them (57, 58, 60) can be accounted for as $ABCS_1^3(DEF)$. The remaining nine terms all have the third variable in the zero state and so can be expressed as $C'f(ABDEF)$.

Repeating the technique on the nine-term, five-variable function results in the 0XXXX, 1XXXX, XXXX0, and XXXX1 submaps shown in Fig. 18. The first two submaps concur on a 1111 COS and the

* Note that the terms already identified appear as "don't care's" (D) in the submaps even though, in this case, they do not result in simplifying the remaining imbedded functions.



00XXXX SUBMAP

	00	01	11	10	00	01	11	10
00		1	D			1	D	
01	1	D	1	D	1	D	1	D
11	D			1	D			1
10	1	D	1	D	1	D	1	D
00		1	D			1	D	
01	1	D	1	D	1	D	1	D
11	D			1	D			1
10	1	D	1	D	1	D	1	D

COS = 0111

XXXXII SUBMAP

	00	01	11	10	00	01	11	10
00	D	1		1	D	1		1
01								
11								
10	1				1			
00	D	1		1	D	1		1
01								
11								
10	1				1			

COS = 0010/1101,0100/1011,1000/0111

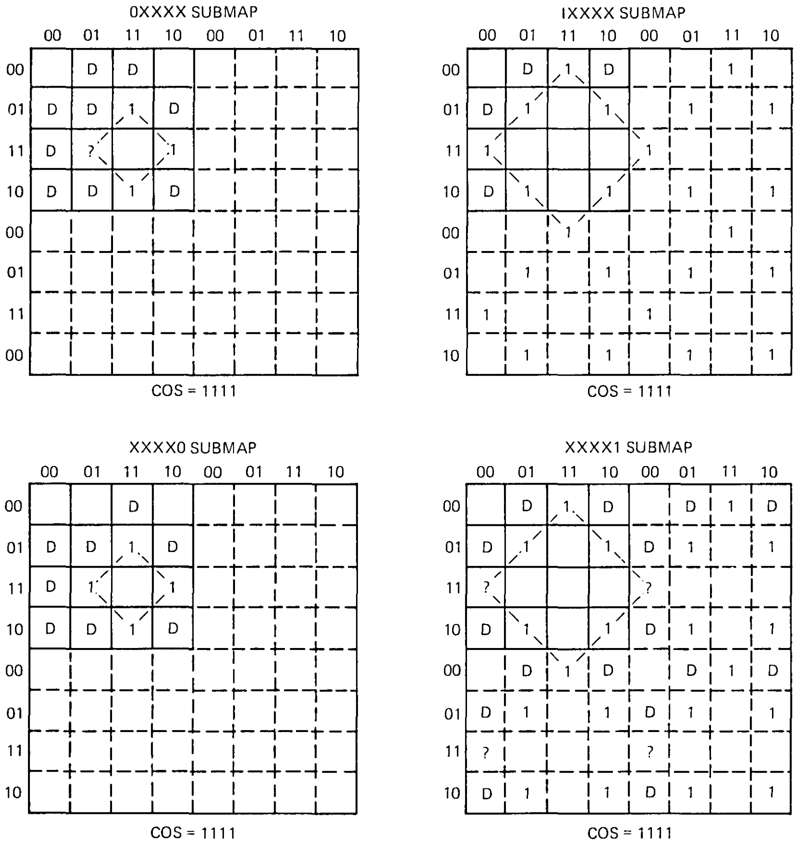
Fig. 16—Karnaugh submaps of remainder of function.

latter two submaps also result in a 1111 COS. As a result, a composite COS of 11111 is tried. Nine of the 10 needed terms for $S_3^5(31)$ are present with only the term 00111 missing. Thus, the original function has been broken up into the sum of four smaller functions, namely:

(*i*) $S_2^6(ABCDEF)$.
(*ii*) $S_2^6(ABC'DEF) - A'BC'D'E'F' - AB'C'D'E'F'$.
(*iii*) $ABCS_1^3(DEF)$.
(*iv*) $C'S_3^5(ABDEF) - A'B'C'DEF$.

TERM	N	INCLUDED IN FUNCTION
111000	56	✓
100000	32	
101100	44	✓
101010	42	✓
101001	41	✓
010000	16	
011100	28	✓
011010	26	✓
011001	25	✓
000100	4	✓
000010	2	✓
000001	1	✓
001110	14	✓
001101	13	✓
001011	11	✓

Fig. 17—Terms of symmetric S_2^6 (55).



0XXXX SUBMAP

	00	01	11	10	00	01	11	10
00		D	D					
01	D	D	1	D				
11	D	?		1				
10	D	D	1	D				
00								
01								
11								
00								

COS = 1111

IXXXX SUBMAP

	00	01	11	10	00	01	11	10
00		D	1	D			1	
01	D	1		1		1		1
11	1				1			
10	D	1		1		1		1
00			1				1	
01		1		1		1		1
11	1				1			
10		1		1		1		1

COS = 1111

XXXX0 SUBMAP

	00	01	11	10	00	01	11	10
00			D					
01	D	D	1	D				
11	D	1		1				
10	D	D	1	D				
00								
01								
11								
10								

COS = 1111

XXXX1 SUBMAP

	00	01	11	10	00	01	11	10
00		D	1	D		D	1	D
01	D	1		1	D	1		1
11	?				?			
10	D	1		1	D	1		1
00		D	1	D		D	1	D
01	D	1		1	D	1		1
11	?				?			
10	D	1		1	D	1		1

COS = 1111

COS = 11111 FOR ABDEF
MISSING TERM = 00111

Fig. 18—Karnaugh submaps for final nine terms of function.

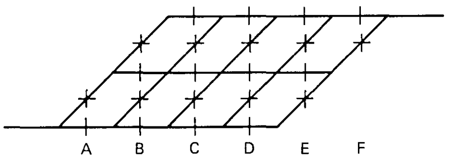


Fig. 19—Circuit for S_2^6 (63).

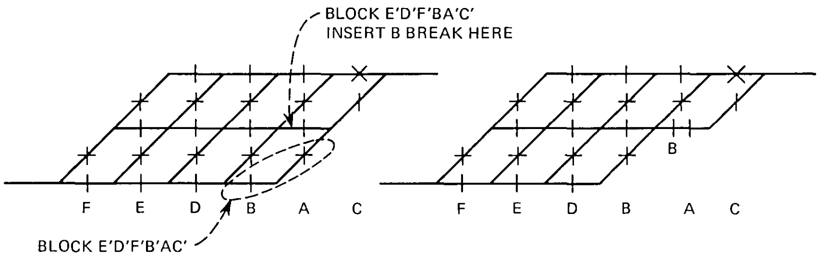


Fig. 20—Construction of circuit for S_2^6(55)-16-32.

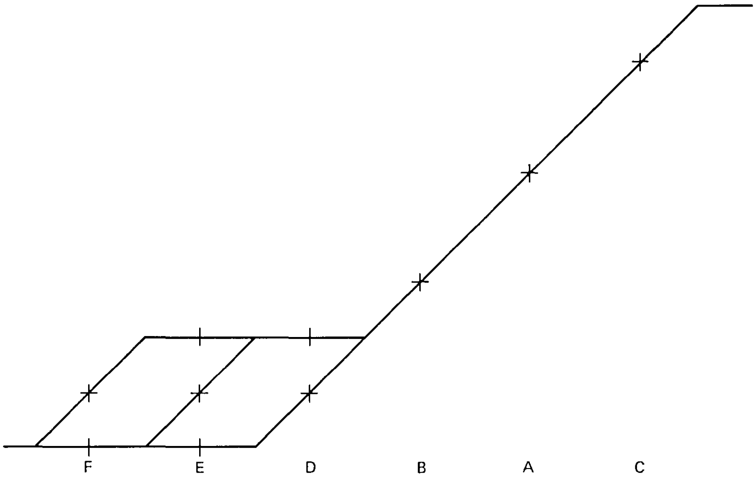


Fig. 21—Symmetric for $ABCS_1^3(DEF)$.

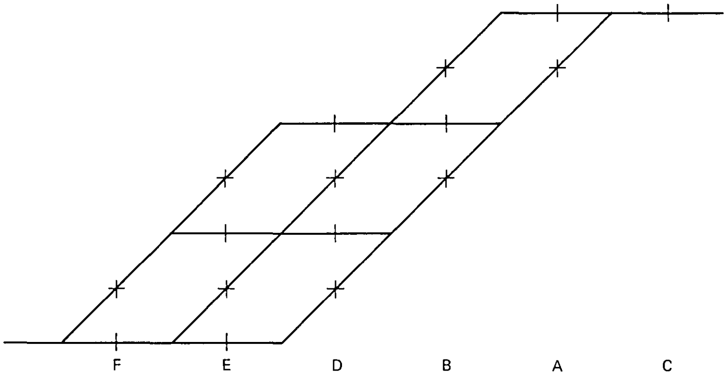


Fig. 22—Symmetric for $C'S_3^5(ABDEF)$—$A'B'C'DEF$.

RELAY	LOAD
A	3T
B	2T
C	1T
D	2T,1B
E	2T
F	1T
TOTAL	11T,1B

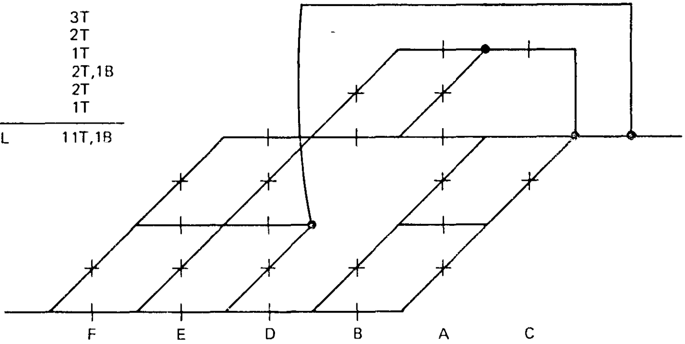


Fig. 23—Final circuit for Born and Scidmore problem.

These represent

(*i*) An SF.

(*ii*) An incomplete SF.

(*iii*) A minterm in three variables ANDed with an SF in the remaining three variables.

(*iv*) A single variable ANDed with an incomplete SF in the remaining five variables.

Figures 19 through 22 portray a relay configuration for each of these smaller functions, while Fig. 23 is the resulting circuit configuration when the preceding four circuits are combined. This resulting circuit uses 11 transfers and 1 break-contact, or less than half of the number required for the circuit resulting from the Born and Scidmore technique.

REFERENCES

1. S. H. Caldwell, "The Recognition and Identification of Symmetric Switching Functions," Trans. AIEE, *73*, Part 1 (May 1954), pp. 142–147.
2. S. H. Caldwell, *Switching Circuits and Logical Design,* New York: John Wiley and Sons, 1958.
3. R. C. Born and A. K. Scidmore, "Transformation of Switching Functions to Completely Symmetric Switching Functions," IEEETEC, *C-17,* No. 6 (June 1968), pp. 596–599.
4. C. E. Shannon, "A Symbolic Analysis of Relay and Switching Circuits," AIEE Trans., *57* (1938), pp. 713–723.
5. M. P. Marcus, *Switching Circuits for Engineers,* 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1967.
6. R. F. Arnold and E. L. Lawler, "On the Analysis of Functional Symmetry," Proc. 4th Ann. IEEE Symp. Switching Theor. Logical Design (September 1963), pp. 53–62.
7. H. L. Garner, "Generalized Parity Checking," IRETEC, *EC-7,* No. 3 (September 1958), pp. 207–213.
8. E. J. McCluskey, Jr., "Detection of Group Invariance or Total Symmetry of a Boolean Function," B.S.T.J., *35,* No. 6 (November 1956), pp. 1445–1453.

9. E. J. McCluskey, *Introduction to the Theory of Switching Circuits*, New York: McGraw-Hill, 1965.
10. M. P. Marcus, "The Detection and Identification of Symmetric Switching Functions With the Use of Tables of Combinations," IRETEC, *5*, No. 4 (December 1956), pp. 237–239.
11. M. Karnaugh, "Discussion on S. H. Caldwell's 'The Recognition and Identification of Symmetric Switching Functions,' " Trans. AIEE, *73*, Part 11 (May 1954), p. 146.
12. G. Epstein, "Synthesis of Electronic Circuits for Symmetric Functions," IRETEC, *EC-7*, No. 1 (March 1958), pp. 57–60.
13. This theorem was communicated to S. H. Caldwell by R. Grea of Graphic Arts Research Foundation, Cambridge, Massachusetts, and reported by the former in *Switching Circuits and Logical Design*, New York: John Wiley and Sons, 1958.
14. R. F. Arnold and M. A. Harrison, "Algebraic Properties of Symmetric and Partially Symmetric Boolean Functions," IEEETEC, *EC-12*, No. 3 (June 1963), pp. 244–251.
15. R. L. Ashenhurst, "The Decomposition of Switching Functions," Int. Symp. Theor. Switching, Part 1 (April 1959), pp. 74–116.
16. H. J. Beuscher, A. H. Budlong, M. B. Haverty, and G. Waldbaum, *Electronic Switching Theory and Circuits*, New York: Van Nostrand Reinhold, 1971.
17. N. N. Biswas, "On Identification of Totally Symmetric Boolean Functions," IEEETEC, *C-19*, No. 7 (July 1970), pp. 645–648.
18. S. R. Das, "On Detecting Total and Partial Symmetry of Switching Functions," IEEE Proc., *58*, No. 5 (May 1970), pp. 840–841.
19. B. Elspas, "Self-Complementary Symmetry Types of Boolean Functions," IRETEC, *9*, No. 2 (June 1960), pp. 264–266.
20. M. A. Fischler and M. Fannenbaum, "Assumptions in the Threshold Synthesis of Symmetric Switching Functions," IEEETEC, *C-17*, No. 3 (March 1968), pp. 273–279.
21. H. Hellerman, *Digital Computer Systems Principles*, New York: McGraw-Hill, 1967.
22. F. J. Hill, and G. R. Peterson, *Introduction to Switching Theory and Logical Design*, New York: John Wiley and Sons, 1968.
23. W. H. Kautz, "The Realization of Symmetric Switching Functions With Linear-Input Logical Elements," IRETEC, *EC-10*, No. 3 (September 1961), pp. 371–378.
24. W. Keister, A. E. Ritchie, and S. H. Washburn, *The Design of Switching Circuits*, New York: Van Nostrand Reinhold, 1951.
25. R. D. Merrill, Jr., "Symmetric Ternary Switching Functions, Their Detection and Realization With Threshold Logic," IEEETEC, *EC-16*, No. 5 (October 1967), pp. 624–637.
26. A. Mukhopadhyaz, "Detection of Disjuncts of Switching Functions and Multilevel Circuit Design," J. Elec. Control, *10* (January 1961), pp. 45–55.
27. A. Mukhopadhyaz, "Detection of Total or Partial Symmetry of a Switching Function With the Use of Decomposition Charts," IEEETEC, *EC-12*, No. 5 (October 1963), pp. 553–557.
28. A. Mukhopadhyaz, "Symmetric Ternary Switching Functions," IEEETEC, *EC-15*, No. 5 (October 1966), pp. 731–739.
29. G. M. Pavarov, "On a Method of Analyzing Symmetric Switching Circuits," Auto. i Tele., *16* (January 1955), pp. 364–365.
30. P. K. S. Roy, "Synthesis of Symmetric Switching Functions Using Threshold Logic Elements," IEEETEC, *EC-16*, No. 3 (June 1967), pp. 359–364.
31. C. L. Sheng, "Detection of Totally Symmetric Boolean Functions," IEEETEC, *EC-14*, No. 6 (December 1965), pp. 924–926.
32. C. L. Sheng, "A Graphical Interpretation of Realization of Symmetric Boolean Functions With Threshold Logic Elements," IEEETEC, *EC-14*, No. 1 (February 1965), pp. 8–18.
33. T. Singer, "Some Uses of Truth Tables," Int. Symp. Theor. Switching, Part 1 (April 1959), pp. 125–133.
34. S. H. Washburn, "Relay 'Trees' and Symmetric Circuits," Trans. AIEE, *68* (1949), pp. 582–586.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 5, May–June 1974
Printed in U.S.A.

Finite-Element Method for the Determination of Thermal Stresses in Anisotropic Solids of Revolution

By S. KAUFMAN

(Manuscript received October 19, 1973)

This paper discusses stresses and deflections in anisotropic solid structures of revolution. It presents two methods based on finite-element techniques: one, a solid-of-revolution method in which material properties, applied forces, and temperatures are independent of angle, and two, a long-cylinder method in which these properties are independent of the longitudinal coordinate. These methods are postulated on uniform stress fields within the element, rather than on the usual functional displacement description within the element. A Fortran program has been written for both these methods, and ample test problems are presented to validate the methods. An application is presented for thermal stresses induced during the post-growth cooling stage of Czochralski-grown lithium tantalate crystals.

I. INTRODUCTION

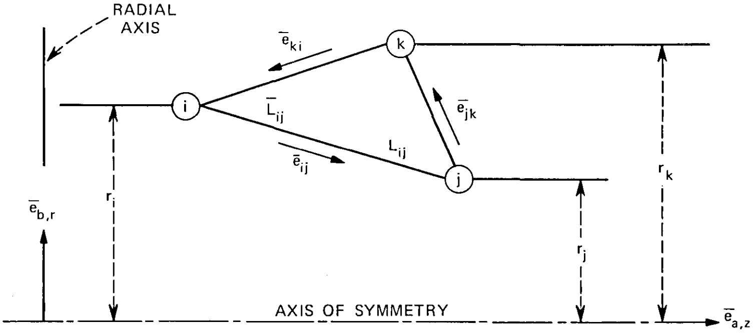
This paper considers two finite-element networks. The first network consists of triangular annuli forming a solid of revolution of any arbitrary cross section in the radial-longitudinal plane. The network has 2π symmetry with regard to material properties, external loadings, and temperatures. The second network consists of trapezoidal and triangular elements in the plane of the circle forming a right circular (actually, a polygon cross section) cylinder long in the longitudinal direction. The restriction of 2π symmetry is lifted for the second network and is replaced with the restriction that material properties, external loadings, and temperatures are independent of the longitudinal direction of the cylinder.

Both methods were programmed on the IBM 370 computer. For the first method, two test plane strain problems are presented: one, a hollow cylinder subjected to a negative radial pressure and two, a solid cylinder subjected to a linear radial temperature gradient. Comparisons of stresses and deflections with known plane strain solutions are excellent in both cases. In the second method, the solution of a long cylinder subjected to a linear radial thermal gradient is presented. Comparison with a known theoretical solution again was excellent.

Results are presented for thermal stresses induced in a lithium tantalate crystal during the post-growth cooling stage. Lithium tantalate is in crystal class C_{3v}. By aligning the trigonal axis of the crystal with the longitudinal axis of the cylinder, the problem can be analyzed by both methods. Comparison of thermal stresses obtained from the two methods was very good.

II. SOLID-OF-REVOLUTION METHOD

The basic element for this method is the triangular annulus shown in Fig. 1. The element, defined in the radial-longitudinal plane ($r - z$), has 2π symmetry. A network of these elements will comprise any desired solid of revolution, whether it is solid or hollow, cylindrical or conical, or any combination thereof. Material properties, temperature distributions, and external forces must be independent of angle. The material properties are then limited to an orthotropic system with isotropic properties in the plane of the circle. This limitation is lifted in the case of the long cylinder method presented in the next section.



$\bar{A} = \tfrac{1}{2}\bar{L}_{ij} \times \bar{L}_{jk} \qquad A = |A|$

$V = 2\pi A(r_i + r_j + r_k)/3$

$\bar{e}_{ij} = \bar{L}_{ij}/L_{ij}$

$A_{ij} = V/L_{ij}$[HALF ALTITUDE AREA (k TO ij)]

$a_{ij} = \bar{e}_{ij} \cdot \bar{e}_a \qquad b_{ij} = \bar{e}_{ij} \cdot \bar{e}_b$

Fig. 1—Properties of triangular annulus.

Solid triangular elements have been used successfully by other authors;[1,2] however, a different approach is presented here. Rather than to assume a displacement function,[1,2] a simpler and more direct approach is to postulate uniform stress fields in the annulus. This method was applied successfully in the mid-60's by David B. Hall for the triangular membrane, but unfortunately his work has not been published. Hall's triangular membrane is extended here to a three-dimensional model by incorporating a uniform hoop stress in the annulus.

The initial strategy is to relate equilibrium between forces at the three points of the triangle i, j, k, and stress fields parallel to the sides of the triangle σ_{ij}, σ_{jk}, σ_{ki}, and the hoop stress σ_θ. Before we do this, let us first compute a few properties of the triangular annulus. These properties include the area of the triangle A, length of each side L_{ij}, etc., volume V, areas A_{ij}, etc. associated with the stress fields σ_{ij}, etc., and direction cosines a_{ij}, b_{ij}, etc. These properties are presented in Fig. 1.

In matrix notation, equilibrium for annulus n between grid point forces and the four stress fields described above is given as follows.

$$\{F_n\} = [B]\{\sigma_o\}, \tag{1}$$

where

$$\{F_n\} = \{F_{ir}F_{iz}F_{jr}F_{jz}F_{kr}F_{kz}\},$$
$$\{\sigma_o\} = \{\sigma_{ij}\sigma_{jk}\sigma_{ki}\sigma_\theta\},$$

and

$$[B] = \begin{bmatrix} -A_{ij}b_{ij} & 0 & A_{ki}b_{ki} & 2\pi A/3 \\ -A_{ij}a_{ij} & 0 & A_{ki}a_{ki} & 0 \\ A_{ij}b_{ij} & -A_{jk}b_{jk} & 0 & 2\pi A/3 \\ A_{ij}a_{ij} & -A_{jk}a_{jk} & 0 & 0 \\ 0 & A_{jk}b_{jk} & -A_{ki}b_{ki} & 2\pi A/3 \\ 0 & A_{jk}a_{jk} & -A_{ki}a_{ki} & 0 \end{bmatrix}.$$

The r and z subscripts attached to the forces F_{ir}, F_{iz}, etc. refer to the radial and longitudinal directions, respectively.

The skew stress field σ_{ij}, σ_{jk}, σ_{ki} is merely an artifice to allow us to readily establish the equilibrium relationship. We are actually interested in the orthogonal field σ_z, σ_r, τ_{rz}, as shown in Fig. 2. The relationship between the two fields including the hoop stress is given below.

$$\{\sigma_n\} = [D]\{\sigma_0\}, \tag{2}$$

where

$$\{\sigma_n\} = \{\sigma_z\sigma_r\tau_{rz}\sigma_\theta\}$$

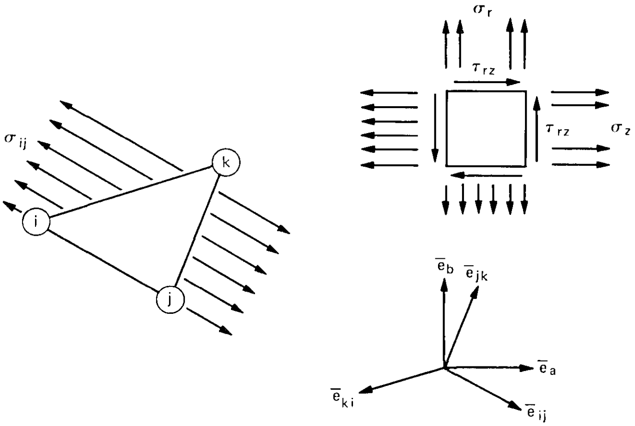


Fig. 2—Stress fields.

and

$$[D] = \begin{bmatrix} a_{ij}^2 & a_{jk}^2 & a_{ki}^2 & 0 \\ b_{ij}^2 & b_{jk}^2 & b_{ki}^2 & 0 \\ a_{ij}b_{ij} & a_{jk}b_{jk} & a_{ki}b_{ki} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} .$$

After inverting eq. (2) and substituting it into eq. (1), the following relationship is obtained.

$$\{F_n\} = [F]\{\sigma_n\}, \tag{3}$$

where

$$[F] = [B][D]^{-1}.$$

A conjugate relationship to eq. (3) by an application of the principle of virtual work can be stated as

$$\{\epsilon_n\} = \frac{1}{V}[F]^t\{U_n\}, \tag{4}$$

where the strains are

$$\{\epsilon_n\} = \{\epsilon_z \epsilon_r \gamma_{rz} \epsilon_\theta\},$$

the deflections conjugate to $\{F_n\}$ are

$$\{U_n\} = \{U_{ir}U_{iz}U_{jr}U_{jz}U_{kr}U_{kz}\},$$

and the superscript t denotes a transposition.

The stress-strain relationship, including the thermal strains $\int \alpha dT$, is given as

$$\{\sigma_n\} = [C]\left(\{\epsilon_n\} - \left\{\int \alpha dT\right\}\right), \tag{5}$$

where $[C]$ is a symmetric 4 × 4 stress-strain matrix such that

$$C_{11} = C_{12}, \quad C_{14} = C_{24}, \quad C_{13} = C_{23} = C_{43} = 0,$$

and

$$\left\{\int \alpha dT\right\} = \left\{\int \alpha_z dT \;\int \alpha_r dT \;\; 0 \int \alpha_\theta dT\right\}, \quad \alpha_r = \alpha_\theta.$$

Substituting eq. (5) into eq. (4) obtains the stresses in terms of the unknown deflections and known thermal strains as

$$\{\sigma_n\} = [C]\left(\frac{1}{V}[F]^t\{U_n\} - \left\{\int \alpha dT\right\}\right). \tag{6}$$

Substituting eq. (6) into eq. (3) obtains contributions to the network stiffness matrix and thermal load vector for annulus n or

$$\{F_n\} = [K_n]\{U_n\} - \{E_n\}, \tag{7}$$

where

$$[K_n] = \frac{1}{V}[F][C][F]^t$$

is a symmetric 6 × 6 stiffness matrix and

$$\{E_n\} = [F][C]\left\{\int \alpha dT\right\}$$

is a 6 × 1 thermal load vector.

Annuli contributions to the network stiffness matrix and thermal load vector are additive, and hence eq. (7) can be written for the network as

$$\{F\} = [K]\{U\} - \{E\}, \tag{7a}$$

where

$$[K] = \sum_n [K_n] \quad \text{and} \quad \{E\} = \sum_n \{E_n\},$$

and where the number of equations equals twice the number of points of the network (one longitudinal and one radial degree of freedom per point).

The $[K]$ matrix in eq. (7a) is singular, as there is no constraint to prevent rigid body motion in the longitudinal direction. Rigid-body motion in the radial direction is prevented by the hoop stress field, as can be seen in eq. (1). (Note that in Eq. (1) σ_θ is the only non-self-equilibrating stress field.) In addition to at least one reference longitudinal constraint, radial constraints must be provided for solid cylinders at points of zero radius to prevent radial (or hoop) motions at these singularities. The degrees of freedom can now be partitioned into an unconstrained set denoted by "a" and a constrained set denoted by "b." This partitioning can be represented schematically as

$$\begin{Bmatrix} F_a \\ F_b \end{Bmatrix} = \begin{bmatrix} K_{aa}K_{ab} \\ K_{ab}^t K_{bb} \end{bmatrix} \begin{Bmatrix} U_a \\ U_b = 0 \end{Bmatrix} - \begin{Bmatrix} E_a \\ E_b \end{Bmatrix}. \tag{8}$$

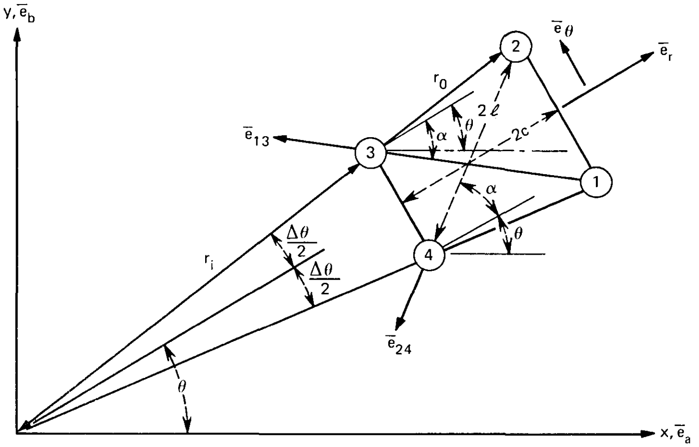
From the upper equation of (8), the previously unknown deflections can now be obtained in terms of known external forces and thermal loads, or

$$\{U_a\} = [K_{aa}]^{-1}(\{F_a\} + \{E_a\}). \tag{9}$$

From the lower equation of (8) and with the help of eq. (9), the forces of constraint, if desired, can be obtained as

$$\{F_b\} = [K_{ab}]^t[K_{aa}]^{-1}\{F_a\} - \{F_b^*\} - (\{E_b\} - [K_{ab}]^t[K_{aa}]^{-1}\{E_a\}), \tag{9a}$$

where $\{F_b^*\}$ represents external forces applied at the constraint points.



$$\alpha = \tan^{-1}[(r_0+r_i)\sin\frac{\Delta\theta}{2}/(r_0-r_i)] \qquad a_r = \bar{e}_r \cdot \bar{e}_a \qquad b_r = \bar{e}_r \cdot \bar{e}_b$$

$$\bar{e}_{13} = [-\cos(\alpha-\theta)\bar{e}_a, \sin(\alpha-\theta)\bar{e}_b] \qquad a_\theta = \bar{e}_\theta \cdot \bar{e}_a \qquad b_\theta = \bar{e}_\theta \cdot \bar{e}_b$$

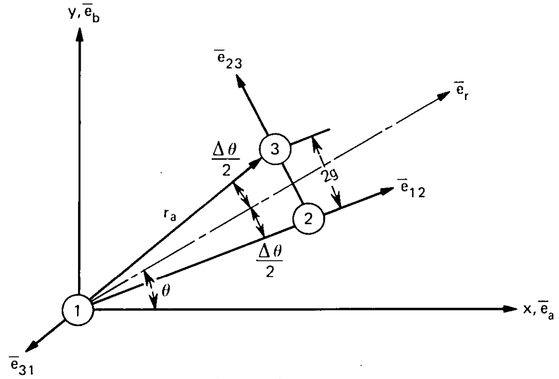
$$\bar{e}_{24} = [-\cos(\alpha+\theta)\bar{e}_a, -\sin(\alpha+\theta)\bar{e}_b] \qquad a_{13} = \bar{e}_{13} \cdot \bar{e}_a \qquad b_{13} = \bar{e}_{13} \cdot \bar{e}_b$$

$$2\ell = [\sin^2\frac{\Delta\theta}{2}(r_0+r_i)^2 + (r_0+r_i)^2]^{1/2} \qquad a_{24} = \bar{e}_{24} \cdot \bar{e}_a \qquad b_{24} = \bar{e}_{24} \cdot \bar{e}_b$$

$$\bar{e}_r = (\cos\theta\,\bar{e}_a, \sin\theta\,\bar{e}_b) \qquad A = (r_0-r_i)(r_0+r_i)\sin\frac{\Delta\theta}{2}\cos\frac{\Delta\theta}{2} \text{ (AREA)}$$

$$\bar{e}_\theta = (-\sin\theta\,\bar{e}_a, \cos\theta\,\bar{e}_b) \qquad 2c = (r_0-r_i)\cos\frac{\Delta\theta}{2} \qquad f = A/4c$$

Fig. 3—Properties of the trapezoidal element.



$$A = r_a^2 \sin\frac{\Delta\theta}{2} \cos\frac{\Delta\theta}{2} \text{ (AREA)}$$

$$2g = r_a \sin \Delta\theta$$

$$\bar{e}_{12} = [\cos(\theta - \frac{\Delta\theta}{2})\,\bar{e}_a, \sin(\theta - \frac{\Delta\theta}{2})\,\bar{e}_b]$$

$$a_{12} = \bar{e}_{12} \cdot \bar{e}_a \quad b_{12} = \bar{e}_{12} \cdot \bar{e}_b$$

$$\bar{e}_{23} = (-\sin\theta\,\bar{e}_a, \cos\theta\,\bar{e}_b)$$

$$a_{23} = \bar{e}_{23} \cdot e_a \quad b_{23} = \bar{e}_{23} \cdot \bar{e}_b$$

$$\bar{e}_{31} = [-\cos(\theta + \frac{\Delta\theta}{2})\bar{e}_a, -\sin(\theta + \frac{\Delta\theta}{2})\bar{e}_b]$$

$$a_{31} = \bar{e}_{31} \cdot \bar{e}_a \quad b_{31} = \bar{e}_{31} \cdot \bar{e}_b$$

Fig. 4—Properties of the triangular element.

III. LONG-CYLINDER METHOD

In this section, the 2π symmetry requirements as to material properties, external loadings, and temperature are lifted and a two-dimensional model is constructed in the plane of the circle. The cylinder is considered long in the z-(longitudinal) direction and material properties, external loadings, and temperatures are assumed independent of z. Shear stresses τ_{yz} and τ_{zx} are assumed to be zero. The most general stress-strain relationship considered is the 4 × 4 submatrix bounded by the dashed lines of eq. (18) (appendix) with the additional proviso of no coupling between the stresses σ_x, σ_y, σ_x, τ_{xy} and the strains γ_{yz}, γ_{zx}. Tentatively, we let ϵ_z vanish during the initial phase of the analysis. This restriction is subsequently removed in a manner similar to that described by Timoshenko[3] for long isotropic cylinders.

The basic building blocks of the long cylinder is the trapezoidal element (Fig. 3) and the isosceles triangular element (Fig. 4, solid cylinders only). The trapezoidal element is subjected to three constant stress fields: a radial stress σ_r, a tangential stress σ_θ, and a shear stress $\tau_{r\theta}$. The triangular element is likewise subjected to three uniform stress fields, each parallel to a side of the triangle. Equilibrium between forces collected at the apices of the elements and three uniform stress fields are obtained as shown below (see Figs. 3 and 4 for the properties of the elements).

Trapezoid:

$$\begin{Bmatrix} F_{x1} \\ F_{y1} \\ F_{x2} \\ F_{y2} \\ F_{x3} \\ F_{y3} \\ F_{x4} \\ F_{y4} \end{Bmatrix} = \begin{bmatrix} fa_r & -ca_\theta & la_{13} \\ fb_r & -cb_\theta & lb_{13} \\ fa_r & ca_\theta & -la_{24} \\ fb_r & cb_\theta & -lb_{24} \\ -fa_r & ca_\theta & -la_{13} \\ -fb_r & cb_\theta & -lb_{13} \\ -fa_r & -ca_\theta & la_{24} \\ -fb_r & -cb_\theta & la_{24} \end{bmatrix} \begin{Bmatrix} \sigma_r \\ \sigma_\theta \\ \tau_{r\theta} \end{Bmatrix}. \qquad (10)$$

Triangle:

$$\begin{Bmatrix} F_{x1} \\ F_{y1} \\ F_{x2} \\ F_{y2} \\ F_{x3} \\ F_{y3} \end{Bmatrix} = \begin{bmatrix} -ga_{12} & 0 & ga_{31} \\ -gb_{12} & 0 & gb_{31} \\ ga_{12} & -r_a/2 & 0 \\ gb_{12} & -r_a/2 & 0 \\ 0 & r_a/2 & -ga_{31} \\ 0 & r_a/2 & -gb_{31} \end{bmatrix} \begin{Bmatrix} \sigma_{12} \\ \sigma_{23} \\ \sigma_{31} \end{Bmatrix}. \qquad (11)$$

The relationship between the orthogonal stress field $(\sigma_r \sigma_\theta \tau_{r\theta})$ and the skew stress field $(\sigma_{12}\ \sigma_{23}\ \sigma_{31})$ is given below for the triangle.

$$\{\sigma_b\} = \begin{Bmatrix} \sigma_r \\ \sigma_\theta \\ \tau_{r\theta} \end{Bmatrix}$$

$$= \begin{bmatrix} \cos^2 \Delta\theta/2 & 0 & \cos^2 \Delta\theta/2 \\ \sin^2 \Delta\theta/2 & 1 & \sin^2 \Delta\theta/2 \\ -\sin \Delta\theta/2 \cos \Delta\theta/2 & 0 & \sin \Delta\theta/2 \cos \Delta\theta/2 \end{bmatrix} \begin{Bmatrix} \sigma_{12} \\ \sigma_{23} \\ \sigma_{31} \end{Bmatrix}. \qquad (11a)$$

After inverting eq. (11a) and substituting it into eq. (11), equilibrium is established between the orthogonal stress field and the forces (denoted as $\{F_n\}$) at the apices of the element or the points of the network. Denote this relationship as:

$$\{F_n\} = [H_n]\{\sigma_b\}. \qquad (12)$$

Similarly, for the trapezoid, eq. (12) will be the shorthand matrix notation for eq. (10).

Conjugate to the forces $\{F_n\}$ are deflections $\{U_n\}$ and therefore the conjugate relationship to eq. (12) can be constructed as

$$\{\epsilon_b\} = 1/A[H_n]^t\{U_n\}, \tag{13}$$

where A is the volume (area times unit thickness) of the trapezoid or triangle.

Upon substituting eq. (13) into eq. (26), the stresses in the element in terms of the unknown deflections at the apices and the known thermal strains are obtained.

$$\{\sigma_b\} = [C]\left(1/A[H_n]^t\{U_n\} - [B_b]\left\{\int \alpha dT\right\}\right). \tag{14}$$

Upon substituting eq. (14) into eq. (12), contributions to the network stiffness matrix and thermal load vector are obtained for the element

$$\{F_n\} = [K_n]\{U_n\} - \{E_n\}, \tag{15}$$

where

$$[K_n] = 1/A[H_n][C][H_n]^t$$

is a symmetric 8 × 8 or 6 × 6 stiffness matrix and

$$\{E_n\} = [H_n][C][E_b]\left\{\int \alpha dT\right\}$$

is an 8 × 1 or 6 × 1 thermal load vector.

Contributions of each element to the network stiffness matrix and thermal load vector are additive, and hence eq. (7a) can represent the force balance at all the points of the long-cylinder network as well as those of the solid-of-revolution network.

Noting the material restrictions outlined in the beginning of this section, a network of triangles and trapezoids need only occupy one-quarter of a circle with planes of symmetry at $\theta = 0$ and $\pi/2$. Hence, tangential deflections must vanish at all points along these planes. The previously unknown deflections and constraint forces can now be solved in terms of known external forces and thermal loads in the same manner as was accomplished in the preceding section [see eqs. (8), (9), and (9a)].

We are not quite finished. Recall that we have let $\epsilon_z = 0$ throughout the cylinder, resulting in an axial stress σ_z [eq. (23)] applied to the ends of the cylinder. If we superimpose a uniform axial stress $\langle\sigma_z\rangle$ such that the resulting force on the ends of the cylinder is zero, the self-equilibrating distribution remaining on the ends will, by St. Venant's principle,[3] give rise only to local effects at the ends. The uniform axial stress correction to be added to eq. (23) is given as

$$\langle \sigma_z \rangle = - \frac{\int \sigma_z \,[\text{eq. } (23)] dA}{\int dA}. \tag{16}$$

This correction results in added radial and tangential strains obtained from eq. (22).

$$\langle \epsilon_r(\theta) \rangle = E_{13}(\theta) \langle \sigma_z \rangle = (E_{xz} \cos^2\theta + E_{yz} \sin^2\theta) \langle \sigma_z \rangle \tag{16a}$$

$$\langle \epsilon_\theta(\theta) \rangle = E_{23}(\theta) \langle \sigma_z \rangle = (E_{xz} \sin^2\theta + E_{yz} \cos^2\theta) \langle \sigma_z \rangle. \tag{16b}$$

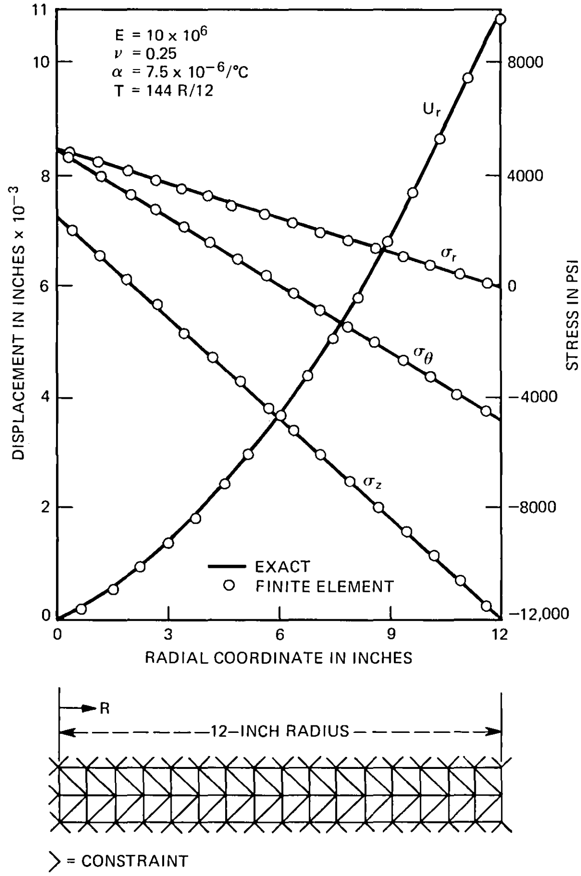


Fig. 5—Thermal stresses plane strain problem.

The resulting correction to the radial and tangential deflections are next obtained.

$$\langle U_r(r,\theta)\rangle = r\langle \epsilon_r(\theta)\rangle = r(E_{xy}\cos^2\theta + E_{yz}\sin^2\theta)\langle\sigma_z\rangle \tag{17}$$

$$\langle U_\theta(r,\theta)\rangle = \int_0^\theta (r\langle\epsilon_\theta(\theta)\rangle - \langle U_r(r,\theta)\rangle)d\theta$$

$$= r\int^\theta (E_{23}(\theta) - E_{13}(\theta))\langle\sigma_z\rangle d\theta$$

$$= r\sin\theta\cos\theta\,(E_{yz} - E_{xz}). \tag{17a}$$

Note that the correction $\langle U_\theta(r,\theta)\rangle$ vanishes at $\theta = 0$ and $\theta = \pi/2$, agreeing with the stipulated boundary conditions stated previously.

IV. COMPARISONS WITH THEORETICAL SOLUTIONS

Two plane strain problems are presented for the solid-of-revolution method: one, a linear radial thermal gradient applied to an isotropic solid cylinder (Fig. 5) and two, a negative pressure applied to the outside circumference of a hollow cylinder (Fig. 6). Comparisons were

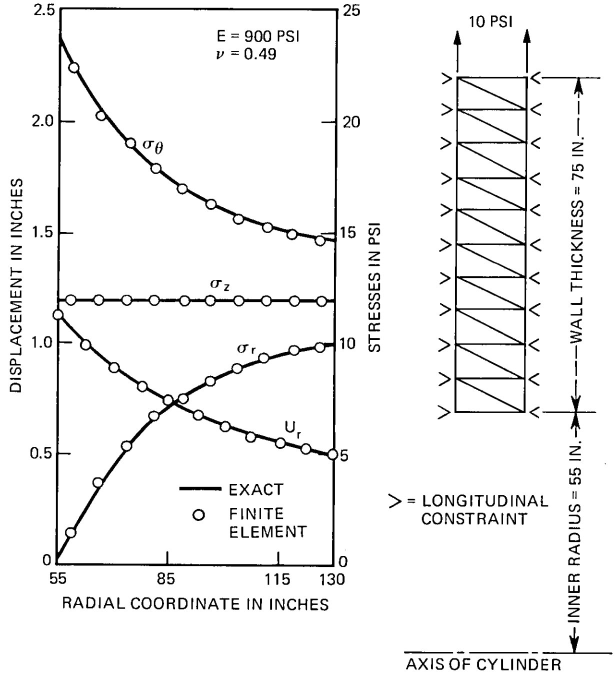


Fig. 6—Plane strain solution—thick-walled cylinder.

made of the finite element computer results with known theoretical[3] solutions. The agreement was excellent.

A thermal problem is presented for the long-cylinder method in which an isotropic cylinder is subjected to a constant temperature plus a radial thermal gradient (Fig. 7). Comparisons were made of the finite-element computer results with a known theoretical solution.[3] The known solution on page 410 of Ref. 3 was found to be in error. In the expression for radial displacement (U), insert $(1 - 3\nu)/(1 - \nu)$ for $(1 - 2\nu)$ and in the expression for σ_z, replace $2r$ with 2. After the above corrections were made to the theoretical solution, excellent comparisons resulted as shown in Fig. 7.

V. THERMAL STRESSES IN LITHIUM TANTALATE CRYSTALS

Thermal stresses were computed for a lithium tantalate crystal, class C_{3v}, during the post-growth cooling stage. The crystal was analyzed by the two methods presented in this paper on the IBM 370 computer. For the solid-of-revolution method (Fig. 8), the model con-

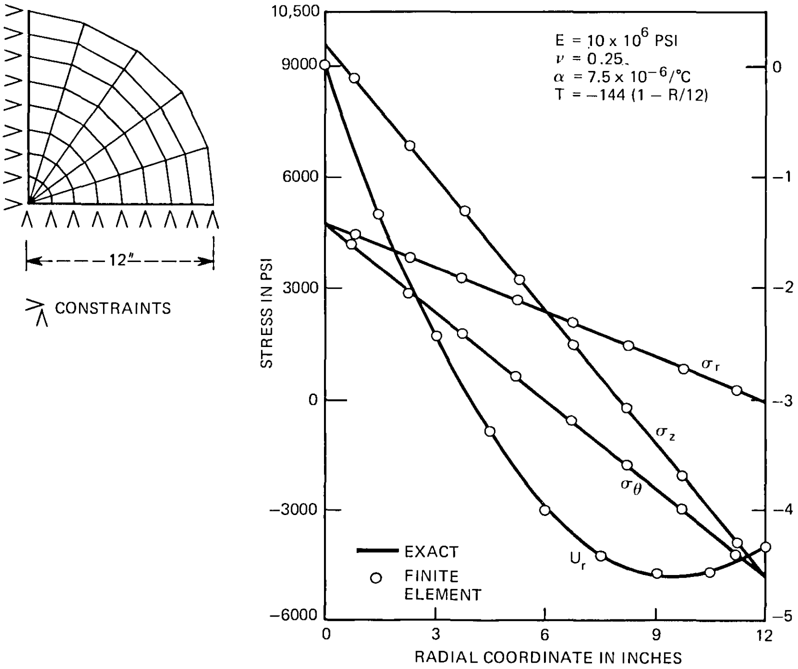


Fig. 7—Thermal stresses—long cylinder.

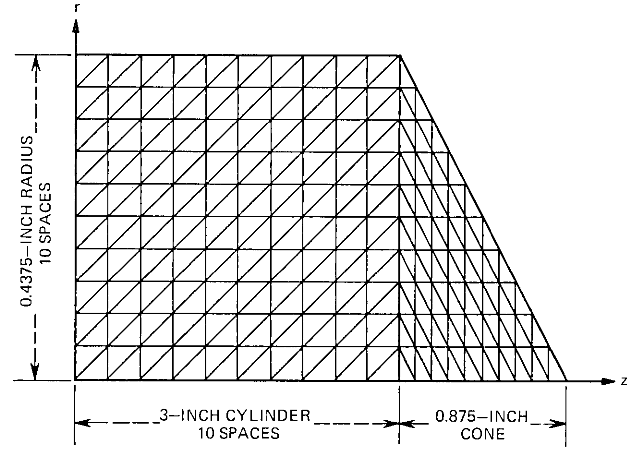


Fig. 8—Network of triangular annuli—lithium tantalate crystal—solid-of-revolution model.

sists of a right cylinder of radius $a = 0.4375$ inch and length $b = 3$ inches plus a cone of length 0.875 inch attached to the far (cold) end. Radial constraints are provided at all zero radius points. For the long-cylinder method (Fig. 9), the model comprises one-quarter of a circle with tangential constraints at all points along the two planes of symmetry ($\theta = 0$ and $\pi/2$).

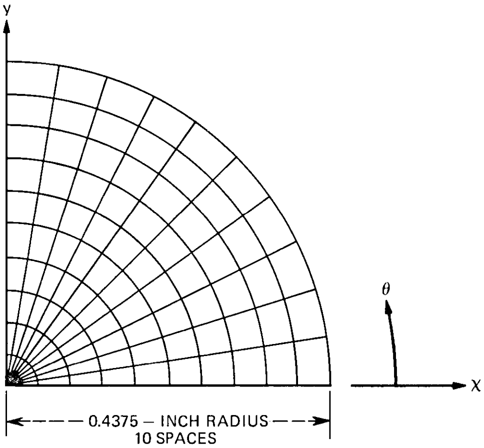


Fig. 9—Network of triangles and trapezoids—lithium tantalate crystal—long-cylinder model.

Elastic properties[4] of lithium tantalate are shown in Fig. 10. Axis 1 is perpendicular to the vertical mirror plane and axis 3 corresponds to the trigonal axis of the crystal. If we strike out the small shear coupling terms, c_{14}, in Fig. 10, then we have met the stress-strain relationship [eqs. (5) and (18)] requirements of this paper. The coefficients of thermal expansion will be assumed constant in the 1, 2 directions and piecewise linear along the trigonal axis 3. These coefficients[5] are given below.

$$\begin{aligned}
\alpha_1 = \alpha_2 &= 21.9 \times 10^{-6}/^\circ\mathrm{C} & & \\
\alpha_3 &= 8.35 \times 10^{-6}/^\circ\mathrm{C} & & 625^\circ\mathrm{C} \text{ and higher} \\
&\quad -3.1 \times 10^{-6}/^\circ\mathrm{C} & & 500^\circ\mathrm{C} - 625^\circ\mathrm{C} \\
&\quad 0 & & 400^\circ\mathrm{C} - 500^\circ\mathrm{C} \\
&\quad 1.16 \times 10^{-6}/^\circ\mathrm{C} & & \text{room temperature} - 400^\circ\mathrm{C}.
\end{aligned}$$

The crystal is assumed to be in a stress-free condition at an elevated temperature distribution $T1$, and elastic material behavior is assumed linear from the initial state of strain at temperature distribution $T1$ to the final state of stress and strain at room temperature (26°C). This assumption rules out plasticity and stress relaxation. It is felt that ignoring these nonlinear effects plus ignoring the initial state of stress at $T1$ does not distort the qualitative picture of the stress pattern after the cooling-down process. The elevated temperature distribution in degrees centigrade is given below.

$$\begin{aligned}
T1 &= 1500 - 100\,(r/a)^2 - 500\,z/b, \text{ for solid of revolution} \\
&= 1500 - 100\,(r/a)^2, \text{ for the long cylinder.}
\end{aligned}$$

The addition of the longitudinal gradient for the solid-of-revolution method was found to induce negligible stress in the crystal.

For the solid-of-revolution method (z-growth), material axes 1, 2, and 3 correspond to r, θ, and z, respectively. For the long-cylinder method, two cases were investigated: one, z-growth where material axes 1, 2, and 3 correspond to x, y, and z, respectively, and two, y-growth where material axes 1, 2, and 3 correspond to z, x, and y,

$$\left\{\begin{matrix}\sigma_1\\ \sigma_2\\ \sigma_3\\ \tau_{23}\\ \tau_{31}\\ \tau_{12}\end{matrix}\right\} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & \xcancel{c_{14}} & 0 & 0\\ c_{12} & c_{11} & c_{13} & \xcancel{-c_{14}} & 0 & 0\\ c_{13} & c_{13} & c_{33} & 0 & 0 & 0\\ \xcancel{c_{14}} & \xcancel{-c_{14}} & 0 & c_{44} & 0 & 0\\ 0 & 0 & 0 & 0 & c_{44} & \xcancel{c_{14}}\\ 0 & 0 & 0 & 0 & \xcancel{c_{14}} & c_{66}\end{bmatrix} \left\{\begin{matrix}\epsilon_1\\ \epsilon_2\\ \epsilon_3\\ \gamma_{23}\\ \gamma_{31}\\ \gamma_{12}\end{matrix}\right\} \qquad \begin{aligned} c_{33} &= 2.75 \times 10^{11}\ \mathrm{N/m^2}\\ c_{11} &= 2.33\\ c_{14} &= 0.11\\ c_{12} &= 0.47\\ c_{13} &= 0.80\\ c_{44} &= 0.94\\ c_{66} &= 0.93 = \tfrac{1}{2}(c_{11} - c_{12})\end{aligned}$$

Fig 10—$LiTaO_3$ stress-strain.

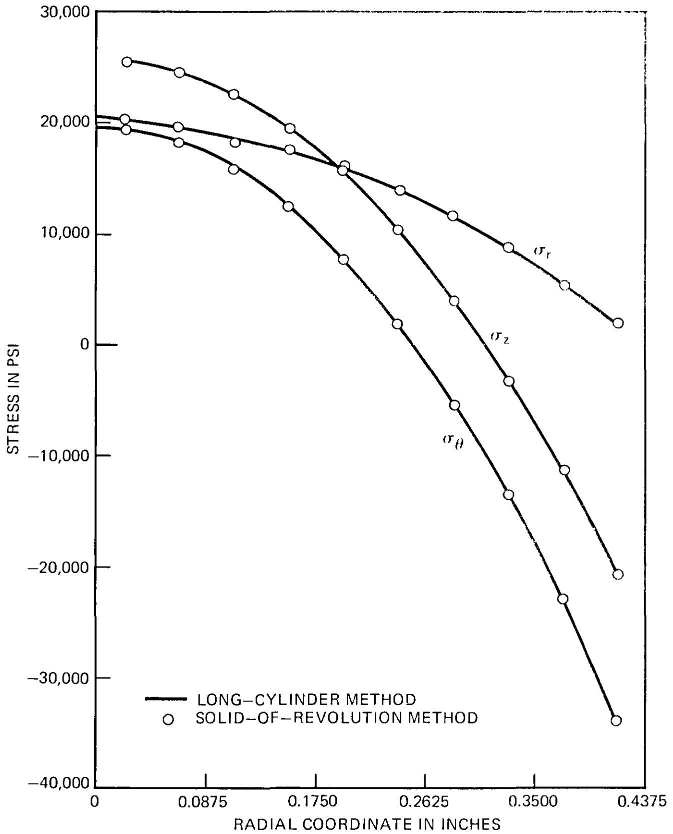


Fig. 11—Thermal stress comparison z-growth $LiTaO_3$ crystal.

respectively. A plot of thermal stresses for the z-growth crystals by both methods is compared in Fig. 11. Agreement is very good. A comparison of thermal stresses by the long-cylinder method of both z-growth and y-growth is shown in Fig. 12 at $\theta = 45°$. (Typically, θ-stress variations are less than 5 percent.) The maximum stress recorded in absolute value is about 39,000 psi for both z-growth and y-growth. The x-growth (material axis 3 corresponds to x) would yield the same results as y-growth. This can be inferred from examination of the material properties of Fig. 10.

VI. CONCLUSIONS

Excellent comparisons were obtained for stresses and deflection in isotropic cylinders between the methods described in this paper and known theoretical solutions (Figs. 5 to 7). Stress comparisons again

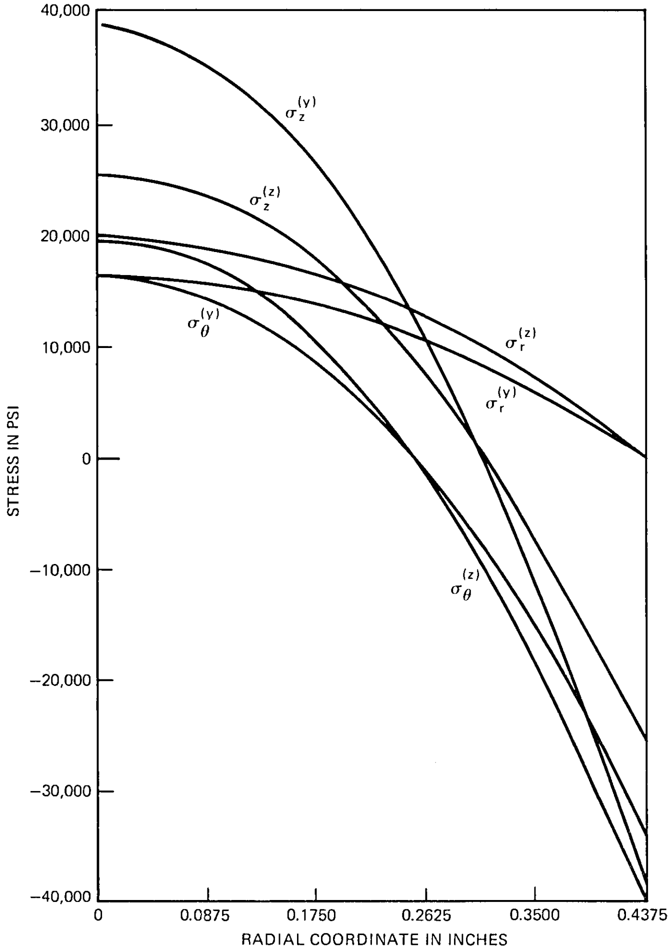


Fig. 12—Thermal stress comparison z-growth and y-growth (at $\theta = 45°$) $LiTaO_3$ crystals—long-cylinder method.

were excellent between the two methods described in this paper for z-growth $LiTaO_3$ crystals (isotropic in the plane of the circle) during the post-growth cooling stage (Fig. 11). Comparisons of thermal stresses between a y-(or x-) growth $LiTaO_3$ crystal (both anisotropic in the plane of the circle) and a z-growth crystal are presented (Fig. 12). Differences in the stress distributions were not great enough to favor either growth direction; the more important consideration is to slow the cooling process down enough to keep radial thermal gradient to a minimum.

VII. ACKNOWLEDGMENT

The author gratefully acknowledges B. E. Nevis for his critical review of this paper and his assistance in estimating rational thermal gradients during the cooling stage of the lithium tantalate crystal.

APPENDIX

Stress-Strain Relationships—Long Cylinder

The most general stress-strain relationship considered is the 4 × 4 submatrix bounded by the dotted lines of eq. (18) as shown below, with no coupling between the stresses (σ_x, σ_y, σ_z, τ_{xy}) and the strains (γ_{yz}, γ_{zx}).

$$\begin{Bmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ \tau_{xy} \\ \\ \tau_{yz} \\ \tau_{zx} \end{Bmatrix} = \left[\begin{array}{cccc|cc} C_{xx} & C_{xy} & C_{xz} & 0 & 0 & 0 \\ C_{xy} & C_{yy} & C_{yz} & 0 & 0 & 0 \\ C_{xz} & C_{yz} & C_{zz} & 0 & 0 & 0 \\ 0 & 0 & 0 & C_{ss} & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & & \\ 0 & 0 & 0 & 0 & & \end{array}\right] \begin{Bmatrix} \epsilon_x \\ \epsilon_y \\ \epsilon_z \\ \gamma_{xy} \\ \\ \gamma_{yz} \\ \gamma_{zx} \end{Bmatrix}. \tag{18}$$

After inclusion of thermal strains, eq. (18) can be restated as

$$\{\sigma_0\} = [C_0]\left(\{\epsilon_0\} - \left\{\int \alpha dT\right\}\right), \tag{19}$$

where

$$\{\sigma_0\} = \{\sigma_x \; \sigma_y \; \sigma_z \; \tau_{xy}\},$$
$$\{\epsilon_0\} = \{\epsilon_x \; \epsilon_y \; \epsilon_z \; \gamma_{xy}\},$$
$$\left\{\int \alpha dT\right\} = \left\{\int \alpha_z dT \; \int \alpha_y dT \; \int \alpha_z dT \;\; 0\right\},$$

and $[C_0]$ is the 4 × 4 submatrix of eq. (18).

Now consider a rotation about z by the angle θ, and let $\{\sigma_a\} = \{\sigma_r \; \sigma_\theta \; \sigma_z \; \tau_{r\theta}\}$ be the radial, tangential, longitudinal, and shear stress, respectively. The relationship between this rotated stress field and $\{\sigma_0\}$ can be expressed as

$$\{\sigma_a\} = [B]\{\sigma_0\}, \tag{20}$$

where

$$[B] = \begin{bmatrix} \cos^2\theta & \sin^2\theta & 0 & 2\sin\theta\cos\theta \\ \sin^2\theta & \cos^2\theta & 0 & -2\sin\theta\cos\theta \\ 0 & 0 & 1 & 0 \\ -\sin\theta\cos\theta & \sin\theta\cos\theta & 0 & \cos^2\theta - \sin^2\theta \end{bmatrix}.$$

Conjugate to eq. (20) is the following relationship between the strains

$$\{\epsilon_0\} = [B]^t\{\epsilon_a\}, \tag{20a}$$

where

$$\{\epsilon_a\} = \{\epsilon_r\ \epsilon_\theta\ \epsilon_z\ \gamma_{r\theta}\}.$$

Eqs. (19), (20), and (20a) can be combined as

$$\{\sigma_a\} = [C_a]\{\epsilon_a\} - [B][C_0]\{\alpha dT\}, \tag{21}$$

where

$$[C_a] = [B][C_0][B]^t.$$

After multiplying eq. (21) by $[C_a]^{-1}$ and rearranging, the following result is obtained.

$$\{\epsilon_a\} = [E]\{\sigma_a\} + ([B]^t)^{-1}\left\{\int \alpha dT\right\}, \tag{22}$$

where

$$[E] = [C_a]^{-1} = ([B]^t)^{-1}[C_0]^{-1}[B]^{-1}. \tag{22a}$$

Tentatively, let $\epsilon_z = 0$. This results in a longitudinal stress applied at the ends of the cylinder. From the third line of eq. (22), the longitudinal stress can be obtained as

$$\sigma_z = 1/E_{33}\left(E_{31}\,\sigma_r + E_{32}\,\sigma_\theta + E_{34}\,\tau_{r\theta} + \int \alpha_z dT\right), \tag{23}$$

where, from eq. (22a),

$$\begin{aligned}
E_{31} &= E_{xz}\cos^2\theta + E_{yz}\sin^2\theta,\\
E_{32} &= E_{xz}\sin^2\theta + E_{yz}\cos^2\theta,\\
E_{33} &= E_{zz},\\
E_{34} &= 2\sin\theta\cos\theta\,(E_{yz} - E_{xz}),
\end{aligned}$$

and E_{xz}, E_{yz}, and E_{zz} are obtained from $[C_0]^{-1}$.

From eq. (23) we obtain

$$\{\sigma_a\} = [D]\{\sigma_b\} - \left\{0\ 0\ \ 1/E_{33}\int \alpha_z dT\ \ 0\right\}, \tag{24}$$

where

$$\{\sigma_b\} = \{\sigma_r\ \sigma_\theta\ \tau_{r\theta}\}$$

and

$$[D] = \left\{\begin{matrix} 1 & 0 & 0\\ 0 & 1 & 0\\ -E_{31}/E_{32} & -E_{32}/E_{33} & -E_{34}/E_{33}\\ 0 & 0 & 1 \end{matrix}\right\}.$$

After substituting eq. (24) into eq. (22) and premultiplying by $[D]^t$ (remember $\epsilon_z = 0$), we obtain

$$\{\epsilon_b\} = [E_b]\{\sigma_b\} + [B_b]\left\{\int \alpha dT\right\}, \tag{25}$$

where

$$\{\epsilon_b\} = \{\epsilon_r\ \epsilon_\theta\ \gamma_{r\theta}\},$$
$$[E_b] = [D]^t[E][D]$$

and

$$\begin{aligned}[B_b] &= [D]^t([B]^t)^{-1} \\ &= \left\{\begin{matrix} \cos^2\theta & \sin^2\theta & -E_{31}/E_{33} & \sin\theta\cos\theta \\ \sin^2\theta & \cos^2\theta & -E_{32}/E_{33} & -\sin\theta\cos\theta \\ -2\sin\theta\cos\theta & 2\sin\theta\cos\theta & -E_{34}/E_{33} & \cos^2\theta - \sin^2\theta \end{matrix}\right\}.\end{aligned}$$

Let $[C] = [E_b]^{-1}$. After inverting eq. (25) we obtain a desired result.

$$\{\sigma_b\} = [C]\left(\{\epsilon_b\} - [B_b]\left\{\int \alpha dT\right\}\right) \tag{26}$$

subject to the restriction $\epsilon_z = 0$, which will be removed after an initial solution is obtained.

REFERENCES

1. R. W. Clough, F. Asce, and Y. Rashid, "Finite Element Analysis of Axi-Symmetric Solids," JEMD of ASME, *EM1*, 1965, pp. 71–85.
2. E. L. Wilson, "Structural Analysis of Axisymmetric Solids," JAIAA, 1965, pp. 2269–2274.
3. S. P. Timoshenko and J. N. Goodier, *Theory of Elasticity*, 2nd Edition, New York: McGraw-Hill 1951, pp. 60, 65–66, 408–410.
4. A. W. Warner, M. Onoe, and G. A. Coquin, "Determination of Elastic and Piezoelectric Constants for Crystals in Class (3m)," J. Acoust. Soc. Amer., *6*, 1967, pp. 1223–1231.
5. H. Iwasaki, S. Miyagawa, T. Yamada, N. Uchida, and N. Niizeki, "Single Crystal Growth and Physical Properties of LiTaO$_3$," Rev. Elec. Commun. Laboratories (Japan), *20*, January–Feburary, 1972, pp. 129–137.

Copyright © 1974 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 5, May–June 1974
Printed in U.S.A.

Mean-Squared-Error Equalization Using Manually Adjusted Equalizers

By Y.-S. CHO

(Manuscript received October 19, 1973)

This paper describes an equalization procedure for systems using manually adjustable bump equalizers that is based on a mean-squared error criterion. We show that, in accordance with a Gauss-Seidel iteration process, gain adjustment always converges to the optimal value at which the minimum MSE of the equalized channel is obtained. Both zero forcing and MSE algorithms based on the Gauss-Seidel iteration method are derived, and hardware implementation of these algorithms is discussed. According to the error reduction analysis, an equalizer composed of orthogonal networks requires only one iteration to bring the equalizer to the optimum state. For the bump equalizers used in the latest L5 Coaxial Carrier Transmission System whose Bode networks are semi-orthogonal, two to three iterations are shown to be sufficient to achieve the optimum gain settings in the mean-squared error sense.

I. INTRODUCTION

In this paper, a manual adjustment of bump equalizers is described which uses a mean-squared error (MSE) criterion. In the existing L4 Coaxial Transmission System, the bump equalizers (realized with Bode equalizer networks[1]) are used for line equalization and adjusted according to a zero forcing (ZF) algorithm.[2] This method results in an optimum equalization in the MSE sense, but only under certain very restrictive conditions with respect to the transmission response of the channel. The latest L5 Coaxial Transmission System, which provides up to 10,800 toll-grade long-haul message channels on a pair of 0.375-inch coaxial cables over 4000 miles, also includes bump equalizers for equalization of the 65-MHz bandwidth channel. The equalization method used in the L5 Coaxial Carrier Transmission System, however, minimizes the MSE of the channel deviation. It is found in practice that the MSE algorithm has given a better equalization result than the ZF algorithm.

Since a coaxial cable system shows relatively stable channel characteristic, the bulk of the L5 line equalization is accomplished by manually adjustable equalizers. Normally, the time-varying channel deviation in a cable system is mostly the result of seasonal temperature variation and aging of the components in the system. In such case, a usage of complex automatic equalizers in the system is not economically desirable.

In Ref. 3, two MSE methods are discussed for the equalizer adjustment. Both methods are based on the steepest descent algorithm and could be easily implemented in an automatic equalizer, but not in a manual equalizer.

In this paper, an MSE algorithm based on the Gauss-Seidel iteration method is described for the gain adjustment of the manual bump equalizers. Under the specified assumptions, this method guarantees the convergence of the following iterative process. If a visual display of the gradient of the MSE with respect to each gain setting is available, adjust the first gain setting until the gradient becomes zero; next, adjust the second gain setting until the associated gradient becomes zero; similarly, adjust the third gain setting and all others up to the last one, thus completing one iteration. As the number of iterations increases, the residual error in the channel will be minimized in the MSE sense.

While the Gauss-Seidel iteration method may seem quite complicated, it has several distinct advantages. First of all, the Gauss-Seidel iteration method requires only one gradient at a time, which can simplify the hardware involvement, particularly for manual equalizers. As is shown in Section III, the number of iterations needed to bring the equalizers to the optimum state is not large. When the equalizer is composed of orthogonal networks, a single iteration is sufficient. Since most of the equalizer networks used in transmission systems are orthogonal or semi-orthogonal in nature, the number of iterations will usually be small. For the bump equalizers, which consist of semi-orthogonal terms, two to three iterations are satisfactory for the optimal equalization according to the error analysis given in Section III. This result has been verified experimentally in the field.

II. MSE ADJUSTING ALGORITHMS FOR BUMP EQUALIZERS

In this section, several assumptions are made before the ZF and MSE algorithms are presented.

2.1 Characterization of coaxial channel

The channel assumed in this section is represented by an infinite sin x/x series on the frequency domain. (Since the transfer function of the Bode network is symmetric on the log f plane, where f is the natural frequency in hertz, the frequency used throughout is defined by $w = \log f$.)

Let $M(w, t)$ represent the time-varying channel misalignment which is a real valued function of frequency in decibels. From the practical point of view, however, the channel can be assumed to be simply $M(w)$ since the time variation is negligible during the equalization interval. Further, assume that the Fourier transform of the channel is limited in the time domain by a certain positive constant. (It should be noted that the Fourier transform of the channel does not result in an impulse response of the channel because the channel is measured in dB. However, there is an implicit dual relationship between the channel studied in this paper and that involving a time-domain equalizer, e.g., a transversal equalizer, in which a frequency band limitation of the channel is implied.) Hence, the channel can be characterized on the frequency domain by the following series:

$$M(w) = \sum_{n=0}^{\infty} C_n \frac{\sin[2\pi p(w - w_n)]}{2\pi p(w - w_n)} \qquad (\text{dB}), \tag{1}$$

where C_n, p, and w_n are real numbers and

$$w_{n+1} - w_n = \frac{1}{2p} \quad \text{for all } n = 0, 1, \cdots.$$

Note that w_n is equally spaced.

Equation (1) can also be written in the following way.

$$\begin{aligned} M(w) &= \int_0^1 \sum_{n=0}^{\infty} C_n \cos[2\pi p(w - w_n)x]dx \\ &= \int_0^1 \left\{ \sum_{n=0}^{\infty} C_n \cos(2\pi p w_n x) \cos(2\pi p w x) \right. \\ &\qquad\qquad \left. + \sum_{n=0}^{\infty} C_n \sin(2\pi p w_n x) \sin(2\pi p w x) \right\} dx \\ &= \int_0^1 \{F(x) \cos(2\pi p w x) + H(x) \sin(2\pi p w x)\} dx, \end{aligned} \tag{2}$$

where

$$F(x) = \sum_{n=0}^{\infty} C_n \cos(2\pi p w_n x) \quad \text{and} \quad H(x) = \sum_{n=0}^{\infty} C_n \sin(2\pi p w_n x).$$

Since $0 \leqq x \leqq 1$, eq. (2) implies that the shortest frequency domain ripple period found in the channel $M(w)$ is $1/p$.

2.2 Representation of bump equalizers

The bump equalizer considered in this paper is a linear combination of adjustable-loss Bode networks. The input-output transfer function of an equalizer composed of N Bode networks can be represented by

$$\mathrm{EQL}(w) = \sum_{k=1}^{N} g_k B_k(w) \qquad \text{(dB)}, \tag{3}$$

where N is the number of networks and g_k and B_k represent the gain and response respectively of the kth Bode network.

A typical Bode network is shown in Fig. 1a, where the loss is controlled by the resistor R. The transfer function, $B_k(w)$, can be analytically derived and, with a suitable flat gain amplifier, it can be expressed by the following equation:

$$B_k(w) = \frac{[E_k(1 + E_k) + D_k]^2 - D_k}{[(1 + E_k)^2 + D_k]^2} \qquad \text{(dB)}, \tag{4}$$

where

$$E_k = \frac{R_{0k}}{R_{1k}},$$

$$D_k = \frac{(w/w_k)^2 H_k}{[(w/w_k)^2 - 1]^2},$$

$$H_k = \frac{C_k}{L_k},$$

and

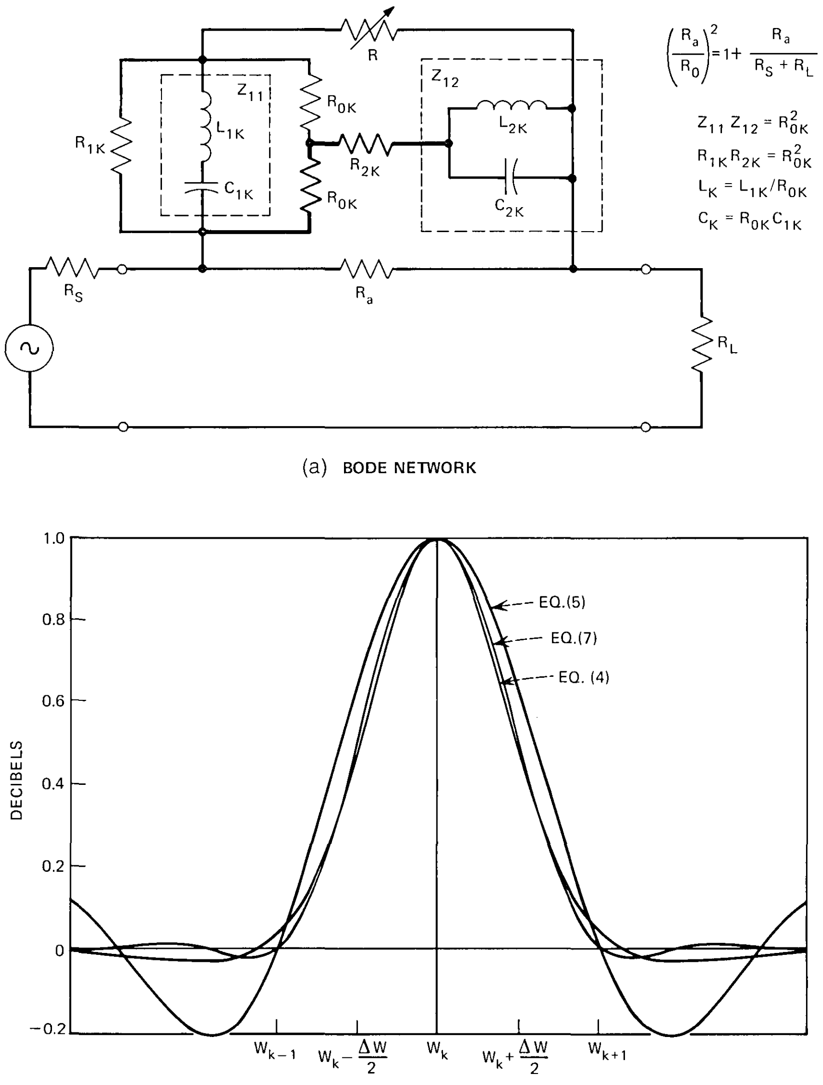
$$w_k = \log \frac{1}{2\pi\sqrt{L_k C_k}}.$$

Since eq. (4) shows $B_k(w)$ to be a quite complicated function of w, one of the following assumptions is used while analyzing the equalizer in detail.

Assumption 1: Let $B_k(w)$ be approximated by

$$\mathrm{sinc}\left[\frac{\pi}{\Delta w}(w - w_k)\right] = \frac{\sin[\pi(w - w_k)/\Delta w]}{\pi(w - w_k)/\Delta w} \qquad \text{(dB)}. \tag{5}$$

Since there are N Bode networks in the equalizer, which for the



(a) BODE NETWORK

(b) TRANSFER CHARACTERISTICS OF BODE NETWORK

Fig. 1—Adjustable Bode network.

present analysis are spaced equally on the w-axis with interval Δw (see Fig. 2), then the transfer function of equalizer can be expressed by

$$\mathrm{EQL}(w) = \sum_{k=1}^{N} g_k \operatorname{sinc}\left[\frac{\pi}{\Delta w}(w - w_k) \right] \qquad \text{(dB)}. \tag{6}$$

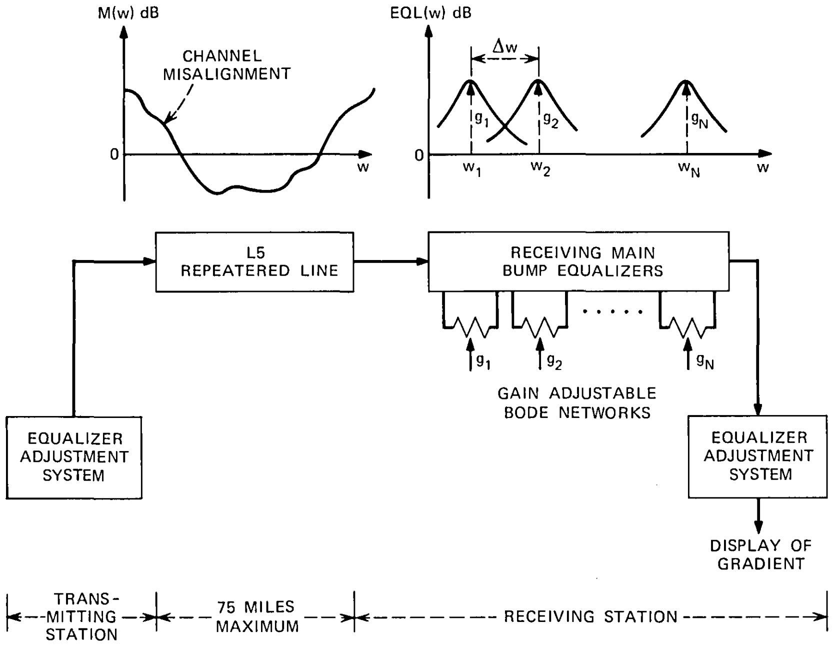


Fig. 2—Manual equalization of L5 coaxial system.

To permit a comparison between $B_k(w)$, as represented by eqs. (4) and (5), the two equations are plotted in Fig. 1b. The maximum differences between the two best matched curves are 0.165 and 0.183 dB when $|w - w_k| \leqq \Delta w$ and $|w - w_k| > \Delta w$, respectively.

Assumption 2: Let $B_k(w)$ be approximated by

$$\text{cosinc}\left[\frac{\pi}{\Delta w}(w - w_k)\right] = \frac{\sin[\pi(w - w_k)/\Delta w]}{\pi(w - w_k)/\Delta w} \cdot \frac{\cos[\pi(w - w_k)/\Delta w}{1 - 4[(w - w_k)/\Delta w]^2}. \qquad (7)$$

Under the same conditions listed in Assumption 1, the transfer function of equalizer can be expressed by

$$\text{EQL}(w) = \sum_{k=1}^{N} g_k \,\text{cosinc}\left[\frac{\pi}{\Delta w}(w - w_k)\right] \qquad (\text{dB}). \qquad (8)$$

Expression (7) is also plotted in Fig. 1b, and it can be seen that cosinc $[\pi(w - w_k)/\Delta w]$ approximates quite well the actual transfer function of the Bode network as expressed by eq. (4). The maximum difference between the two best-matched curves is 0.0327 dB when $|w - w_k| \leqq \Delta w$ and 0.0404 dB for $|w - w_k| > \Delta w$.

2.3 Mean-squared error algorithms (Gauss-Seidel iteration method)

The definition of optimal equalization used in this paper is the minimization of the MSE of the equalized channel as a result of adjusting the gain parameters, g_k.

On a decibel scale, the equalized error will be

$$E(w) = \sum_{k=1}^{N} g_k B_k(w) - M(w). \tag{9}$$

Then the MSE can be represented in the frequency domain by

$$\text{MSE} = \int_{-\infty}^{\infty} |E(w)|^2 dw. \tag{10}$$

Theorem 1: If the equalizer described by eq. (3) is composed of linearly independent networks, then there exists a unique set of g_k's which nulls all the gradients G_k; i.e.,

$$G_k = \frac{\partial MSE}{\partial g_k} = 0 \quad \textit{for all } k = 1, 2, \cdots, N, \tag{11}$$

where MSE is defined in eq. (10), and the corresponding set of g_k's results in minimum MSE.

The proof is given in Appendix A.

The bump equalizers considered in this paper belong to the class of the equalizers defined in Theorem 1.

As derived in eq. (25) of Appendix A, the gradient vector is given by

$$\mathbf{G} = \mathbf{Bg} - \mathbf{M}, \tag{12}$$

where **B**, **g**, and **M** are the system matrix, gain vector, and correlation vector, respectively, and are defined as follows. Defining an inner product

$$\langle A, B \rangle \equiv \int_{-\infty}^{\infty} A(w)B(w)dw,$$

$$\mathbf{G} = [G_1, G_2, \cdots, G_N]^T,$$

where T indicates the transpose,

$$\mathbf{g} = [g_1, g_2, \cdots, g_N]^T,$$
$$\mathbf{M} = 2[\langle B_1(w), M(w)\rangle, \langle B_2(w), M(w)\rangle, \cdots, \langle B_N(w), M(w)\rangle]^T,$$

and

$$\mathbf{B} = 2\begin{bmatrix} \langle B_1, B_1\rangle, \langle B_1, B_2\rangle, \cdots & \langle B_1, B_N\rangle \\ \langle B_2, B_1\rangle, \langle B_2, B_2\rangle, \cdots & \langle B_2, B_N\rangle \\ \vdots & \\ \langle B_N, B_1\rangle, \langle B_N, B_2\rangle, \cdots & \langle B_N, B_N\rangle \end{bmatrix}.$$

Now the gain vector is

$$\mathbf{g} = \mathbf{B}^{-1}(\mathbf{G} + \mathbf{M}), \tag{13}$$

provided that $\mathbf{B}^{-1}$ exists. The optimum gain, $\mathbf{g}^*$, which results in the minimum MSE is obtained by solving eq. (13) with $\mathbf{G} = \mathbf{0}$. The MSE algorithm given in Ref. 3 solves (13) with $\mathbf{G} = \mathbf{0}$ by the steepest descent method, which can be readily implemented in an automatic equalizer control circuit. For manually adjusted equalizers, however, the steepest descent algorithm cannot be easily implemented because the algorithm requires simultaneous adjustment of all the gain settings. A manual equalizer adjustment algorithm should have the following properties:

(*i*) The several gain settings can be adjusted one at a time, within a specified sequence.

(*ii*) Repeating step (*i*), $\mathbf{g}$ approaches $\mathbf{g}^*$.

The converging rate of the initial $\mathbf{g}$ to the final $\mathbf{g}^*$ depends on the type of algorithm used and the system matrix $\mathbf{B}$. For the bump equalizer, this question is discussed in Section 3.2.

Theorem 2 (Gauss-Seidel iteration algorithm): If the system matrix $\mathbf{B}$ *in (12) has dominant diagonal elements such that*

$$|\langle B_k, B_k\rangle| > \sum_{j=1}^{N}{}' |\langle B_k, B_j\rangle| \tag{14}$$

$$\textit{for all } k = 1, 2, \cdots, N,$$

where $\sum'$ *indicates the summation of all terms excluding the case* $j = k$, *then every* $\mathbf{g}$ *converges to the optimum gain* $\mathbf{g}^* = \mathbf{B}^{-1}\mathbf{M}$ *by the following iteration process:*

Iteration 1: Let $g_{k(i)}$ *indicate the kth gain at the ith iteration; thus, the initial gain settings are* $g_{1(0)}, g_{2(0)}, \cdots, g_{N(0)}$. *Adjust* $g_{1(0)}$ *until its corresponding gradient* $G_1 = 0$ *and designate the resultant gain* $g_{1(1)}$. *The gain settings are then* $g_{1(1)}, g_{2(0)}, \cdots, g_{N(0)}$. *Adjust* $g_{2(0)}$ *until the gradient* $G_2 = 0$, *resulting in the gain settings* $g_{1(1)}, g_{2(1)}, g_{3(0)}, \cdots, g_{N(0)}$. *Repeating the operation for each setting results in* $g_{1(1)}, g_{2(1)}, g_{3(1)}, \cdots, g_{N(1)}$, *and completes the first iteration.*

Iteration 2: Adjust $g_{1(1)}$ until $G_1 = 0$, resulting in the gain settings $g_{1(2)}$, $g_{2(1)}$, $g_{3(1)}$, $\cdots$, $g_{N(1)}$. Obtain $g_{2(2)}$, $g_{3(2)}$, $\cdots$, $g_{N(2)}$ by similar operation, completing the second iteration. Similarly, iterations 3, 4, $\cdots$, n can be carried out as required.

The proof is given in Appendix B.

If equalizer networks satisfy the inequality (14), they are called, in this paper, semi-orthogonal terms. The bump equalizers defined in this paper satisfy the inequality (14), and hence Theorem 2 can be used as a manual-equalizer adjusting algorithm. To implement this algorithm, a visual display of each gradient is required before the corresponding gain is adjusted. The following two theorems provide simple ways of determining the gradient.

Theorem 3 (ZF algorithm): Let the channel be represented by eq. (1) and the equalizer satisfy assumption 1. If the interval Δw between two adjacent Bode networks is no greater than half the shortest ripple period ($= 1/p$) found in the channel, i.e.,

$$\Delta w \leqq \frac{1}{2p}, \tag{15}$$

then the optimum gain setting is obtained by repeating the Gauss-Seidel iteration process defined in Theorem 2 with the gradient given by

$$G_k = 2E(w_k), \tag{16}$$

where $k = 1, 2, \cdots, N$ and $E(w_k)$ is the frequency domain error value measured at frequency w_k, the center frequency of the kth Bode network.

The proof is given in Appendix C.

Thus, if signals are transmitted at a set of frequencies equal to the center frequencies, w_k, of the Bode networks, and if the errors are measured at these frequencies at the receiving station, the gradients, G_k, can be obtained directly. Then each gain, g_k, would be adjusted until its gradient, G_k, reduced to zero. This is the well-known "zero-forcing" technique used in Ref. 2; it also achieves the optimum equalized channel in the MSE sense, if the stipulated assumptions apply.

Theorem 4 (MSE algorithm): Let the bump equalizer in this case satisfy assumption 2 and assume that the interval, Δw, between adjacent Bode networks is no greater than the shortest frequency domain ripple period in the channel, i.e.,

$$\Delta w \leqq \frac{1}{p}. \tag{17}$$

Then the optimum gain setting is obtained by repeating the Gauss-Seidel *iteration process with the gradient in this instance given by*

$$G_k = \tfrac{1}{2}E\left(w_k - \frac{\Delta w}{2}\right) + E(w_k) + \tfrac{1}{2}E\left(w_k + \frac{\Delta w}{2}\right) \qquad (18)$$
$$k = 1, 2, \cdots, N,$$

where $E(w_k)$ is the frequency domain error at the center frequency of the kth Bode network, and $E[w_k - (\Delta w/2)]$ and $E[w_k + (\Delta/2)]$ are the frequency domain errors measured at lower and upper frequencies midway between adjacent Bode networks. Equation (18) is derived in Ref. 3.

Proof: In this case,

$$\langle B_k, B_k\rangle = 0.75$$

and

$$\sum_{j=1}^{N}{}' \, |\langle B_k, B_j\rangle| = 0.25$$

for all k, thus satisfying the inequality (14). Hence, the Gauss-Seidel iteration process converges to the optimum gain settings.

To implement the MSE algorithm, a measure of the error at $2N - 1$ points in the frequency domain is required (see Fig. 2). In practice, the MSE technique results in better equalization than that obtained by the ZF method. Note that assumption 2 for the MSE algorithm approximates the actual equalizer more precisely than assumption 1 does. Moreover, inequality (15) for the ZF algorithm derived in this section is a conservative assumption. The channel ripple period allowed by the MSE algorithm can be half the period assumed by the ZF algorithm.

III. CONTROL OF MANUAL BUMP EQUALIZERS

In this section, the Gauss-Seidel iteration process derived in the previous section is applied to the manual equalizer for optimum gain control. The number of iterations required to obtain acceptable gain settings is reflected in inequality (14). The larger the diagonal components [left-hand side of (14)] compared to the off-diagonal components [right-hand side of (14)], the fewer the iterations needed. The rate of convergence of the iteration is described in Section 3.2 based on an error-reduction analysis. It is shown that one iteration is sufficient to obtain the optimum gain settings for the ZF Gauss-Seidel iteration algorithm derived in Theorem 3. When, in the more general case, the channel is initially equalized by the ZF algorithm, one or two more iterations are usually sufficient to achieve a practically optimum equalized channel for the MSE algorithm.

3.1 Hardware realization of Gauss-Seidel iteration process

For the L5 line equalization, an equalizer adjustment system has been developed for the adjustment of bump equalizers by the Gauss-Seidel iteration process. It is composed of a precision transmission measuring set, 90G oscillator—90H detector—digital control unit (Ref. 4), and a hardwired, special-purpose computer which contains a programmed memory and an arithmetic unit called an equalizer adjustment unit (EAU) (see Fig. 3). Referring to Fig. 2, we assume that the equalizers in the receiving station are to be adjusted by the MSE algorithm. By selecting the particular Bode network to be adjusted, the EAU in the transmitting station causes the 90G oscillator to generate sequentially an appropriate set of three frequencies (f_{k1}, f_{k2}, and f_{k3}). The EAU in the receiving station measures the channel error at the same three frequencies and computes the gradient, G_k,

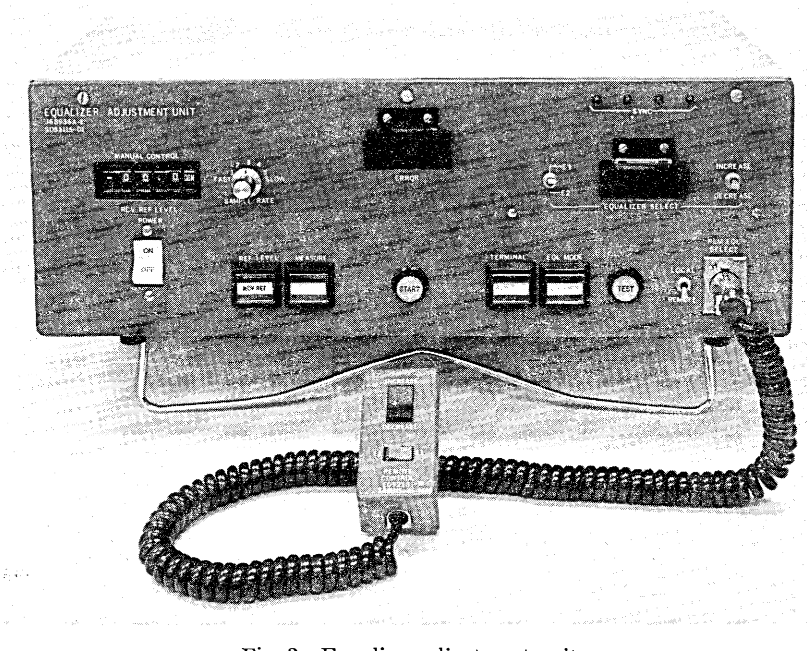


Fig. 3—Equalizer adjustment unit.

by the following relationship:

$$G_k = B_{k1}E(f_{k1}) + B_{k2}E(f_{k2}) + B_{k3}E(f_{k3}).$$

Normally, $B_{k1} = B_{k3} = \frac{1}{2}$ and $B_{k2} = 1$. The resultant is displayed in a digital readout of the EAU and the operator subsequently adjusts g_k until $G_k = 0$. Then the next Bode network is selected and the transmitting EAU causes the generation of the required set of three frequencies, and the receiving EAU processes the received error signals and displays the calculated gradient. Again, the corresponding gain adjustment is made. When the ZF algorithm is selected, the gradient displayed is simply the error at the center frequency of the Bode network. Hence, as far as the operator is concerned, the adjustment procedure for the ZF and MSE algorithms is identical.

In practice, it is found that the equalizer should be initially adjusted by the ZF algorithm to bring the system near the optimum state. In this way, any large initial gain deviations from the optimum value are quickly reduced to within about ±0.5 dB. Then the MSE algorithm is used for the "fine tuning" of the gain adjustments. Usually, one or two iterations with the MSE algorithm will be sufficient for the equalizer to reach the optimum MSE state when starting from the ZF state. According to inequality (23) derived in the following section, and given initial gain settings within ±0.5 dB of the optimum value, the gain settings after two iterations are within ±0.036 dB of the ideal values in the worst case. The actual deviation from optimum in most cases will be smaller than 0.036 dB and in any case will be less than the inherent accuracy limitations of the 90-type transmission measuring sets to be used.

3.2 Error analysis

The Gauss-Seidel iteration provides fast convergence of initial gain settings to the optimum values for the bump equalizers. As derived in Appendix B, the Gauss-Seidel iteration can be expressed by

$$\mathbf{g}_{(i+1)} = -\mathbf{L}^{-1}\mathbf{U}\mathbf{g}_{(i)} + \mathbf{L}^{-1}\mathbf{M}, \tag{19}$$

where $\mathbf{L}$ is the lower triangular portion of the system matrix $\mathbf{B}$ including the diagonal and $\mathbf{U}$ is the upper triangular portion of $\mathbf{B}$ not including the diagonal.

Defining an error vector after the ith iteration to be

$$\mathbf{e}_{(i)} = \mathbf{g}^* - \mathbf{g}_{(i)}, \tag{20}$$

where $\mathbf{g}^*$ is the optimum value, then

$$\begin{aligned} \mathbf{e}_{(i+1)} &= \mathbf{g}^* - \mathbf{g}_{(i+1)} \\ &= \mathbf{g}^* + \mathbf{L}^{-1}\mathbf{U}\mathbf{g}_{(i)} - \mathbf{L}^{-1}\mathrm{M} \\ &= -\mathbf{L}^{-1}\mathbf{U}\mathbf{e}_{(i)}. \end{aligned} \tag{21}$$

To derive (21), the equality, $\mathbf{L}^{-1}\mathbf{M} = \mathbf{g}^* + \mathbf{L}^{-1}\mathbf{U}\mathbf{g}^*$, was used. Hence, the magnitude of the eigenvalues of $\mathbf{L}^{-1}\mathbf{U}$ determines the speed of convergence.

If the equalizer belongs to the class defined in assumption 1, $\mathbf{B}$ is a positive definite diagonal matrix and $\mathbf{L}^{-1}\mathbf{U}$ is a null matrix. Hence, $\mathbf{e}_{(i+1)} = \mathbf{0}$ for all $i = 0, 1, 2, \cdots$. In other words, we obtain the optimum gain settings by the ZF Gauss-Seidel iteration algorithm in one iteration. The same result can be obtained from eq. (19), since $\mathbf{L}^{-1}\mathbf{U}$ is a null matrix and $\mathbf{L}^{-1} = \mathbf{B}^{-1}$.

If the equalizer satisfies assumption 2, the MSE Gauss-Seidel iteration algorithm can be used. Now the system matrix is

$$\mathbf{B} = 2\begin{bmatrix} b_{11}, & b_{12}, & b_{13}, & \cdots, & b_{1N} \\ b_{21}, & b_{22}, & b_{23}, & \cdots, & b_{2N} \\ \vdots & & & & \\ b_{N1}, & b_{N2}, & b_{N3}, & \cdots, & b_{NN} \end{bmatrix},$$

where

$$\begin{aligned} b_{ij} &= 0.75 && \text{if} \quad i = j \\ b_{ij} &= 0.125 && \text{if} \quad |i - j| = 1 \end{aligned}$$

and

$$b_{ij} = 0 \qquad \text{if} \quad |i - j| \geqq 2$$

for all $i, j = 1, 2, \cdots, N$. Splitting $\mathbf{B}$ into two parts, $\mathbf{L}$ and $\mathbf{U}$, which are defined above, and performing some algebra,

$$\mathbf{L}^{-1}\mathbf{U} = \frac{4}{3}\begin{bmatrix} 0, & 6^{-1}, & 0, & 0, & \cdots, & 0 & \\ 0, & -6^{-2}, & 6^{-1}, & 0, & \cdots, & 0 & \\ 0, & 6^{-3}, & -6^{-2}, & 6^{-1}, & \cdots, & 0 & \\ \vdots & & & & & & \\ 0, & (-1)^{N+1}6^{-N}, & (-1)^{N}6^{-(N-1)}, & \cdots & 6^{-3}, & -6^{-2} & \end{bmatrix}.$$

Hence, one can calculate the new error vector by eq. (21). After one iteration, the upper bound on the maximum residual error becomes

$$\begin{aligned} |e_{k(1)}|_{\max} &\leqq \tfrac{4}{3}(6^{-1} + 6^{-2} + 6^{-3} + \cdots)|e_{j(0)}|_{\max} \\ &= 0.26667|e_{j(0)}|_{\max} \end{aligned} \tag{22}$$

for all $j, k = 1, 2, \cdots, N$. Similarly, after the second iteration,

$$|e_{k(2)}|_{\max} \leqq 0.26667|e_{j(1)}|_{\max} \leqq 0.07111|e_{i(0)}|_{\max} \tag{23}$$

for all $i, j, k = 1, 2, \cdots, N$. The equality in eqs. (22) and (23) will be obtained if and only if

$$e_{1(0)} = - e_{2(0)} = e_{3(0)} = - e_{4(0)} = \cdots = |e_{k(0)}|_{\max}.$$

This in general will not be the case, and the maximum setting error after the second iteration usually will be less than 0.07111 times the maximum setting error prior to the first iteration. Consequently, if the gain settings are within 0.5 dB of optimum at the start (which a single ZF iteration will establish), the deviation from optimum settings in the MSE sense is within a few hundredths of a decibel after two additional iterations.

IV. CONCLUSION

This paper shows that a manual equalization process which can be described by a Gauss-Seidel iteration method provides optimal control for bump equalizers in the MSE sense. Compared to the steepest descent method discussed in Ref. 3, the Gauss-Seidel iteration method can be more economically implemented for manual equalization such as in the L5 system. The Gauss-Seidel iteration process requires knowledge of the gradient of the MSE with respect to the gain setting for each Bode network to be adjusted. The ZF algorithm derived in this paper requires just one Gauss-Seidel iteration, but in practice the ZF equalized channel is not optimum and can be further improved by the MSE algorithm. This is because the gradient obtained by the MSE algorithm is more accurate than the one obtained by the ZF algorithm for realizable equalizer shapes. It should be noted that three tones are required to determine the gradient of the MSE with respect to the gain setting of each Bode network for the MSE algorithm, while only one tone is used to obtain the gradient for the ZF algorithm. The number of iterations that are necessary to bring the equalizer to the optimum state depends on how close the initial settings are to optimum. When the channel is initially ZF-equalized, only one or two more iterations are needed to optimize the channel with the MSE algorithm. This result agrees completely with experiments conducted in the L5 field trial.

APPENDIX A

Proof of Theorem 1

Substituting eqs. (9) and (10) into eq. (11), we obtain

$$G_k = 2 \int_{-\infty}^{\infty} B_k(w) \left\{ \sum_{j=1}^{N} g_j B_j(w) - M(w) \right\} dw. \tag{24}$$

Defining an inner product

$$\langle A, B\rangle = \int_{-\infty}^{\infty} A(w)B(w)dw,$$

$$\mathbf{G} = [G_1, G_2, \cdots, G_N]^T,$$

where T indicates the transpose,

$$\mathbf{g} = [g_1, g_2, \cdots, g_N]^T$$

$$\mathbf{M} = 2[\langle B_1(w), M(w)\rangle, \langle B_2(w), M(w)\rangle, \cdots, \langle B_N(w), M(w)\rangle]^T$$

and

$$\mathbf{B} = 2\begin{bmatrix} \langle B_1, B_1\rangle, & \langle B_1, B_2\rangle, & \cdots, & \langle B_1, B_N\rangle \\ \langle B_2, B_1\rangle, & \langle B_2, B_2\rangle, & \cdots, & \langle B_2, B_N\rangle \\ \vdots & & & \\ \langle B_N, B_1\rangle, & \langle B_N, B_2\rangle, & \cdots, & \langle B_N, B_N\rangle \end{bmatrix},$$

a simultaneous equation of the type of eq. (24) for all k from 1 to N can be written as

$$\mathbf{G} = \mathbf{Bg} - \mathbf{M}$$

or

$$\mathbf{Bg} = \mathbf{G} + \mathbf{M}. \tag{25}$$

Since eq. (25) is a nonhomogeneous system of N equations and $\mathbf{G} + \mathbf{M}$ is a vector with N real-numbered components in the case considered, for the given $\mathbf{G}$, the unknown $\mathbf{g}$ is uniquely obtained by

$$\mathbf{g} = \mathbf{B}^{-1}(\mathbf{G} + \mathbf{M}), \tag{26}$$

provided that $\mathbf{B}$ is a nonsingular matrix.

However, if $\mathbf{B}$ were a singular matrix, then a linear combination of the columns could be made zero, i.e.,

$$\sum_{k=1}^{N} h_k\langle B_k, B_j\rangle = 0$$

for each $j = 1, 2, \cdots, N$ where h_k are real nonzero numbers.

In addition, the following relationship could also hold:

$$\sum_{k=1}^{N} h_1 h_k\langle B_k, B_1\rangle + \sum_{k=1}^{N} h_2 h_k\langle B_k, B_2\rangle + \cdots + \sum_{k=1}^{N} h_N h_k\langle B_k, B_N\rangle = 0. \tag{27}$$

But (27) can also be written as

$$\int_{-\infty}^{\infty} \left\{ \sum_{k=1}^{N} h_k B_k(w) \right\}^2 dw = 0. \tag{28}$$

Equation (28) contradicts the assumption that $B_k(w)$'s are linearly independent. Hence, **B** is indeed a nonsingular matrix and there exists but one set of g_k's for which $\mathbf{G} = \mathbf{0}$ in eq. (26). It is yet necessary to prove that this stationary point is the global minimum of the MSE defined in eq. (10), which is established if the following relationship can be proved:

$$\int_{-\infty}^{\infty} \left\{ M(w) - \alpha \sum_{k=1}^{N} g_k^* B_k(w) - \beta \sum_{k=1}^{N} g_k^{**} B_k(w) \right\}^2 dw$$
$$< \alpha \int_{-\infty}^{\infty} \left\{ M(w) - \sum_{k=1}^{N} g_k^* B_k(w) \right\}^2 dw$$
$$+ \beta \int_{-\infty}^{\infty} \left\{ M(w) - \sum_{k=1}^{N} g_k^{**} B_k(w) \right\}^2 dw, \quad (29)$$

where

$$\sum_{k=1}^{N} g_k^* B_k(w) \quad \text{and} \quad \sum_{k=1}^{N} g_k^{**} B_k(w)$$

indicate distinct equalizer settings, $\alpha + \beta = 1$ and $\alpha, \beta > 0$, then MSE is a strict convex function of gain settings g_k's and has a global minimum.

Subtracting the left-hand side from the right-hand side of inequality (29), we obtain the following:

$$-2\alpha\beta \int_{-\infty}^{\infty} \left\{ \left[\sum_{k=1}^{N} g_k^* B_k(w) \right]^2 + \left[\sum_{k=1}^{N} g_k^{**} B_k(w) \right]^2 \right\} dw. \quad (30)$$

Since $B_k(w)$'s are linearly independent and at least one equalizer setting,

$$\sum_{k=1}^{N} g_k^* B_k(w) \quad \text{or} \quad \sum_{k=1}^{N} g_k^{**} B_k(w),$$

is not zero, (30) is negative. Hence, inequality (29) is correct, and the proof of Theorem 1 is complete.

APPENDIX B

Proof of Theorem 2

The gradient of MSE with respect to the gain settings is represented by the following equation:

$$\mathbf{G} = \mathbf{Bg} - \mathbf{M}, \quad (31)$$

where **B**, **g**, and **M** are defined in (24). Splitting the **B** matrix as follows

$$\mathbf{B} = \mathbf{L} + \mathbf{U}, \quad (32)$$

where $\mathbf{L}$ is the lower triangular portion of $\mathbf{B}$ including the diagonal and $\mathbf{U}$ is the remainder of $\mathbf{B}$,

$$\mathbf{G} = \mathbf{Lg} + \mathbf{Ug} - \mathbf{M}. \tag{33}$$

According to the iterative procedure described in the theorem, $\mathbf{g}_{(i+1)}$ is obtained from $\mathbf{g}_{(i)}$ by setting $\mathbf{G} = \mathbf{0}$. With the aid of eq. (33), this procedure can be expressed by

$$\mathbf{0} = \mathbf{Lg}_{(i+1)} + \mathbf{Ug}_{(i)} - \mathbf{M}$$

or

$$\mathbf{g}_{(i+1)} = -\mathbf{L}^{-1}\mathbf{U}\mathbf{g}_{(i)} + \mathbf{L}^{-1}\mathbf{M}. \tag{34}$$

By successive calculation, eq. (34) can be modified by

$$\mathbf{g}_{(i+1)} = [-\mathbf{L}^{-1}\mathbf{U}]^{i+1}\mathbf{g}_{(0)} + \sum_{k=0}^{i} [-\mathbf{L}^{-1}\mathbf{U}]^{k}\mathbf{L}^{-1}\mathbf{M}, \tag{35}$$

where $\mathbf{g}_{(0)}$ is the initial value.

If

$$[-\mathbf{L}^{-1}\mathbf{U}]^{i} \rightarrow [\mathbf{0}] \quad \text{as} \quad i \rightarrow \infty,$$

$$\sum_{k=0}^{i} [-\mathbf{L}^{-1}\mathbf{U}]^{k}\mathbf{L}^{-1} \rightarrow [\mathbf{L} + \mathbf{U}]^{-1} = \mathbf{B}^{-1}.$$

Hence, eq. (35) becomes

$$\mathbf{g}_{(i+1)} = \mathbf{0} + \mathbf{B}^{-1}\mathbf{M},$$

which is the desired result.

Hence, the theorem is proved if $\mathbf{L}^{-1}\mathbf{U}$ is a convergent matrix, i.e., eigenvalues of the matrix $\mathbf{L}^{-1}\mathbf{U}$ are all less than one in absolute value. However, if condition (14) is satisfied, $\mathbf{L}^{-1}\mathbf{U}$ is a convergent matrix and $[-\mathbf{L}^{-1}\mathbf{U}]^{i} \rightarrow [\mathbf{0}]$ as $i \rightarrow \infty$ (see Theorems 3.3 and 3.4 in Ref. 5).

Note: The iteration process defined by (34) is known as the Gauss-Seidel or, simply, the Seidel iteration.

APPENDIX C

Proof of Theorem 3

When assumption 1 is satisfied,

$$\begin{aligned} \langle B_k, B_j \rangle &= 1, \quad k = j \\ &= 0, \quad k \neq j \end{aligned}$$

for all j and k.

Hence, Theorem 2 is satisfied and we have to prove now

$$G_k = 2E(w_k). \tag{36}$$

From eq. (24),

$$\begin{aligned} G_k &= 2\int_{-\infty}^{\infty} B_k(w)\left\{\sum_{j=1}^{N} g_jB_j(w) - M(w)\right\}dw \\ &= 2\int_{-\infty}^{\infty} \operatorname{sinc}\left[\frac{\pi}{\Delta w}(w-w_k)\ \right]\sum_{j=1}^{N} g_jB_j(w)dw \\ &\qquad - 2\int_{-\infty}^{\infty} \operatorname{sinc}\left[\frac{\pi}{\Delta w}(w-w_k)\right]M(w)dw. \end{aligned} \tag{37}$$

Since

$$\int_{-\infty}^{\infty} \operatorname{sinc}\left[\frac{\pi}{\Delta w}(w-w_k)\right]\operatorname{sinc}\left[\frac{\pi}{\Delta w}(w-w_j)\right]dw = 0 \quad \text{if} \quad k \neq j$$

and

$$= 1 \quad \text{if} \quad k = j,$$

the first integration in (37) is simply $2g_k$.

Substituting $M(w)$ of (2) into (37), the second integration of (37) becomes

$$\begin{aligned} &2\int_{-\infty}^{\infty} \operatorname{sinc}\left[\frac{\pi}{\Delta w}(w-w_k)\right]\int_0^1 \{F(x)\cos(2\pi pwx) \\ &\qquad\qquad + H(x)\sin(2\pi pwx)\}dxdw \\ &= \int_{-\infty}^{\infty} \operatorname{sinc}\left[\frac{\pi}{\Delta w}(w-w_k)\right]\int_0^1 \{f(x)\cos[2\pi p(w-w_k)x] \\ &\qquad\qquad + h(x)\sin[2\pi p(w-w_k)x]\}dxdw, \end{aligned} \tag{38}$$

where

$$f(x) = F(x)\cos(2\pi pw_kx) + H(x)\sin(2\pi pw_kx)$$

and

$$h(x) = H(x)\cos(2\pi pw_kx) - F(x)\sin(2\pi pw_kx).$$

Since

$$\int_0^1\int_{-\infty}^{\infty} h(x)\operatorname{sinc}\left[\frac{\pi}{\Delta w}(w-w_k)\right]\sin[2\pi p(w-w_k)x]dwdx = 0,$$

Eq. (38) becomes

$$2\int_{-\infty}^{\infty} \operatorname{sinc}\left[\frac{\pi}{\Delta w}(w-w_k)\right]\int_0^1 \{f(x)\cos[2\pi p(w-w_k)x]\}dxdw. \tag{39}$$

Replacing $w = u + w_k$ and changing the "cos" into "exponential" form, (39) becomes

$$\int_{-\infty}^{\infty} \operatorname{sinc}\left(\frac{\pi}{\Delta w}u\right) \int_0^1 \{f(x)[\exp(i2\pi pux) + \exp(-i2\pi pux)]\} dxdu$$

$$= \int_0^1 f(x) \int_{-\infty}^{\infty} \operatorname{sinc}\left(\frac{\pi}{\Delta w}u\right) \{\exp(i2\pi pux) + \exp(-i2\pi pux)\} dudx, \qquad (40)$$

where $i^2 = -1$. Since $0 \leqq x \leqq 1$ and $2p \leqq 1/\Delta w$ by assumption, integration of (40) is simply

$$2 \int_0^1 f(x)dx.$$

Note that the inner integration in (40) is the Fourier transformation of the sinc function.

Combining the results, G_k in (37) becomes

$$G_k = 2g_k - 2 \int_0^1 f(x)dx. \qquad (41)$$

However,

$$M(w_k) = \int_0^1 f(x)dx \quad \text{and} \quad \mathrm{EQL}(w_k) = g_k.$$

Hence, (41) becomes

$$\begin{aligned} G_k &= 2[\mathrm{EQL}(w_k) - M(w_k)] \\ &= 2E(w_k). \end{aligned}$$

This proves the theorem.

REFERENCES

1. H. W. Bode, "Variable Equalizers," B.S.T.J., *17*, No. 2 (April 1938), pp. 229–244.
2. F. C. Kelcourse, W. G. Scheerer, and R. J. Wirtz, "Equalizing and Main Station Repeaters," B.S.T.J., *48*, No. 4 (April 1969), pp. 889–925.
3. Cho, Y.-S., "Optimal Equalization of Wideband Coaxial Cable Channels Using 'Bump' Equalizers," B.S.T.J., *51*, No. 6 (July–August 1972), pp. 1327–1345.
4. N. H. Christiansen, "New Instruments Simplify Carrier System Measurements," Bell Laboratories Record, *48*, No. 8 (September 1970), pp. 232–238.
5. R. S. Varga, *Matrix Iterative Analysis*, Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1962.

Copyright © 1974 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 5, May–June 1974
Printed in U.S.A.

Mathematical Analysis of an Adaptive Quantizer

By DEBASIS MITRA

(Manuscript received December 4, 1973)

This paper presents a mathematical analysis of an adaptive quantizer, a pulse code modulator, which is used for coding speech and other continuous signals with a large dynamic range into digital form. The device is a two-bit quantizer in which the step size is modified at every sampling instant with the object of adapting the range of the device to the intensity level of the signal. In the adaptation algorithm analyzed in the paper, the encoded information of the previous sampling instant is used either to increase or to decrease the step size by fixed, but not necessarily equal, proportions.

Initially, the stochastic stability of the device is established by constructing a stochastic Liapunov function. Various basic identities and bounds on aspects of the behavior of the device are obtained. The qualitative results obtained indicate the nature of the trade-offs between the quality of the steady state and the transient performance of the device. Also, formulas are developed for the purpose of evaluating the mean time required for the step size to adapt from arbitrary initial conditions to certain optimal values.

I. INTRODUCTION

A mathematical analysis of an adaptive quantizer is presented in this paper. The coding thresholds of the device, also referred to as the step sizes, are not fixed but adapt according to a particular alogrithm. The object of the algorithm is to modify the threshold to larger or smaller levels, depending on whether the signal intensity level is high or low, in a manner that allows a decoder at the receiving end to effectively reconstruct the continuous signal. The basic two-bit quantizer, i.e., quantizers with four output levels with codes 01, 00, 10, and 11, is characterized by a particular function of the following form at each sampling instant.

Input refers to the nth sample of the continuous signal, $x(n)$, $n = 0, 1, 2, \cdots$; output refers to the coded signal to be transmitted at that instant; and Δ is the step size. In adaptive quantizers of the type to be investigated here, the step size is variable and the step size at the nth sampling instant is denoted by $\Delta(n)$. The step size uniquely defines the entire function in the manner indicated by Fig. 1; hence, the complete adaptive quantizer is associated with a sequence of functions. The adaptive quantizers that are the subject of this paper are basically characterized by the following adaptation algorithm

$$\Delta(n+1) = M_1\Delta(n) \quad \text{if} \quad |x(n)| \leqq \Delta(n) \tag{1a}$$

$$= M_2\Delta(n) \quad \text{if} \quad |x(n)| > \Delta(n), \tag{1b}$$

where M_1 and M_2, called multiplier coefficients, are fixed constants satisfying* $0 < M_1 < 1 < M_2$. Variations on (1) are considered in the main text, although the discussion in the introductory section is in terms of (1). Results on adaptive quantizers with output levels more numerous than 4 will be considered in a future publication.

The adaptation algorithm in (1) is due to Cummiskey, Flanagan, and Jayant.[1,2] In Ref. 1 Jayant presents the results of extensive computer simulations undertaken to determine the multiplier coefficients which maximize various performance functionals. A class of random inputs $\{x(n)\}$ that is considered is obtained by passing a discrete, white, Gaussian process through a filter with a single pole. In Ref. 2, Cummiskey, Jayant, and Flanagan consider a differential PCM coder in which the adaptive quantizer is used together with a fixed first-order predictor in the feedback loop. Their work has its direct antecedents in the various schemes[3,4,5] for adapting step sizes in delta-modulators, a one-bit quantizer, and in the work of Wilkinson.[6] Wilkinson's paper on a two-bit adaptive quantizer, largely concerned with hardware implementation, is particularly interesting. In his scheme, the step size is controlled by a moving fraction obtained by keeping a tally of the number of times the input falls in the lower slot of the quantizer. Goodman and Gersho[7] have independently looked at the adaptive quantizer from a theoretical standpoint and their work complements the work described here.

In this paper we make a number of simplifying assumptions about the input sequence $\{x(n)\}$, the most restrictive being the assumption

* Since the absolute value of the input in Fig. 1 is partitioned into $[0, \Delta]$ and $(\Delta, \infty]$, we shall loosely refer to the event leading to (1a) as "the input falling in the lower slot."

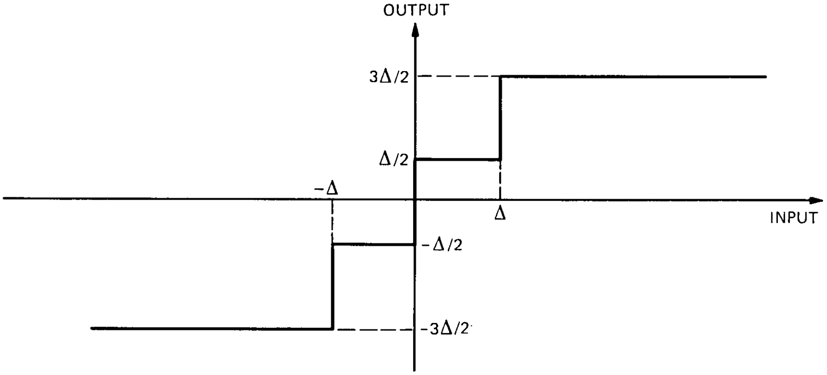


Fig. 1—The quantizer function.

that it is a sequence of independent random variables. However, we have obtained for the idealized model precise results which indicate rather fully the trade-offs involved in the choice of the multiplier coefficients. Also, we have developed formulas for efficiently computing functionals as aids in the design problem. We believe that the broad qualitative features of the device that are found to hold in this model carry over for more realistic input processes. It is hoped too that the techniques developed here will provide a point of reference for future work.

The mathematical analysis, for the main part, is of a random walk on the integers, whose complexity is due to the dependence of the state transition probabilities on the states. The structure of the random walk which is exploited here is rather general, and for this reason the model is of independent interest; to our knowledge, the main mathematical results have not appeared in the literature on random walks.

The organization of the paper is as follows. In Section 1.1 we continue the discussion on the adaptation algorithm in the context of a particular idealized model of the sequence $\{x(n)\}$, and we discuss some of the results to be derived later and what is already known about optimal quantization in the nonadaptive framework. In Sections 1.2 and 1.3 we give the basic equations of the process arising from (1), and certain modifications of it, when the input sequence $\{x(n)\}$ is independent and identically distributed. In Section II the stochastic stability of the device is established under general conditions. The existence and uniqueness of the stationary distribution of the step size is proved by constructing a stochastic Liapunov function for the random process. Section III examines in detail the stationary step size distribution. In Section 3.2 we prove an identity which explicitly gives the stationary probability of the input falling in the lower slot of the quantizer, i.e., $\Pr_s\,[|x(n)| \leqq \Delta(n)]$. In Section 3.3 sharp bounds are obtained on the stationary probabilities. It is shown that for almost all values of the multiplier coefficients there exists a natural center of the distribution and that the stationary probabilities fall off at least geometrically with increasing distances from the natural center. In Section 3.5 results are obtained on a particular limiting behavior, namely, the effect of the stationary distribution of making both multiplier coefficients close to unity. Section IV is devoted to the transient response of the device. In Section 4.1 we develop formulas for the efficient computation of the time required for the step size to adapt from an arbitrary initial value to the desired step size. Section 4.2 by giving an explicit bound on this time provides some insight into the dependence of the adaptive time on the choice of the multiplier coefficients. Finally, we report some computational results.

1.1 Background

In an idealized model for the samples, $x(n)$, of the continuous signal process, assume that $\{x(n)\}$ is a sequence of independent random variables with zero mean. Assume further that the distribution of $x(n)$ for every n is an element of the same equivalence class of distributions in which the distributions are equivalent to within a scaling operation. The scaling or intensity level changes slowly with n. For instance, the equivalence class of distributions may be the family of Gaussian distributions and only the variance, indicating the intensity level, changes with n.

It is necessary to recall at this stage some known facts concerning the design of quantizers in the nonadaptive framework[8] where $\{x(n)\}$ is a sequence of independent, identically distributed random variables and the step size is fixed. Suppose that $E[\{y(n) - x(n)\}^2]$ measures the performance of the quantizer where $y(n)$ is the nth output of the device.* The step size which minimizes this functional, $\hat{\Delta}$, is in principle easy to establish, and $\hat{\Delta}$ is uniquely characterized by the probability of the input falling in the lower slot, i.e., $\Pr\,[|x(n)| \leqq \hat{\Delta}]$. Another observation that is equally easy to verify is that the optimal step size has the property that if the distribution of $\{x(n)\}$ is scaled, then the optimal step size is obtained by an identical scaling of the previous optimal step size. A convenient way of stating this observation is: a

* It is not essential that the performance functional be of that form.

property of the optimal step size that is invariant to scaling of the distribution of $\{x(n)\}$ is the probability that the absolute value of the input $x(n)$ does not exceed the optimal step size. For instance, when the distribution is Gaussian it is known that this probability is close to 0.68.[8]

An intermediate step in proceeding from the nonadaptive case to the more general model described prior to it, in which the identically distributed condition does not hold, is provided by the following model. Assume that the sequence $\{x(n)\}$ is indeed independent and identically distributed, and that the equivalence class of distributions to which the particular distribution belongs is known. However, the scaling parameter is unknown. It is relatively straightforward to state the requirements on a well-behaved algorithm operating in this simple framework, and, if these requirements are always satisfied, then it is possible to conclude that the device will operate satisfactorily for the more general model. The requirements are: (*i*) for arbitrary initial step size guesses, the step size rapidly converges to the optimal step size, and (*ii*) it is thereafter localized in a small neighborhood of that point. This paper separately analyzes the two requirements in the simple framework just described. Considerations related to (*i*) and (*ii*) are lumped respectively under the terms "transient response" and "steady-state response," since the latter property is effectively investigated in terms of the stationary distribution of the step size, assuming one exists. A good reason for the division is that they lead, in some ways, to quite opposite requirements for the multiplier coefficients.

Consider, in the light of what is known about optimal quantization in the nonadaptive framework, what is required for the localization property, requirement (*ii*), to hold. When the stationary distribution has both of the following properties, it is possible to establish an effective correspondence and infer that (*ii*) holds: (*a*) the stationary probability of the step size falling in the lower slot, i.e., $\Pr_s[|x(n)| \leqq \Delta]$ equals the known value associated with the particular family of distributions; and (*b*) the mass of the stationary distribution is concentrated in the small neighborhood of a point. In Section III we show that by appropriate choice of the multiplier coefficients it is possible to achieve both requirements.

1.2 Basic assumptions and equations

We consider only quantizers with multiplier coefficients having the following structure:

$$M_1 = \gamma^{-k} \quad \text{and} \quad M_2 = \gamma^{l}, \tag{2}$$

where γ is some real number greater than 1 and k and l are positive integers. We shall further make k and l relatively prime, i.e., their greatest common factor is 1. If, as we shall assume, the initial step size is of the form γ^i, with i an integer, then the step size is always of that form and the space of possible step sizes forms a lattice.[*]

There is a step size with, as we shall see, certain claims to being the central step size for a particular distribution of $\{x(n)\}$ and choice of parameters k and l; this step size is used as a reference point. There exists an integer $\bar{i}$ such that[†]

$$\Pr\left[|x(n)| \leqq \gamma^{\bar{i}-1}\right] < \frac{l}{k+l} \leqq \Pr\left[|x(n)| \leqq \gamma^{\bar{i}}\right]. \tag{3}$$

We denote $\gamma^{\bar{i}}$ by C and refer to it as the *central step size*; all step sizes are considered to be of the form $C\gamma^i$, $i = 0, \pm 1, \pm 2, \cdots$.

Obviously, it is more convenient to work with the log transform of the step size, so let

$$\omega(n) \triangleq \log_\gamma \Delta(n) - \log_\gamma C. \tag{4}$$

From the original algorithm we have

$$\begin{aligned} \omega(n+1) &= \omega(n) - k \quad \text{if} \quad |x(n)| \leqq C\gamma^{\omega(n)} \\ &= \omega(n) + l \quad \text{if} \quad |x(n)| > C\gamma^{\omega(n)}. \end{aligned} \tag{5}$$

We have in (5) a Markov chain with states $0, \pm 1, \pm 2, \cdots$. The state transition probabilities are obtained from the distribution of $x(n)$: for all integers i let

$$b_i \triangleq \Pr\left[|x(n)| \leqq C\gamma^i\right] \tag{6}$$

and

$$a_i \triangleq 1 - b_i.$$

The "b" is a mnemonic for backward probabilities since it is associated with a transition backwards from the generic state i to $(i-k)$. The diagram in Fig. 2 represents the Markov chain. Denoting by $p_i(n)$ the probability that $\omega(n) = i$, we have

$$\boxed{p_i(n+1) = b_{i+k}p_{i+k}(n) + a_{i-l}p_{i-l}(n).} \tag{7}$$

Although the transition probabilities depend on the distribution of $x(n)$, the two following properties of the sequence $\{b_i\}$, on which we

[*] D. J. Goodman suggested the above structure on the multiplier coefficients with the object of obtaining a discrete Markov process.

[†] We are tacitly assuming that $\Pr\left[|x(n)| = 0\right] \leqq l/(k+l) - \epsilon$, $\epsilon > 0$.

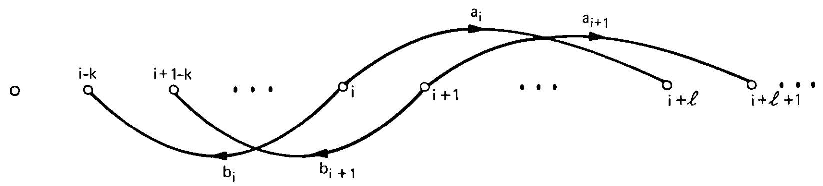


Fig. 2—The Markov chain.

base our results, hold irrespective of the distribution:

$$0 \leqq b_i < b_{i+1} \leqq 1 \quad \text{for all } i, \tag{8}$$

and

$$b_{-1} < \frac{l}{k+l} \leqq b_o. \tag{9}$$

That the strict inequality in (8) holds for all i is a mild restriction on the distribution of $x(n)$; however, certain straightforward modifications may be made to obtain corresponding results when the strict inequality does not hold for all i.

The property of the 0 state to which we alluded earlier may be loosely stated, thus: there is a net drift to the left (right) from states to the right (left) of the 0 state. Formally,

$$\boxed{\begin{aligned} E[\omega(n+1)\,|\,\omega(n)=i]-i=-(k+l)\left[b_i-\frac{l}{k+l}\right] &< 0 \text{ if } i>0 \\ &> 0 \text{ if } i<0. \end{aligned}} \tag{10}$$

The above super- and submartingale properties are the basis for the existence of a stochastic Liapunov function (Section 2.2) and the bound obtained in Section 4.2.

Remarks: The random walk in (5) with $k = l = 1$ is also the model for the delta-modulator subject to random, independent, identically distributed inputs. The stationary behavior of the model was treated in an elegant paper by Fine.[9] Gersho[10] has established the stochastic stability of the delta-modulator for a larger class of input processes. Some of our results, particularly those in Section IV on transient response, appear to be new and of some interest in this context.

1.3 The saturating adaptive quantizer

For the algorithm in (1) and, say, Gaussian distributions of the input, there is a small, positive probability of the step size exceeding any large prespecified level. A model which reflects more accurately the practical algorithm for adapting the step size is one which does not allow the step size to become unbounded. One way of implementing this is to make the step size saturate at some suitably large level, i.e., if $\Delta(n) < |x(n)|$, then

$$\Delta(n+1) = \min\,[M_2\Delta(n), \bar{L}]; \quad \bar{L} \gg 0; \tag{11}$$

i.e., in the log transformed variables,

$$\omega(n+1) = \min\,[\omega(n) + l, L]; \quad L \gg 0. \tag{12}$$

The model of this device, which we shall refer to as the saturating adaptive quantizer, is useful not only for the reasons given but also on theoretical grounds since the results obtained for the saturating adaptive quantizer yield, in the limit as $L \to \infty$, corresponding results for the adaptive quantizer. We carry both models with us throughout the paper and at least indicate along the way the main correspondences.

For similar reasons we expect that in practice the step size will also be bounded from below in the obvious manner. This case is not formally dealt with in the text since the main results may be readily inferred from the saturating adaptive quantizer.

For the saturating adaptive quantizer, the following equations govern the evolution of $\{p_i(n) = \Pr[\omega(n) = i]\}$, $i \leqq L$:

$$\boxed{\begin{aligned} p_i(n+1) &= b_{i+k}p_{i+k}(n) + a_{i-l}p_{i-l}(n) \quad && i \leqq L-k \\ p_i(n+1) &= a_{i-l}p_{i-l}(n) \quad && L-k+1 \leqq i \leqq L-1 \\ p_L(n+1) &= \sum_{j=L-l}^{L} a_j p_j(n). && \end{aligned}} \tag{13}$$

The important super- and sub-martingale properties of the random walk, as expressed by the inequalities in eq. (10), apply as well to the saturating adaptive quantizer.

II. THE EXISTENCE AND UNIQUENESS OF THE STATIONARY DISTRIBUTION

We examine in this section questions related to the stochastic stability of the adaptive quantizer. We establish theoretically that certain acute types of erratic operations such as the unboundedness of the evolving random variable, namely, the step size, do not occur. We begin by establishing that the process has the basic properties of a well-behaved process, namely, irreducibility and recurrence. We thereby establish the existence and uniqueness of a finite stationary distribution. We then proceed to the saturating adaptive quantizer, the more realistic model of the adapting algorithm, which in addition to the above properties, is also aperiodic. Here, the entire state space is a single ergodic class. The main result of this section is obtained from the construction of a stochastic Liapunov function for the process; and the theory of stochastic Liapunov functions is fairly well known.[11,12]

2.1 Irreducibility of the Markov chain

The chain is irreducible if and only if every state communicates with both the neighboring states. This occurs if and only if there exists nonnegative integers m, m', n, n' such that

$$ml - nk = 1 \tag{14a}$$

and

$$m'l - n'k = -1. \tag{14b}$$

It is an elementary fact from number theory that this occurs if and only if k and l are relatively prime, i.e., their greatest common divisor is unity. In fact, Euclid's algorithm yields the unknown quantities in eq. (14).

2.2 Recurrence

Consider the following nonnegative function of the states:

$$V(i) = |i| \qquad i = 0, \pm 1, \cdots. \tag{15}$$

This function is a stochastic Liapunov function[12] if the following holds: if $D(i)$ is defined as follows,

$$E[V\{\omega(n+1)\}\,|\,\omega(n) = i] - V(i) \triangleq D(i), \tag{16}$$

then (i) $D(i)$ is uniformly bounded from above and (ii) $D(i) \leqq -\epsilon < 0$ for all but a finite set of states i. Condition (i) is trivially true for the process. Also, for all $i \geqq k$

$$D(i) = -(k+l)\left(b_i - \frac{l}{k+l}\right) \leqq -(k+l)\left(b_k - \frac{l}{k+l}\right) < 0 \tag{17}$$

and, for all $i \leqq -l$,

$$D(i) = (k+l)\left(b_i - \frac{l}{k+l}\right) \leqq (k+l)\left(b_{-l} - \frac{l}{k+l}\right) < 0. \tag{18}$$

Therefore, condition (ii) is verified, and $V(i)$ is a stochastic Liapunov function for the process.

From Kushner's Theorem 7[12] we have recurrence* and we can infer further, from Theorem 4, that there exists at least one finite invariant measure, i.e., stationary distribution. Also, as we have shown earlier there does not exist two or more disjoint self-contained subsets of the state space; hence, we have from Theorem 5 that there is at most one invariant probability measure. Thus, the existence and uniqueness of a finite stationary distribution for the step size of the adaptive quantizer is established.

2.3 The saturating adaptive quantizer

We will circumvent the technical nuisance† posed by periodicity by proceeding to the saturating adaptive quantizer. In this case the above arguments leading to irreducibility and recurrence are intact. In addition, the end state L has period 1 and, since periodicity is a class concept (i.e., every state in a particular communicating class has the same periodicity), the entire Markov chain is aperiodic. We have, then, $p(n) \rightarrow p$ for any $p(0)$ and $p_i > 0$ for all i. Also, the state space is a single ergodic class. Hence, the statistical average of the step sizes approach a limit given by the unique, finite, stationary distribution.

III. SOME PROPERTIES OF THE STATIONARY DISTRIBUTIONS

In this section we investigate in detail properties of the stationary distribution of the step size. In eq. (7) if we set $p_i(n + 1) = p_i(n) = p_i$, then the stationary distribution is given by $\{p_i\}$. Thus, the stationary probabilities are the solutions of

$$\boxed{p_i = b_{i+k}p_{i+k} + a_{i-l}p_{i-l}} \tag{19}$$

with, of course, the normalization,

$$\sum_{-\infty}^{\infty} p_i = 1. \tag{20}$$

For the saturating adaptive quantizer, we have from eq. (13) that the basic recursion in (19) holds for all $i \leqq (L - k)$. The remaining

* A Markov chain is recurrent if and only if every state is recurrent; and state i is recurrent if and only if, starting from state i, the probability of returning to state i after some finite length of time is one.

† Feller[13] writes: "The classification into persistent and transient states is fundamental, whereas the classification into periodic and aperiodic states concerns a technical detail."

equations are (20) and the following:

$$p_i = a_{i-l}p_{i-l} \quad L - k + 1 \leqq i \leqq L - 1 \tag{21a}$$

$$p_L = \sum_{L-l}^{L} a_j p_j \tag{21b}$$

and, of course, $p_i = 0,\ i > L$.

3.1 A useful reduction of the equations for the stationary probabilities

To provide some insight into the motivation for the step we undertake here, consider the recursion, analogous to (19), that would arise from a Markov chain with uniform transition probabilities:

$$p_i = bp_{i+k} + ap_{i-l}, \quad a + b = 1. \tag{22}$$

A particular solution of the above recursion is $p_i \equiv c$, a constant. Since, in probability theory, interest is restricted to solutions with bounded sums, one would proceed in the case of (22) by factoring the root at unity from the characteristic polynomial:

$$b\lambda^{k+1} - \lambda^l + a = 0,$$

and thus obtain a new, and reduced, polynomial and an associated recursion. This operation is paralleled for the more general recursion in (19) by the following: from (19),

$$p_i - p_{i-l} = b_{i+k}p_{i+k} - b_{i-l}p_{i-l}.$$

Hence, for all j,

$$\sum_{-\infty}^{j} (p_i - p_{i-l}) = \sum_{-\infty}^{j} (b_{i+k}p_{i+k} - b_{i-l}p_{i-l}) \tag{22a}$$

which reduces to

$$\boxed{\sum_{j+1}^{j+k} b_i p_i = \sum_{j-l+1}^{j} (1 - b_i)p_i.} \tag{23}$$

Remarks:

(*i*) Observe that we are justified in carrying out the operation in (22a) in the case of solutions of (19) for which $\sum_{-\infty}^{j} p_i$ is bounded and which we have established, in Section II, to be the case for the stationary probabilities.

(*ii*) The reduction alluded to earlier refers to the fact that the largest difference in variable indices in (23) is $k + l$, while the largest difference in (19) is $k + l + 1$.

(*iii*) Observe that when $k = l = 1$, (23) gives the solution in closed form: $p_{j+1} = (a_j/b_{j+1})p_j$ and $\sum p_j = 1$. This is a previously known fact; see Feller[14] and Fine.[9] However, neither author gave any indication of the possible generalization to the form in (23).

For the saturating adaptive quantizer, (23) holds for all $j \leqq (L - k)$. Hence, the range over which (23) is valid is such that every state is included in at least one component of the recursion.

3.2 An identity involving the stationary distribution

We use eq. (23) to show that the *stationary* probability of the nth input sample, $x(n)$, falling in the lower slot, $\Pr_s [|x(n)| \leqq \Delta(n)] \equiv l/(k + l)$. The significance of this identity from the point of view of optimal steady-state operation (see Section 1.1) is that by appropriate choice of k and l the above quantity may be matched to the corresponding probability for the optimal nonadaptive step size. This, of course, has the effect of locating the central step size, eq. (3), close to the optimal nonadaptive step size. In the case of independent Gaussian inputs, the above quantity is close to 0.68 and a reasonable approximation is obtained by making $k = 1$ and $l = 2$.

From (23),

$$\sum_{j-l+1}^{j+k} b_i p_i = \sum_{j-l+1}^{j} p_i.$$

Hence,

$$\sum_{j=-\infty}^{\infty} \sum_{i=j-l+1}^{j+k} b_i p_i = \sum_{j=-\infty}^{\infty} \sum_{i=j-l+1}^{j} p_i. \tag{24}$$

The left-hand side equals $(k + l) \sum_{-\infty}^{\infty} b_j p_j$, while the right-hand side equals l. Hence,

$$\boxed{\sum_{-\infty}^{\infty} b_j p_j = \frac{l}{k + l}.} \tag{25}$$

Consider what the above equality implies in terms of step size behavior. The stationary probability of the input falling in the lower slot,

$$\Pr_s [|x(n)| \leqq \Delta(n)] = \sum_{i=-\infty}^{\infty} \Pr_s [\Delta = C\gamma^i \quad \text{and} \quad |x| \leqq C\gamma^i]$$

$$= \sum_{i=-\infty}^{\infty} b_i \Pr_s [\Delta = C\gamma^i] \tag{26}$$

from the independence of $\{x(n)\}$. Hence, from (25),

$$\Pr_s [|x(n)| \leqq \Delta(n)] = \frac{l}{k+l}. \tag{27}$$

Immediately on substituting $M_1 = \gamma^{-k}$ and $M_2 = \gamma^l$ we have an identity with a rather appealing and natural interpretation*:

$$M_1^{\rho_1} M_2^{\rho_2} = 1 \tag{28}$$

where ρ_1 and ρ_2 are respectively the two stationary probabilities of the input falling in the lower and upper slots.

For the saturating adaptive quantizer, it can be shown that

$$\sum_{i \leqq L} b_i p_i < \frac{l}{k+l}. \tag{29}$$

However, the quantity $[(l/k + l) - \sum b_i p_i]$ depends only on $(k + l)$ terms involving the end probabilities $p_L, \cdots, p_{L-k-l}$ and it goes to zero with these probabilities. Now we will prove in Section 3.3 certain results which indicate that these probabilities are relatively small if L is large.

3.3 Geometric bounds on the stationary probabilities

In this section we prove a fundamental property of the stationary distribution of the step size which holds for *all* values of γ. We obtain sharp bounds on almost all of the stationary probabilities—the bounds apply as well to the saturating adaptive quantizer—which show that the stationary probability of the random walk being in a particular state falls off at least geometrically with the distance of that state from the 0 state. The actual bounds obtained are substantially stronger and they indicate that a localization property on the stationary distribution is inherent for the random walk. As discussed in Section 1.1 this localization property is important in understanding the basis for the satisfactory behavior of the adaptive quantizer.

We obtain the following point-wise bound: for every $i > 0$ we give positive constants $r > 1$ and c such that for all $j \geqq i$,

$$p_j \leqq c \left(\frac{1}{r} \right)^{j-i}. \tag{30}$$

The quantities r and c depend on i. The quantity r which we call the

* D. J. Goodman first conjectured the existence of (28) in the context of the adaptive quantizer. Earlier, N. S. Jayant[5] made a related conjecture in connection with an adaptive delta-modulator.

local steepness factor is a monotonic increasing function of i for nonnegative i. Of course, a corresponding result holds for $i < 0$ and all $j \leqq i$,

Let $\mathbf{P}_i$ denote the $(k + l - 1)$-dimensional column vector* with the following components

$$\mathbf{P}_i \triangleq [p_i, p_{i+1}, \cdots, p_{i+k+l-2}]^t. \tag{31}$$

Then, from (23), we obtain $(k + l - 1) \times (k + l - 1)$ transition matrices $\mathbf{A}_i$, where

$$\mathbf{P}_{i+1} \triangleq \mathbf{A}_i\mathbf{P}_i. \tag{32}$$

The leading $(k + l - 2)$ components of $\mathbf{P}_{i+1}$ are obtained from $\mathbf{P}_i$ by merely shift operations. The nontrivial information in $\mathbf{A}_i$ is in the last row which is obtained from (23); clearly, $\mathbf{A}_i$ depends on i.

We will show that there exist a constant weight vector $\boldsymbol{\lambda}$, every element of $\boldsymbol{\lambda}$ being positive, and a constant $r > 1$ depending only on $\mathbf{A}_i$, such that for all $j \geqq i$

$$\boldsymbol{\lambda}^t\mathbf{A}_j^{-1} \geqq r\boldsymbol{\lambda}^t \tag{33}$$

in the sense that every element of the left vector is not less than the corresponding element of the right vector. Since $\mathbf{P}_{j+1}$ is a vector with nonnegative elements, we have

$$r\boldsymbol{\lambda}^t\mathbf{P}_{j+1} \leqq \boldsymbol{\lambda}^t\mathbf{A}_j^{-1}\mathbf{P}_{j+1} = \boldsymbol{\lambda}^t\mathbf{P}_j. \tag{34}$$

Hence,

$$\boxed{\boldsymbol{\lambda}^t\mathbf{P}_j \leqq \left(\frac{1}{r}\right)^{j-i}(\boldsymbol{\lambda}^t\mathbf{P}_i) \qquad j \geqq i.} \tag{35}$$

Remarks: Equation (35) is a strong result if $\boldsymbol{\lambda}^t\mathbf{P}_j$ is viewed as a norm of the vector $\mathbf{P}_j$ of the L_1-type: $|\mathbf{x}| = \sum \lambda_i|x_i|$, which is a valid interpretation since the latter reduces to $\boldsymbol{\lambda}^t\mathbf{x}$ whenever every element of $\mathbf{x}$ is nonnegative. By standard methods we can obtain upper bounds for $\mathbf{P}_j$ in norms other than the one used in (35). In particular, (30) follows trivially.

It is necessary now to discuss the structure of the matrix $\mathbf{A}_i^{-1}$. Directly from (23) we obtain the first row:†

$$\left[\overbrace{-\frac{a_{i+1}}{a_i}, -\frac{a_{i+2}}{a_i}, \cdots, \frac{-a_{i+l-1}}{a_i}}^{(l-1)\text{ terms}}, \overbrace{\frac{b_{i+l}}{a_i}, \frac{b_{i+l+1}}{a_i}, \cdots, \frac{b_{i+l+k-1}}{a_i}}^{k\text{ terms}}\right].$$

* The superscript t denotes the transpose.

† Observe that neither $\mathbf{A}_i$ nor $\mathbf{A}_i^{-1}$ is a stochastic matrix (nonnegative elements, columns sum to unity).

The remaining rows of $\mathbf{A}_i^{-1}$ reflect shift operations: for $m = 2, 3, \cdots, (k + l - 1)$,

$$\begin{aligned}(\mathbf{A}_i^{-1})_{mn} &= 0 \quad \text{if} \quad n \neq (m-1) \\ &= 1 \quad \text{if} \quad n = (m-1).\end{aligned}$$

Before proceeding to prove (33) we need the following lemma. This lemma concerns the matrix $\tilde{\mathbf{A}}_i^{-1}$ which is obtained from $\mathbf{A}_i^{-1}$ by merely replacing the first $(l - 1)$ elements of the first row by -1.

Lemma 1: For every $i \geqq 0$

(*i*) $\tilde{\mathbf{A}}_i^{-1}$ has a unique positive real eigenvalue r, say. Furthermore, $r > 1$.
(*ii*) Every element of the corresponding left eigenvector $\boldsymbol{\lambda}$ is of the same sign and nonzero; hence, $\boldsymbol{\lambda}$ may be taken to be a positive vector.
(*iii*) r, which depends on i, is monotonic, strictly increasing with i.

We give the proof of Lemma 1 in Appendix A.

We need one further observation to prove (33) with the help of the lemma. For $j \geqq i$,

$$\begin{aligned}\boldsymbol{\lambda}^t\mathbf{A}_j^{-1} &= \boldsymbol{\lambda}^t(\mathbf{A}_j^{-1} - \tilde{\mathbf{A}}_i^{-1}) + \boldsymbol{\lambda}^t\tilde{\mathbf{A}}_i^{-1} \\ &= \boldsymbol{\lambda}^t(\mathbf{A}_j^{-1} - \tilde{\mathbf{A}}_i^{-1}) + r\boldsymbol{\lambda}^t.\end{aligned}$$

The bound in (33) follows if $\boldsymbol{\lambda}^t(\mathbf{A}_j^{-1} - \tilde{\mathbf{A}}_i^{-1}) \geqq 0$. Since $\boldsymbol{\lambda}$ is a positive vector it is sufficient to show that the elements of the matrix $(\mathbf{A}_j^{-1} - \tilde{\mathbf{A}}_i^{-1})$ are nonnegative. The only nonzero elements of the matrix $(\mathbf{A}_j^{-1} - \tilde{\mathbf{A}}_i^{-1})$ are in the first row. That every term of the first row is nonnegative is implied by the following: for $s \geqq 1$

$$1 - \frac{a_{j+s}}{a_j} \geqq 0 \tag{36}$$

and

$$\frac{b_{j+s}}{a_j} - \frac{b_{i+s}}{a_i} \geqq 0. \tag{37}$$

This concludes the proof of (33) and, hence, of (35).

Remarks:

(*i*) The reader may now appreciate the reason for replacing some of the elements of $\mathbf{A}_i^{-1}$ by -1 to form $\tilde{\mathbf{A}}_i^{-1}$: a_{j+s}/a_j although bounded by 1 can come arbitrarily close to 1.

The reader is also due an explanation for our having worked with $\mathbf{A}_j^{-1}$ after defining the natural transformation $\mathbf{A}_j$, especially since (34) may be put in the form $\boldsymbol{\lambda}^t[\mathbf{I} - r\mathbf{A}_j]\mathbf{P}_j \geqq 0$. The reason is that r and λ, depending only on i, do not exist such that for $j \geqq i$, $\boldsymbol{\lambda}^t[\mathbf{I} - r\mathbf{A}_j] \geqq \mathbf{0}$, although, as we have shown, $\boldsymbol{\lambda}$ and r do exist such that $\boldsymbol{\lambda}^t[\mathbf{I} - r\mathbf{A}_j]\mathbf{A}_j^{-1} \geqq \mathbf{0}$. In working this step the assumption of $\mathbf{P}_{j+1} \geqq \mathbf{0}$, rather than $\mathbf{P}_j \geqq \mathbf{0}$, appears to be critical.

(*ii*) The interesting quantity $r = r(i)$ may reasonably be called the local steepness factor, since for $i \geqq 0$ it is a local measure of the rapidness with which the stationary distribution falls off. From statement (*iii*) of the lemma we have the fact that the distribution tends to get steeper with increasing distances from the natural center of the distribution, the 0 state.

(*iii*) The theoretical interest in the inequality in (35) results from the fact that we cannot expect to obtain a significantly better value than r for the geometric factor in geometrical bounds on p_j for all $j \geqq i$. The reason for this is that by making b_{j+1} very close to b_j over a fairly large set of j's, it is possible to make the solution of (23) close to the stationary probabilities of a random walk with uniform transition probabilities, which in turn may be obtained in terms of r as the unique positive real root of the characteristic polynomial $C(\mu)$ given in eq. (56), Appendix A.

(*iv*) From symmetry we expect results similar to (35) to hold for $i < 0$. Perhaps the simplest way to show this is by means of the following transformations which have the effect of making the direction of decreasing i the forward direction. Let

$$p'_{-i} = p_i, \quad b'_{-i} = a_i, \quad a'_{-i} = b_i.$$

The basic recursion (23), stated in terms of the new variables, is

$$\sum_{j+1}^{j+l} b'_i p'_i = \sum_{j-k+1}^{j} (1 - b'_i) p'_i.$$

Now $\{b'_i\}$ is a monotonic, increasing sequence with i and $i > 0 \Rightarrow lb'_i > ka'_i$. (Observe the interchange of l and k, i.e., $l' = k$ and $k' = l$.) This transformation makes the transfer of results holding for $i > 0$ to $i < 0$ fairly straightforward.

(*v*) In considering the application of (35) to the saturating adaptive quantizer we note that the basic recursion (23) holds over the entire range of states, i.e., (23) holds for all $j \leqq L - k$. Hence, (35) holds for $L - (l + k) + 2 \geqq j \geqq i \geqq 0$. This observation is the basis for a statement made earlier in Section 3.2, namely, we expect the tail probabilities of the stationary distribution of the step size for the saturating adaptive quantizer to be small.

From (35) we obtain a rather simple point-wise bound on the stationary probabilities. Let λ_m denote the largest element of the vector $\boldsymbol{\lambda}$. Clearly,*

$$\boldsymbol{\lambda}^t\mathbf{P}_i \leqq \lambda_m \mathbf{1}^t\mathbf{P}_i,$$

and, hence, from (35), for all $j \geqq i \geqq 0$

$$\lambda_m p_{j+m-1} \leqq \boldsymbol{\lambda}^t\mathbf{P}_j \leqq \left(\frac{1}{r}\right)^{j-i} \boldsymbol{\lambda}^t\mathbf{P}_i \leqq \left(\frac{1}{r}\right)^{j-i} \lambda_m(\mathbf{1}^t\mathbf{P}_i),$$

i.e.,

$$p_{j+m-1} \leqq \left(\frac{1}{r}\right)^{j-i} (\mathbf{1}^t\mathbf{P}_i).$$

Hence,

$$\boxed{p_{j+k+l-2} \leqq \left(\frac{1}{r}\right)^{j-i} (\mathbf{1}^t\mathbf{P}_i) \leqq \left(\frac{1}{r}\right)^{j-i}, \quad j \geqq i \geqq 0,} \tag{38}$$

where $r = r(i)$.

3.4 Lower bounds on the steepness factors, r(i)

We have associated with every state i a local steepness factor $r(i)$. Here we go back to the definition of $r(i)$ as being the unique positive root of the polynomial $C(\mu)$, eq. (56), to obtain the following bound which has the advantage of being explicit.

$$\boxed{\left[\frac{kb_i}{la_i}\right]^{1/(k+l-1)} \triangleq \rho(i) \leqq r(i), \qquad i \geqq 0.} \tag{39}$$

Observe that $\rho(i) > 1$ for all $i > 0$ and itself forms a monotonic increasing sequence with i. To prove (39) it is enough to show that $C[(kb_i/la_i)] \leqq 0$. The proof is straightforward but tedious and we omit it.

3.5 The effect of γ on the stationary distribution

We show here that the mass of the stationary distribution of the step size can be localized about the central step size to an arbitrary extent by making γ sufficiently close to unity. To do this, we first put

* The column vector with every element equal to unity is denoted by $\mathbf{1}$.

together from the results of the preceding sections a rather explicit bound on the stationary probability of the step size exceeding a particular value for a given γ, i.e., $\Pr_s[\Delta \geqq C\gamma^i]$. This bound is in a form which allows direct comparison with the corresponding probability arising from the choice of $\gamma' = \sqrt{\gamma}$. By successively taking γ to be the square root of the preceding value, the bound on the probability can be made as small as desired. As before, we shall restrict our attention to step sizes which exceed the central step size, i.e., $i > 0$ since a parallel argument holds for $i < 0$.

For $i > 0$ and $r = r(i)$, we have from (35) that

$$(\Sigma\lambda_i) \sum_{j=i+k+l-2}^{\infty} p_j \leqq \sum_{j=i}^{\infty} \boldsymbol{\lambda}^t\mathbf{P}_j \leqq \boldsymbol{\lambda}^t\mathbf{P}_i \sum_{j=0}^{\infty} \left(\frac{1}{r}\right)^j = \boldsymbol{\lambda}^t\mathbf{P}_i \frac{r}{r-1} \cdot \qquad (40)$$

Now, as in (39),

$$r \geqq \rho(i) = \left(\frac{kb_i}{la_i}\right)^{1/(k+l-1)}$$

and

$$\frac{\boldsymbol{\lambda}^t\mathbf{P}_i}{\Sigma\lambda_i} \leqq \max\left[p_i, \cdots, p_{i+k+l-2}\right].$$

Since

$$\Pr_s\left[\Delta \geqq C\gamma^{i+k+l-2}\right] = \sum_{j=i+k+l-2}^{\infty} p_j,$$

we have, from (40),

$$\boxed{\Pr_s\left[\Delta \geqq C\gamma^{i+k+l-2}\right] \leqq \frac{\rho(i)}{\rho(i)-1} \max\left[p_i, \cdots, p_{i+k+l-2}\right].} \qquad (41)$$

Finally, from (38), for $i \geqq k + l - 1$,

$$\boxed{\max\left[p_i, \cdots, p_{i+k+l-2}\right] \leqq \left(\frac{1}{\rho(1)}\right)^{i-k-l+1}.} \qquad (42)$$

Equations (41) and (42) give the bound for the mass of the distribution to the right of a particular state, which we shall now compare with a similar bound that holds for $\gamma' = \sqrt{\gamma}$. The prime superscript will be used on symbols to denote the functional dependence of the associated quantities on γ'. In establishing the reference (central) step size [see eq. (3)], minor differences exist depending on whether

$$(i) \qquad \Pr\left[|x(n)| \leqq \gamma^{i-1}\right] < \frac{l}{k+l} \leqq \Pr\left[|x(n)| \leqq \gamma^{i-1/2}\right]$$

or

$$(ii) \qquad \Pr\,[\,|x(n)|\, \leqq \gamma^{\bar{i}-1/2}] < \frac{l}{k+l} \leqq \Pr\,[\,|x(n)|\, \leqq \gamma^{\bar{i}}].$$

We consider only (*ii*), in which case: $\omega'(n) = 2i \Leftrightarrow \omega(n) = i$ and $b'_{2i} = b_i$ for all $i \geqq 0$.

Repeating the arguments leading to (41) and (42) we have

$$\Pr{}_s\,[\Delta \geqq C\sqrt{\gamma}^{2i+(k+l-2)}] \leqq \frac{\rho'(2i)}{\rho'(2i)-1} \max\,[p'_{2i}, \cdots, p'_{2i+k+l-2}] \tag{43}$$

and

$$\max\,[p'_{2i}, \cdots, p'_{2i+k+l-2}] \leqq \left[\frac{1}{\rho'(2)}\right]^{2i-k-l}. \tag{44}$$

Since $\rho'(2i) = \rho(i)$, we have

$$\Pr{}_s\,[\Delta \geqq C\sqrt{\gamma}^{2i+(k+l-2)}] \leqq \frac{\rho(i)}{\rho(i)-1} \left[\frac{1}{\rho(1)}\right]^{i-k-l+1} \left[\frac{1}{\rho(1)}\right]^{i-1}. \tag{45}$$

Comparison with (41) and (42) completes the demonstration.

IV. TRANSIENT RESPONSE

The preceding section discusses various aspects of the stationary distribution of the step size which effectively describes the steady-state behavior of the device. However, as stated before in Section I, the steady-state response is only of partial interest since the adaptability of the device is tied to quickness of response in the following situations:

(*i*) Start up—we are forced to consider situations in which the initial step size is fairly arbitrary.

(*ii*) Changes in the scaling of the input distribution—the scenario here is that the device has adapted to a particular intensity level (scaling) of the input distribution when a jump occurs to a new intensity level.

In common with both situations, we have an initial step size and a waiting time for the step size to adapt to the desired step size. Recall that with k and l appropriately chosen, the desirable step size is the central step size, which corresponds to the 0 state in the random walk, eq. (5). This aspect of the behavior of the device is also related to the rate at which the evolving step size distribution approaches the stationary distribution.

The main contribution of this section is the development of formulas for the efficient computation of the mean time required for the step size to first reach the central step size for various values of the initial step size. The designer can use the information generated by the methods given here in the following manner. Assuming that the designer has some understanding of the rate of variation of the intensity level of the input distribution, he is in a position to determine the smallest value of γ for which the adaptation algorithm adequately tracks the input process. The parameter γ has to be made sufficiently large for the mean waiting time (time, of course, is used synonymously with number of transitions) for adaptation to be small compared to the changes in the location of the desired step size arising from changes in the intensity level.

4.1 The mean time for first passage to the origin

We will consider the random walk, eqs. (5) and (12), for the saturating adaptive quantizer since in the limit, as L becomes large, the functionals obtained for this model yield corresponding quantities for the adaptive quantizer. Also, we shall consider only the case of the initial state $\omega(0) > 0$ since the results obtained can be transferred to the case of negative initial states in a fairly obvious manner (see Remark (*iv*) of Section 3.3).

Let the initial state $\omega(0) = i > 0$ and let M_i denote the mean time required for the first occurrence of the event $\omega(n) \leqq 0$. We observe that for all values of L, not necessarily finite, the time to first passage is finite with probability 1 as a consequence of the properties of recurrence and irreducibility established earlier in Section II. If the first transition results in a decrease of the step size, the process continues as if the initial state has been $(i - k)$. The conditional expectation of the first passage time, therefore, is $M_{i-k} + 1$. From this argument we deduce that the mean first passage time satisfies the recursion*

$$\boxed{\begin{aligned} M_i = b_i(M_{i-k} + 1) + a_i(M_{i+l} + 1) \\ \text{for} \quad (k+1) \leqq i \leqq (L - l). \end{aligned}} \qquad (46)$$

The relation in (46) may be used to generate the entire sequence $\{M_i\}$ provided the initial conditions are known. Now, by the same argument that led to (46), we have that (46) holds for $1 \leqq i \leqq k$ with

$$M_{1-k} = M_{2-k} = \cdots = M_0 = 0.$$

*There is some similarity between (46) and the equations arising in gambler's ruin problems[15] and sequential analysis,[16] in which generally $k = l = 1$ and the transition probabilities are not variable.

The remaining l boundary conditions, namely,

$$M_1, M_2, \cdots, M_l$$

are hard to obtain and it is necessary to look more deeply into the dynamics of the process to obtain these quantities.

For every sampling instant we define the L-dimensional vector $\mathbf{z}(n)$ with components $z_j(n)$, $1 \leqq j \leqq L$, where

$$z_j(n) \triangleq \Pr\,[\omega(n) = j \quad \text{and} \quad \omega(s) \geqq 1 \quad \text{for all } s \leqq n]. \tag{47}$$

These vectors, $\mathbf{z}(n)$, evolve with time according to

$$\mathbf{z}(n+1) = \mathbf{D}\mathbf{z}(n), \quad n \geqq 0. \tag{48}$$

These equations are given in Appendix B. Here we reproduce the structure of the $L \times L$ matrix $\mathbf{D}$:

$$\mathbf{D} = \begin{array}{r} l\left\{\begin{array}{c} \\ \\ \\ \end{array}\right. \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array}\left[\begin{array}{ccccccc}
\overbrace{0 \;\cdots\; 0}^{k} & b_{k+1} & & & & & \\
\vdots & & b_{k+2} & & & & \\
0 & & & \ddots & & & \\
a_1 & & & & \ddots & & \\
& a_2 & & & & \ddots & \\
& & \ddots & & & & b_L \\
& & & \ddots & & & 0 \\
& & & & & & \vdots \\
& & & & & & 0 \\
& & & a_{L-l} & a_{L-l+1} & \cdots & a_L
\end{array}\right].$$

Putting together various properties of the matrix $\mathbf{D}$ and the random walk, we obtain, in Appendix B, the following result: for $i \geqq 1$

$$\boxed{\begin{array}{l} M_i = \sum\limits_{j \geqq 1} x_j^{(i)}, \\ \text{where} \quad [\mathbf{I} - \mathbf{D}]\mathbf{x}^{(i)} = \mathbf{e}_i \end{array}} \tag{49}$$

and the elements of the vector $\mathbf{e}_i$ are zero everywhere except at the ith location where it is unity. In Appendix B it is shown that $[\mathbf{I} - \mathbf{D}]$ is nonsingular. We observe parenthetically the virtue of the recursion given in (46) in that it allows us to generate rather easily all the M_i's once the l inversions necessary to evaluate $M_1, \cdots, M_l$ are carried out.

The matrix inversion in (49) may be viewed as a mixed boundary value problem with the first l and the final k equations providing the boundary conditions. The bulk of the elements of the vector $\mathbf{x}^{(i)}$

satisfy a recursion that was encountered previously in Section III:

$$x_j^{(i)} = b_{j+k}x_{j+k}^{(i)} + a_{j-l}x_{j-l}^{(i)}. \tag{50}$$

Furthermore, we show in Appendix B that the elements $x_j^{(i)}$ are all nonnegative. Hence, we are in a position to usefully apply, even for infinite L, the techniques and results of Section III.

First, we carry out the reduction of the equations as stated in Section 3.1 where the motivation for this step is discussed. We obtain

$$\sum_{j=r+1}^{r+k} b_j x_j = \sum_{j=r-l+1}^{r} (1 - b_j)x_j, \quad l \leqq r \leqq (L - k). \tag{51}$$

The superscripts on the x's have not been used since (51) holds for all $\mathbf{x}^{(i)}$, $1 \leqq i \leqq l$.

One benefit of the above form is that it involves one less variable than the original recursion (50). In the important case of $k = 1$ and $l \geqq 1$, this reduction is sufficient to transform the original mixed boundary value problem (49) to an initial value problem, i.e., the solution to the matrix inversion problem (49) satisfies a recursion with specified initial conditions. Exact computation in this case becomes quite trivial. The details of this solution are given in Appendix C. Apart from its independent interest, this result is of particular interest in the adaptive quantizer when the distribution of the input sequence is Gaussian. As discussed previously, it is desirable to have in this case $l/(k + l) = 0.68$, and $k = 1$ and $l = 2$ will suffice.

Another property of the solutions $x^{(i)}$ of (49) which holds for all L is that with increasing j, $x_j^{(i)}$ decreases at least geometrically. This conclusion may be drawn from the bounds obtained in Section 3.3, eqs. (35) and (38). From the point of view of numerical inversion of $[\mathbf{I} - \mathbf{D}]$ for large L, this is a critical property in that it is a necessary condition for most numerical techniques. The reader is referred to Richtmyer and Morton[17] for one such technique that we have used successfully and found to be efficient in that it effectively exploits the band structure of the matrix $[\mathbf{I} - \mathbf{D}]$.

Finally, we remark that while we have dealt exclusively with first passage across the 0 state it is clear that generalizations to first crossings across states other than the 0 state is straightforward.

4.2 Bound on the mean first passage time

Two formulas, eqs. (46) and (49), have been given for computing the mean time required for the step size to adapt from an arbitrary initial value to the desired, and also central, step size. However, by examining these formulas it is not easy to gain insights into the rate at which this adaptation time grows with the distance separating the two states and its dependence on γ. Here, by probabilistic reasoning, we obtain an explicit upper bound on this time and this bound does provide some insight. As we have done before, we consider here only the case of positive initial states, i.e., $\omega(0) > 0$. Let M_{ij}, $0 \leqq i < j$, denote the mean first passage time under the following conditions: the initial state $\omega(0) = j$ and first crossing occurs after τ transitions if $\omega(\tau) \leqq i$ and $\omega(n) > i$ for all $n < \tau$; then $M_{ij} = E(\tau)$. In this notation the quantity M_j defined in Section 4.1 is equivalent to M_{0j}.

In Section 1.2, eq. (10), it is given that

$$E[\omega(n+1) \mid \omega(n) = i] - i = -(k+l)\left[b_i - \frac{l}{k+l}\right]. \tag{52}$$

Denote the quantity on the right by $-S_i$ and observe that for $i > 0$, $S_{i+1} > S_i > 0$; hence, the supermartingale property. [For the saturating adaptive quantizer, the supermartingale property holds even more strongly, i.e., for $i > 0$, (52) holds with the equality replaced by $\leqq$.] In fact, the supermartingale property holds for the transformed process: $\omega'(n) = \omega(n) + nS_{i+1}$. i.e.,

$$E[\omega'(n+1) \mid \omega'(n)] \leqq \omega'(n) \tag{53}$$

for all $\omega'(n) \geqq (i+1) + nS_{i+1}$. For the crossing problem, (53) holds for all $(n+1) \leqq \tau$, the crossing time. We can now apply a theorem due to Doob[18] on optional stopping on supermartingales. In this case, the theorem states that

$$E[\omega'(\tau)] \leqq E[\omega'(0)]. \tag{54}$$

Since

$$(i + 1 - k) + S_{i+1}E(\tau) \leqq E[\omega'(\tau)] \leqq E[\omega'(0)] = j,$$

we obtain

$$\boxed{M_{ij} = E(\tau) \leqq \frac{1}{S_{i+1}}[(j-i) + (k-1)].} \tag{55}$$

We gain some insight on the role of γ in determining the transient response of the device by observing the dependence of the above bound on γ. Suppose we are interested in M_{0j}, the waiting time for the initial step size $\Delta(0) = C\gamma^j$ to reach the central step size C. Consider the effects of making $\gamma' = \sqrt{\gamma}$ on this waiting time (the multiplier coefficients of the device are therefore $\sqrt{\gamma}^{-k}$ and $\sqrt{\gamma}^{l}$). We let the prime superscript on symbols indicate a functional dependence on γ'. In establishing the new central step size [see eq. (3)], minor differences exist depending on whether

$$(i) \quad \Pr\left[|x(n)| \leqq \gamma^{i-1}\right] < \frac{l}{k+l} \leqq \Pr\left[|x(n)| \leqq \gamma^{i-\frac{1}{2}}\right]$$

or

$$(ii) \quad \Pr\left[|x(n)| \leqq \gamma^{i-\frac{1}{2}}\right] < \frac{l}{k+l} \leqq \Pr\left[|x(n)| \leqq \gamma^{i}\right].$$

We consider only (ii), in which case the central step sizes are identical: $\omega'(n) = 2i \Leftrightarrow \omega(n) = i$ and $b'_{2i} = b_i$ for all $i \geqq 0$. The waiting time

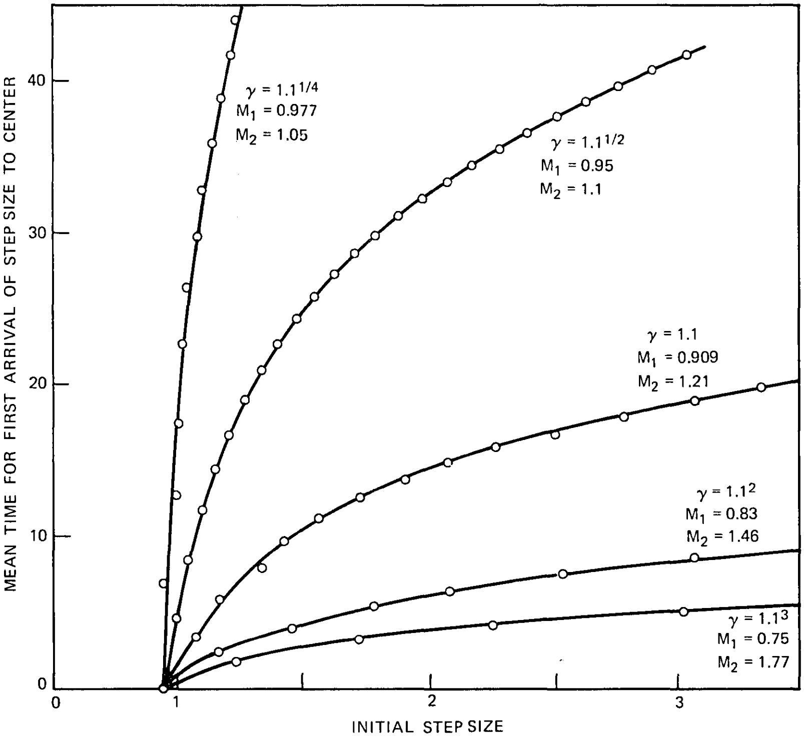


Fig. 3—Transient response of the adaptive quantizer.

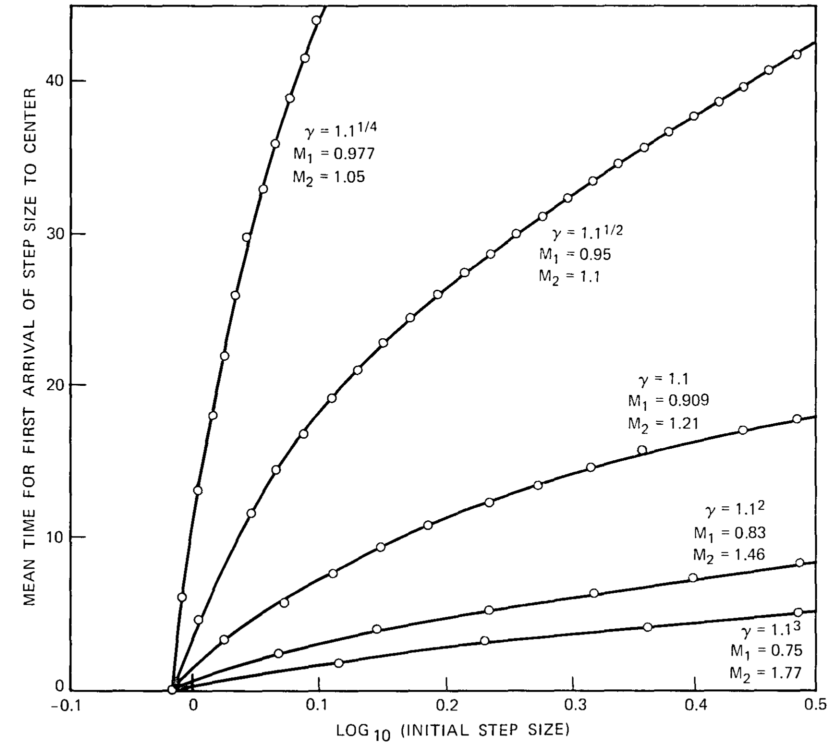


Fig. 4—Transient response of the adaptive quantizer.

for the step size to adapt from identical initial step size $C\gamma^j$ to final step size C is $M'_{0,2j}$. From (55),

$$M'_{0,2j} \leqq \frac{1}{S'_1}[2j - (k - 1)].$$

Now, $S_0 \leqq S'_1 \leqq S_1$; hence, making $\gamma' = \sqrt{\gamma}$ and keeping k and l unchanged has the effect of making the bound on the waiting time at least twice as large for $j \gg k$. This is a conclusion which is plausible in the light of the linear form of the bound (55) since the effect of making $\gamma' = \sqrt{\gamma}$ is to introduce twice as many transitions between the initial and final step sizes.

4.3 Computational results

We present here a sampling of our computational results. It is assumed that for every n, $x(n)$ is normally distributed with unit variance. The optimal step size $\hat{\Delta}$ in this case has the property that $\Pr\{|x(n)| \leqq \hat{\Delta}\} = 0.68$. To center the stationary distribution of the step size close to the optimal step size, we choose $k = 1$ and $l = 2$.

Figure 3 plots the mean time for first passage to the optimal step size vs. initial step size, and the initial step sizes chosen for this figure exceed the optimal step size. Various values of $\gamma(M_1 = \gamma^{-k}, M_2 = \gamma^l)$ were used. Figure 4 provides the same information except that the horizontal axis corresponds to $\log_{10} \Delta(0)$, rather than $\Delta(0)$ as in Fig. 3. The mean first passage times M_1 and M_2 were obtained by the method outlined in Appendix C, and M_i, $i \geqq 3$ were generated by using the recursion in (46). To give some idea of the rate of convergence for $x_j^{(1)}$, eqs. (70) and (71), we tabulate some values of $x_j^{(1)}$ for the case of

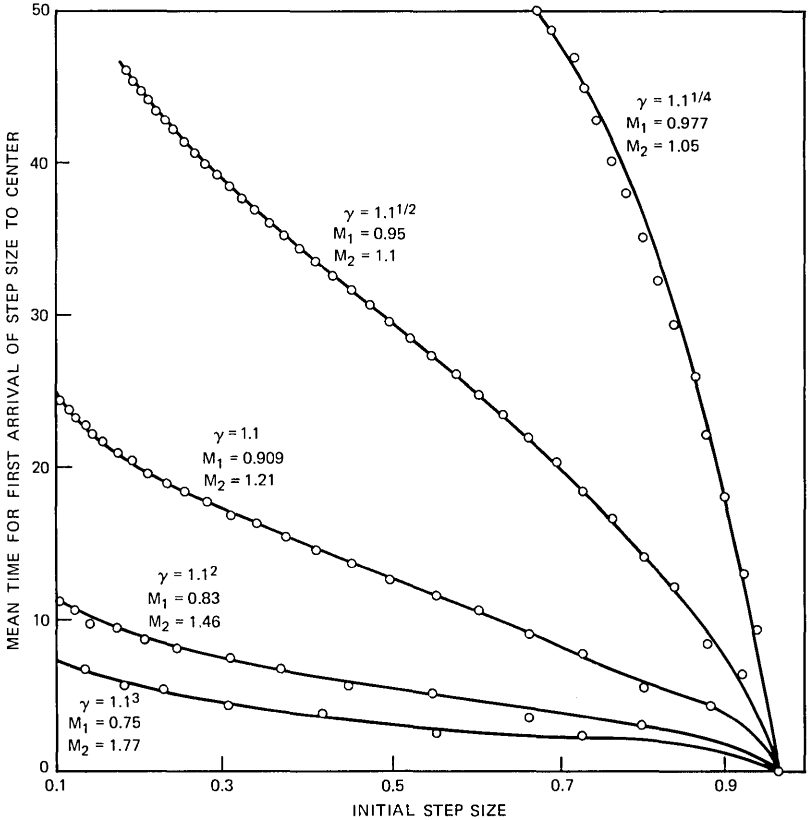


Fig. 5—Transient response of the adaptive quantizer.

$\gamma = 1.1$:

j:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$x_j^{(1)}$:	1.4	0.53	0.66	0.31	0.20	0.08	0.03	0.92×10^{-2}	0.24×10^{-2}	0.41×10^{-3}	0.59×10^{-4}	0.53×10^{-5}	0.35×10^{-6}	0.13×10^{-7}	0.30×10^{-9}	0.31×10^{-11}

Figure 5 is similar to Fig. 3 except that here the initial step sizes are less than the optimal step size. Figure 6 plots the same information with $\log_{10} \Delta(0)$, rather than $\Delta(0)$, on the horizontal axis. The mean first passage time M_1 was obtained by solving (49) by the method given in Ref. 17 and all other first passage times were generated by the recursion in (46).

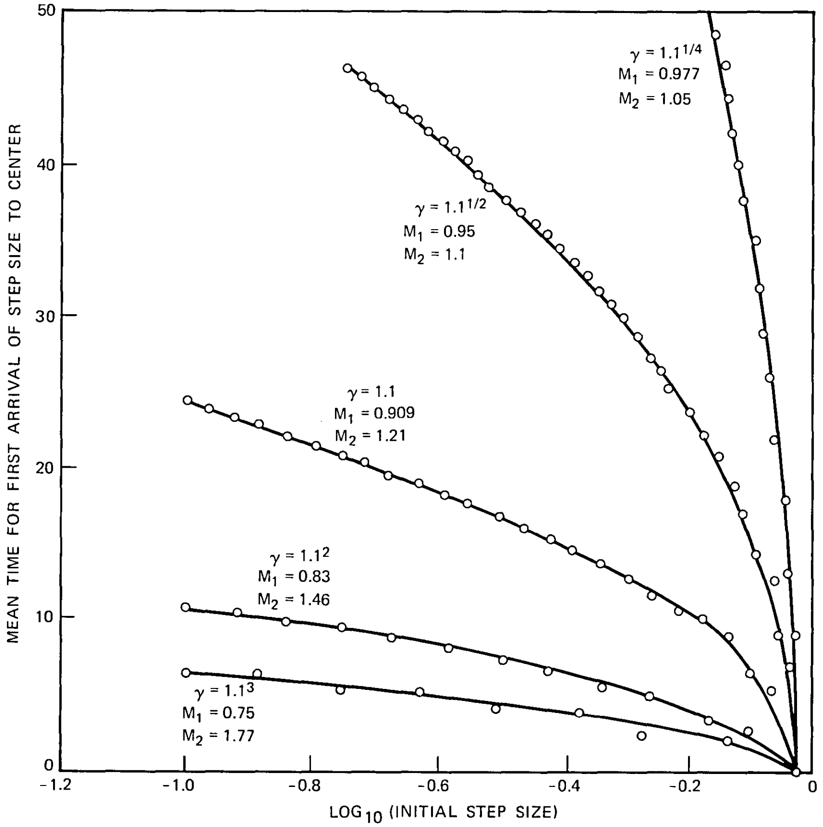


Fig. 6—Transient response of the adaptive quantizer.

V. ACKNOWLEDGMENTS

The criticism and comments of my colleague, Mohan Sondhi, readily and generously given, were always beneficial to me. I benefited greatly from David Goodman's insight while the problem was being formulated. I am grateful to Larry Shepp for the information that Doob's theorem on optional stopping of supermartingales can be used effectively in obtaining the bound on the mean first passage time given in Section 4.2. The benefit of stimulating discussions with N. S. Jayant is acknowledged.

APPENDIX A

Proof of Lemma 1

Proof:

(*i*) $\tilde{\mathbf{A}}_i^{-1}$ being in the form of a companion matrix, the coefficients of the characteristic polynomial of the matrix are the elements of the first row:

$$\begin{aligned} C(\mu) &\triangleq (-1)^{k+l-1} \det [\tilde{\mathbf{A}}_i^{-1} - \mu \mathbf{I}] \\ &= \mu^{k+l-1} + \cdots + \mu^k - [\alpha_1 \mu^{k-1} + \alpha_2 \mu^{k-2} + \cdots + \alpha_k], \end{aligned} \tag{56}$$

where

$$\alpha_1 = \frac{b_{i+l}}{a_i}, \quad \alpha_2 = \frac{b_{i+l+1}}{a_i}, \quad \cdots, \quad \alpha_k = \frac{b_{i+l+k-1}}{a_i}. \tag{57}$$

By Descartes's rule the polynomial $C(\mu)$ has at most one positive real root. Since $C(0) = -\alpha_k < 0$ and $C(\mu) \to \infty$ as $\mu \to \infty$, there exists exactly one positive root. Let r denote this root.

Now $C(1) < 0$ if $la_i < (b_{i+l} + b_{i+l+1} + \cdots + b_{i+l+k-1})$. The latter condition holds for all $i \geqq 0$. Hence, $r > 1$.

(*ii*) The left eigenvector $\boldsymbol{\lambda}$ corresponding to the eigenvalue r satisfies, by definition, $\boldsymbol{\lambda}^t \tilde{\mathbf{A}}_i^{-1} = r\boldsymbol{\lambda}^t$. Examining the component equations we find that

$$\lambda_i = \lambda_1(1 + r + \cdots + r^{i-1}) \quad 1 \leqq i \leqq l. \tag{58}$$

Also,

$$\lambda_{l+k-i} = \frac{\lambda_{l+k-1}}{\alpha_k r^{i-1}} [\alpha_{k-i+1} r^{i-1} + \cdots + \alpha_{k-1} r + \alpha_k] \quad 1 \leqq i \leqq k. \tag{59}$$

Finally, $r\lambda_{l+k-1} = \alpha_k \lambda_1$. Since the α's are positive quantities, the statement is clearly true.

(*ii*) The statement can be verified by inspecting the characteristic polynomial $C(\mu)$ and using the fact that the coefficients $\alpha_1, \cdots, \alpha_k$ each increase with i.

APPENDIX B

Derivation of equations (48) and (49)

The derivation of the equations governing the evolution of the vectors $\mathbf{z}(n)$ defined in eq. (47) proceeds as follows. For convenience, let $X(n)$ denote the event $1 \leqq \omega(\tau) \leqq L$ for all τ, $0 \leqq \tau \leqq n$. Hence, by definition,

$$z_j(n) = \Pr[\omega(n) = j \quad \text{and} \quad X_n] \quad 1 \leqq j \leqq L.$$

Since

$$\begin{aligned} z_j(n) &= \Pr[\omega(n) = j \quad \text{and} \quad X_{n-1}] \\ &= \sum_{i=1}^{L} \Pr[\omega(n) = j \,|\, \omega(n-1) = i, X_{n-1}] z_i(n-1) \\ &= \begin{cases} b_{k+j}\, z_{k+j}(n-1), \quad 1 \leqq j \leqq l, \\ a_{j-l}\, z_{j-l}(n-1) + b_{j+k} z_{j+k}(n-1), \quad (l+1) \leqq j \leqq (L-k) \\ a_{j-l}\, z_{j-l}(n-1), \quad (L-k+1) \leqq j \leqq (L-1), \\ \displaystyle\sum_{i=L-l}^{L} a_i z_i(n-1) \quad j = L. \end{cases} \end{aligned}$$

The above equations define the matrix $\mathbf{D}$ which relates $\mathbf{z}(n)$ to $\mathbf{z}(n-1)$ as in eq. (48).

For the derivation of eq. (49) we proceed as follows. For $i = 1, 2, \cdots, L$, let

$$\begin{aligned} F_i(n+1) &\triangleq \Pr[\text{first passage occurs at } (n+1) \,|\, \omega(0) = i] \\ &= \Pr[\omega(n+1) \leqq 0, \quad X_n \,|\, \omega(0) = i] \\ &= \sum_{j=1}^{k} b_j\, z_j(n) \quad \text{with} \quad \mathbf{z}(0) = \mathbf{e}_i. \end{aligned} \tag{60}$$

The vector $\mathbf{e}_i$ has every element equal to zero except for the ith element which is unity. To express eq. (60) in vector form we let $\mathbf{b} \triangleq [b_1 b_2 \cdots b_k\ 0 \cdots 0]^t$. Then, from (60),

$$F_i(n+1) = \mathbf{b}^t \mathbf{z}(n) \quad \text{with} \quad \mathbf{z}(0) = \mathbf{e}_i.$$

By definition, we have that the mean first passage time conditional on the initial state being i,

$$\begin{aligned} M_i &= \sum_{n \geqq 0} (n+1) F_i(n+1) \\ &= \mathbf{b}^t \sum_{n \geqq 0} (n+1)\mathbf{z}(n) \\ &= \mathbf{b}^t \sum_{n \geqq 0} n\mathbf{z}(n) + \mathbf{b}^t \sum_{n \geqq 0} \mathbf{z}(n). \end{aligned} \tag{61}$$

Now the second term in the above expression is unity since the probability that passage occurs at finite time is unity. Now consider

$$\begin{aligned} [\mathbf{I} - \mathbf{D}] \sum_{n \geqq 1} n\mathbf{z}(n) &= \sum_{n \geqq 1} n\mathbf{z}(n) - \sum_{n \geqq 1} n\mathbf{z}(n+1) \\ &= \sum_{n \geqq 0} \mathbf{z}(n) - \mathbf{z}(0). \end{aligned} \tag{62}$$

Hence, denoting by $\mathbf{1}$ the column vector with every element equal to unity, we have from (62) that

$$\mathbf{1}^t[\mathbf{I} - \mathbf{D}] \sum_{n \geqq 1} n\mathbf{z}(n) = \mathbf{1}^t \sum_{n \geqq 0} \mathbf{z}(n) - 1 \tag{63}$$

$$= \mathbf{b}^t \sum_{n \geqq 1} n\mathbf{z}(n), \tag{64}$$

since $\mathbf{1}^t\mathbf{z}(0) = 1$ and $\mathbf{b}^t = \mathbf{1}^t[\mathbf{I} - \mathbf{D}]$. It only remains to consider

$$\sum_{n \geqq 0} \mathbf{z}(n) = \left[\sum_{i=0}^{\infty} \mathbf{D}^i \right] \mathbf{z}(0).$$

The above series converges since every eigenvalue of the matrix $\mathbf{D}$ lies strictly within the unit circle in the complex plane. The proof of this follows from an old matrix theorem[19] which states that if the diagonal elements of the columns weakly dominate the sum of the absolute values of the off-diagonal elements with strong dominance holding for at least one column and the matrix is irreducible, then the determinant is nonzero. Applying this theorem to $[\mathbf{D} - \lambda\mathbf{I}]$, $|\lambda| \geqq 1$, we note that the irreducibility of the original Markov chain implies irreducibility of the matrix $[\mathbf{D} - \lambda\mathbf{I}]$ and that the weak column dominance property holds everywhere while the strong column dominance property holds for the first k columns. Hence,

$$\sum_{n \geqq 0} \mathbf{z}(n) = \left[\sum_{i \geqq 0} \mathbf{D}^i \right] \mathbf{z}(0) = [\mathbf{I} - \mathbf{D}]^{-1}\mathbf{z}(0). \tag{65}$$

Putting together the above results we have (49), namely,

$$M_i = \sum_{j \geqq 1} x_j^{(i)} \quad \text{where} \quad [\mathbf{I} - \mathbf{D}]\mathbf{x}^{(i)} = \mathbf{e}_i.$$

Observe that $\mathbf{x}^{(i)} = \Sigma \mathbf{z}(n)$ and, from the definition of $\mathbf{z}(n)$, it follows that every element of $\mathbf{x}^{(i)}$ is nonnegative.

APPENDIX C

Mean first passage times for the case $k = 1$, $l \geqq 1$

We have as our starting point eq. (49), namely,

$$M_i = \sum_j x_j^{(i)}, \tag{66}$$

$$\text{where} \quad [\mathbf{I} - \mathbf{D}]\mathbf{x}^{(i)} = \mathbf{e}_i \tag{67}$$

and we are interested only in $1 \leqq i \leqq l$.

The transformation that was made in Section 3.1 is equivalent to the following: add to each row, r, of $[\mathbf{I} - \mathbf{D}]$ all rows $r + 1, r + 2, \cdots$; and do the same to the vector $\mathbf{e}_i$. This operation makes the matrix $[\mathbf{I} - \mathbf{D}]$ lower triangular, the reason being that with the exception of the first column, the elements of all other columns of $[\mathbf{I} - \mathbf{D}]$ sum to zero. The resulting equations are as follows: the first component equation yields

$$b_1 x_1^{(i)} = 1, \tag{68}$$

and the next $(l - 1)$ equations: $2 \leqq r \leqq l$,

$$-\sum_{j=1}^{r-1} a_j x_j^{(i)} + b_r x_r = \begin{cases} 1 & \text{if} \quad r \leqq i \\ \\ 0 & \text{if} \quad r > i. \end{cases} \tag{69}$$

Finally,

$$\boxed{x_r^{(i)} = \frac{1}{b_r} \sum_{j=r-l}^{r-1} a_j x_j^{(i)} \quad \text{for} \quad r > l.} \tag{70}$$

The boundary conditions to the basic recursion in (70) are in (68) and (69) which are, of course, solvable:

$$\boxed{\begin{array}{ll} 1 \leqq r \leqq i & x_r^{(i)} = 1/\prod_{j=1}^{r} b_j \\ (i + 1) \leqq r \leqq l & x_r^{(i)} = (x_i^{(i)} - 1)/\prod_{j=i+1}^{r} b_j. \end{array}} \tag{71}$$

REFERENCES

1. N. S. Jayant, "Adaptive Quantization with a One-Word Memory," B.S.T.J., *52*, No. 7 (September 1973), pp. 1119–1144.
2. P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J. *52*, No. 7 (September 1973), pp. 1105–1118.
3. M. R. Winkler, "High Information Delta Modulation," IEEE Int. Conv. Record, part 8, 1963, pp. 260–265.
4. J. E. Abate, "Linear and Adaptive Delta Modulation," Proc. IEEE, March 1967, pp. 298–307.
5. N. S. Jayant, "Adaptive Delta Modulation with a One-Bit Memory," B.S.T.J., *49*, No. 3 (March 1970), pp. 321–342.
6. R. M. Wilkinson, "An Adaptive Pulse Code Modulator for Speech," Proc. Int. Conf. Commun., Montreal, June 1971, pp. 1-11 to 1-15.
7. D. J. Goodman and A. Gersho, "Theory of an Adaptive Quantizer," Proc. of December 1973 IEEE Symp. on Adaptive Processes, Decision and Control.
8. J. Max, "Quantization for Minimum Distortion," Trans. IRE, *IT-6* (March 1960), pp. 7–12.
9. T. Fine, "The Response of a Particular Nonlinear System with Feedback to Each of Two Random Processes," IEEE Trans. Inform. Theory, *IT-14*, No. 2 (March 1968), pp. 255–264.
10. A. Gersho, "Stochastic Stability of Delta Modulation," B.S.T.J., *51*, No. 4 (April 1972), pp. 821–842.
11. R. S. Bucy, "Stability and Positive Super-Martingales," J. Differential Equations, *1*, 1965, pp. 151–155.
12. H. Kushner, *Introduction to Stochastic Control*, New York: Holt, Rinehart and Winston, 1971.
13. W. Feller, *Introduction to Probability Theory and Its Applications*, vol. 1, 3rd ed., New York: John Wiley & Sons, 1950, p. 387.
14. Ref. 13, p. 402.
15. Ref. 13, pp. 348–349.
16. A. Wald, *Sequential Analysis*, New York: John Wiley & Sons, 1947.
17. R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems*, 2nd ed., New York: John Wiley & Sons, 1957, pp. 198–201.
18. J. L. Doob, *Stochastic Processes*, New York: John Wiley & Sons, 1953, pp. 300–301.
19. M. Marden, "Geometry of Polynomials," Mathematical Surveys 3, American Mathematical Society, Providence, R.I., 1966, pp. 140–141.

Copyright © 1974 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 5, May–June 1974
Printed in U.S.A.

Spectra of Digital Phase Modulation by Matrix Methods

By V. K. PRABHU and H. E. ROWE

(Manuscript received October 31, 1973)

We derive the spectral density of a sinusoidal carrier phase modulated by a random baseband pulse train in which the signaling pulse duration is finite and the signaling pulses may have different shapes. The spectral density is expressed as a compact Hermitian form in which the Hermitian matrix is a function of only the symbol probability distribution, and the associated column vector is a function of only the signal pulse shapes. If the baseband pulse duration is longer than one signaling interval, we assume that the symbols transmitted during different time slots are statistically independent. The applicability of the method to compute the spectral density is illustrated by examples for binary, quaternary, octonary, and 16-ary PSK systems with different pulse overlap. Similar methods yield the spectral density of the output of a nonlinear device whose input is a random baseband pulse train with overlapping pulses.

I. INTRODUCTION

In recent years, digital phase-modulation techniques have been playing an increasingly important role in the transmission of information in radio and waveguide systems. Various methods have been developed recently for computing spectra of a sinusoidal carrier phase modulated by a random baseband pulse train.[1–8]

In this paper, we derive the spectral density of a carrier phase modulated by a random baseband pulse stream in which the signaling pulse duration is finite and the signaling pulses may have different shapes. The spectral density is expressed as a compact Hermitian form in which the Hermitian matrix is a function of only the symbol probability distribution and the associated column vector is a function of only the signal pulse shapes. If the baseband pulse duration is longer than one signaling interval and the pulses from different time slots overlap, we assume that the symbols transmitted during different time slots are statistically independent.

The present method also yields the spectral density of the output of a nonlinear device whose input is a similar baseband pulse train.

The work reported here generalizes and simplifies prior results. The form of the present results provides an appropriate division between analysis and machine computation that enhances physical understanding and simplifies numerical computations. We compute the spectra of binary, quaternary, octonary, and 16-ary PSK systems, with overlapping baseband modulation pulses of several shapes.

II. M-ARY PHASE-MODULATED SIGNALS

We seek the spectrum of the digital phase-modulated wave

$$x(t) = \cos[2\pi f_c t + \phi(t)], \quad f_c > 0, \tag{1}$$

where

$$\phi(t) = \sum_{k=-\infty}^{\infty} g_{s_k}(t - kT), \quad s_k = 1, 2, \cdots, M. \tag{2}$$

The discrete random process s_k is assumed strictly stationary; as noted in (2), it takes on only integer values from 1 to M. The carrier frequency is f_c. The signaling alphabet consists of M time functions, $g_1, g_2, \cdots, g_M$, that may have different shapes; one of these is transmitted for each signaling interval T, to generate the digital baseband phase modulation $\phi(t)$. The different signaling waveforms in (2) may overlap, and may be statistically dependent throughout the present section and Appendix A.

For convenience, we define[9]

$$v(t) \equiv e^{j\phi(t)}; \tag{3}$$

then

$$x(t) = \text{Re}\,\{e^{j2\pi f_c t}\, v(t)\}. \tag{4}$$

Appendix A shows that the spectrum of $x(t)$ is

$$\mathbf{P}_x(f) = \tfrac{1}{4}\mathbf{P}_v(f - f_c) + \tfrac{1}{4}\mathbf{P}_v(-f - f_c),$$
$$2f_c \neq \frac{n}{T}, \quad n = 1, 2, 3, \cdots, \tag{5}$$

where[9]

$$\mathbf{P}_v(f) = \int_{-\infty}^{\infty} \Phi_v(\tau) e^{-j2\pi f \tau} d\tau, \tag{6}$$

$$\Phi_v(\tau) = \overline{\Phi_v(t, \tau)} \equiv \lim_{A \to \infty} \frac{1}{2A} \int_{-A}^{A} \Phi_v(t, \tau) dt, \tag{7}^{\dagger}$$

$$\Phi_v(t, \tau) = \langle v(t + \tau) v^*(t) \rangle = \langle e^{j[\phi(t+\tau) - \phi(t)]} \rangle. \tag{8}$$

† The symbol $\overline{}$ denotes average on t throughout.

The first term of (5) is the spectrum of the complex baseband wave $v(t)$ shifted to the carrier frequency $+f_c$; the second term is the spectrum of $v^*(t)$ shifted to $-f_c$. The spectral relationship of (5) is strictly true as long as the condition on f_c is satisfied, whether or not the two spectral terms overlap; that is, this result applies strictly to both narrow-band [$f_c \gg$ bandwidth of $\mathbf{P}_v(f)$] and wideband modulated waves.

The condition of (5) requires that twice the carrier frequency is not an integral multiple of the signaling rate $1/T$. $\mathbf{P}_v(f)$ has in general both continuous and line spectra, the lines occurring at frequencies n/T for integer n. The condition in (5) guarantees that the line components of the two spectral terms never coincide. In the exceptional case, where the two sets of lines do coincide, it is not surprising that it no longer suffices merely to add powers of the two terms, as in (5); however, if f_c is high enough the modulated wave is narrow band, the two spectral terms of (5) do not overlap significantly, and (5) provides a good approximation even if the condition is violated.

Note that we have not found it necessary to randomize the phase of the unmodulated carrier or the position of the time slots of the digital modulation, as is done in various other studies. Thus, rather than (1) we might have considered

$$x(t) = \cos\left[2\pi f_c t + \phi(t - t_0) + \phi_0\right], \tag{9}$$[†]

with $\phi(t)$ still given by (2). The spectral relation of (5) will again be strictly true when $2f_cT$ = integer if ϕ_0 and t_0 are independent, and either is uniformly distributed over a suitable interval (Appendix A). However, we retain the original formulation of (1) and (2) throughout —corresponding to $\phi_0 = 0$, $t_0 = 0$ in (9)—to determine the effect of deterministic carrier and modulation phase on the statistics of the modulated wave.

Consequently, we study only $\mathbf{P}_v(f)$ throughout the remainder of this paper. This suffices for all cases except a low carrier frequency f_c that is a precise multiple of half the baud rate $1/2T$; the additional calculations necessary for this exceptional case are suggested in Appendix A, but not carried out in detail.

III. NOTATION AND STATISTICAL MODEL

The introduction of vector-matrix notation greatly simplifies the ensuing analysis.

[†] Either of the two parameters, ϕ_0 or t_0, shifts the relative phase between carrier and modulation; thus, only one of them is really necessary.

First, we rewrite the phase modulation of (2) as

$$\phi(t) = \sum_{k=-\infty}^{\infty} [a_k^{(1)}g_1(t-kT) + a_k^{(2)}g_2(t-kT) + \cdots + a_k^{(M)}g_M(t-kT)]. \qquad (10)$$

For a given k (i.e., for a given time slot), one of the a_k's is unity and the rest are zero;

$$\begin{aligned} a_k^{(s_k)} &= 1, \\ a_k^{(i)} &= 0, \quad i \neq s_k, \end{aligned} \qquad (11)$$

where s_k is the strictly stationary, discrete random process of (2), taking on the integer values 1, 2, $\cdots$, M.

Now we define for convenience the row vectors

$$\begin{aligned} \underline{\mathbf{a}_k} &\equiv [a_k^{(1)}\ a_k^{(2)}\ \cdots\ a_k^{(M)}], \\ \underline{\mathbf{g}(t)} &\equiv [g_1(t)\ g_2(t)\ \cdots\ g_M(t)], \end{aligned} \qquad (12)^{\dagger}$$

whose components are respectively the coefficients and pulse shapes of (10). The transposes[‡] of these row vectors are the column vectors

$$\mathbf{a}_k] = \underline{\mathbf{a}_k}' = \begin{bmatrix} a_k^{(1)} \\ a_k^{(2)} \\ \vdots \\ a_k^{(M)} \end{bmatrix}, \quad \mathbf{g}(t)] = \underline{\mathbf{g}(t)}' = \begin{bmatrix} g_1(t) \\ g_2(t) \\ \vdots \\ g_M(t) \end{bmatrix}. \qquad (13)$$

Define the unit basis row vectors

$$\begin{aligned} \underline{\mathbf{e}_1} &\equiv [1\ \ 0\ \ 0\ \ \cdots 0], \\ \underline{\mathbf{e}_2} &\equiv [0\ \ 1\ \ 0\ \ \cdots 0], \\ &\ \vdots \\ \underline{\mathbf{e}_M} &\equiv [0\ \ 0 \cdots\ \ 0\ \ 1], \end{aligned} \qquad (14)$$

with corresponding transpose column vectors $\mathbf{e}_1]$, $\mathbf{e}_2]$, $\cdots$, $\mathbf{e}_M]$. Then $\underline{\mathbf{a}_k}$ takes on only the values $\underline{\mathbf{e}_1}$, $\underline{\mathbf{e}_2}$, $\cdots$, $\underline{\mathbf{e}_M}$ by (11):

$$\underline{\mathbf{a}_k} = \underline{\mathbf{e}_{s_k}}. \qquad (15)$$

Now we rewrite (10) in vector notation as

$$\phi(t) = \sum_{k=-\infty}^{\infty} \underline{\mathbf{a}_k}\cdot\mathbf{g}(t-kT)], \qquad (16)$$

where $\cdot$ signifies ordinary matrix multiplication throughout. The term

[†] Boldface quantities denote matrices throughout. Row and column vectors are distinguished by the additional notation $\sqcup$ and], respectively.

[‡] The transpose of a matrix is indicated by the symbol $'$.

$\lfloor \mathbf{a}_k \rfloor$ is a vector-valued, discrete[†] random process, strictly stationary by the assumed strict stationarity of s_k of (2). We define the first- and second-order probabilities of $\lfloor \mathbf{a}_k \rfloor$ or $\mathbf{a}_k]$ as

$$w_i = \Pr\{\mathbf{a}_k = \mathbf{e}_i\} = \Pr\{s_k = i\} \geqq 0. \tag{17}^{\ddagger}$$

$$W_n(i, j) \equiv \Pr\{\mathbf{a}_k = \mathbf{e}_i, \mathbf{a}_{k+n} = \mathbf{e}_j\} = \Pr\{s_k = i, s_{k+n} = j\} \geqq 0. \tag{18}^{\ddagger}$$

w_i is the probability that the ith signaling waveform g_i is transmitted in any time slot, $W_n(i, j)$ the joint probability that the signaling waveforms g_i and g_j are transmitted in two time slots separated by n signaling intervals. w_i and $W_n(i, j)$ are independent of k by the assumption of stationarity. Then since the marginal probabilities are obtained by summing over the joint probability function,

$$\sum_{i=1}^{M} W_n(i, j) = w_j, \quad \sum_{j=1}^{M} W_n(i, j) = w_i. \tag{19}$$

Normalization of the total probability to 1 requires

$$\sum_{i=1}^{M} w_i = 1, \quad \sum_{i=1}^{M} \sum_{j=1}^{M} W_n(i, j) = 1. \tag{20}$$

Now we introduce vector-matrix notation for the probabilities. Let

$$\lfloor \mathbf{w} \rfloor \equiv [w_1 \ w_2 \cdots w_M] \tag{21}$$

be the probability row vector whose elements give the probabilities of the different signaling waveforms, with transpose column vector

$$\mathbf{w}] = \lfloor \mathbf{w} \rfloor'. \tag{22}$$

Let

$$\mathbf{W}_n \equiv \begin{bmatrix} W_n(1, 1) & W_n(1, 2) & \cdots & W_n(1, M) \\ W_n(2, 1) & W_n(2, 2) & \cdots & W_n(2, M) \\ \vdots & \vdots & \cdots & \vdots \\ W_n(M, 1) & W_n(M, 2) & \cdots & W_n(M, M) \end{bmatrix} \tag{23}$$

be the matrix whose elements specify the joint probabilities of all pairs of signaling waveforms separated by n time slots. Further, define

$$\lfloor \mathbf{1} \rfloor = \mathbf{1}]' \equiv [1 \ 1 \cdots 1] \tag{24}$$

as a vector with all M elements unity. Then (19) and (20) may be written as

$$\lfloor \mathbf{1} \rfloor \cdot \mathbf{W}_n = \lfloor \mathbf{w} \rfloor, \qquad \mathbf{W}_n \cdot \mathbf{1}] = \mathbf{w}] \tag{25}$$

[†] $\lfloor \mathbf{a}_k \rfloor$ is defined only for integer values of its independent variable k.

[‡] It is understood that $\mathbf{a}$, $\mathbf{e}$ are either row or column vectors throughout (17) and (18), and in succeeding equations.

and

$$\underline{\mathbf{1}}\cdot\mathbf{w}] = \underline{\mathbf{w}}\cdot\mathbf{1}] = 1, \qquad \underline{\mathbf{1}}\cdot\mathbf{W}_n\cdot\mathbf{1}] = 1. \tag{26}$$

The probability matrices have the following useful properties. For $n = 0$, the joint probability matrix of (23) and (18) is diagonal, with diagonal elements the first-order symbol probabilities;

$$\mathbf{W}_0 = \begin{bmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_M \end{bmatrix} \equiv \mathbf{w}_d, \tag{27}$$

where we define the diagonal matrix $\mathbf{w}_d$ for later convenience. Then

$$\mathbf{w}_d\cdot\mathbf{1}] = \mathbf{w}], \quad \underline{\mathbf{1}}\cdot\mathbf{w}_d = \underline{\mathbf{w}}. \tag{28}$$

By symmetry,

$$W_n(i, j) = W_{-n}(j, i), \tag{29}$$

or in matrix notation

$$\mathbf{W}_n = \mathbf{W}_{-n}'. \tag{30}$$

We assume that signaling waveforms in widely separated time slots are statistically independent:

$$\lim_{n\to\infty} W_n(i, j) = w_i\, w_j, \tag{31}$$

or in matrix form

$$\lim_{n\to\infty} \mathbf{W}_n = \mathbf{w}]\cdot\underline{\mathbf{w}}. \tag{32}$$

Since

$$(\mathbf{w}]\cdot\underline{\mathbf{w}})' = \mathbf{w}]\cdot\underline{\mathbf{w}}, \tag{33}$$

(30) and (32) yield

$$\mathbf{W}_\infty = \mathbf{W}_\infty' = \mathbf{W}_{-\infty}. \tag{34}$$

Using the above, the mean and covariance of the strictly stationary, vector-valued, discrete random process $\mathbf{a}_k$ are

$$\langle \mathbf{a}_k]\rangle = \mathbf{w}], \quad \langle \underline{\mathbf{a}_k}\rangle = \underline{\mathbf{w}}, \tag{35}$$

$$\tilde{\mathbf{\Phi}}_{\mathbf{a}}(n) \equiv \langle \mathbf{a}_{k+n}]\cdot\underline{\mathbf{a}_k}\rangle = \mathbf{W}_n. \tag{36}$$

The assumption of independence in (31) and (32) implies uncorrelation in (36) as $n \to \infty$. Conversely, uncorrelation implies independence, because the random vectors $\mathbf{a}_k$ are unit basis vectors (15),[10] rendering the covariance and probability matrices identical. Thus, rather than assuming independence in (31) and (32), we might equally well assume that the modulation vectors $\mathbf{a}_k$ become uncorrelated for widely separated time slots.

The spectral density of $\mathbf{a}_k$ is the discrete transform of (36);

$$\tilde{\mathbf{P}}_{\mathbf{a}}(f) = \sum_{n=-\infty}^{\infty} e^{-j2\pi fn}\mathbf{W}_n. \tag{37†}$$

Equations (36) and (37) are the matrix extensions for discrete vector random processes of similar scalar relations for discrete scalar random processes;[11] the symbol $\tilde{}$ indicates that a discrete random process is under consideration. We separate (37) into line and continuous spectral components:

$$\tilde{\mathbf{P}}_{\mathbf{a}}(f) = \tilde{\mathbf{P}}_{\mathbf{a}_l}(f) + \tilde{\mathbf{P}}_{\mathbf{a}_c}(f). \tag{38}$$

Then from (32) and (37):

$$\tilde{\mathbf{P}}_{\mathbf{a}_l}(f) = \mathbf{w}]\cdot\lfloor\mathbf{w}\rfloor \sum_{n=-\infty}^{\infty} \delta(f-n), \tag{39}$$

$$\tilde{\mathbf{P}}_{\mathbf{a}_c}(f) = \sum_{n=-\infty}^{\infty} e^{-j2\pi fn}\{\mathbf{W}_n - \mathbf{w}]\cdot\lfloor\mathbf{w}\rfloor\}. \tag{40}$$

The assumption of (31) and (32) that signaling waveforms in widely separated time slots are independent, or the equivalent assumption that $\mathbf{a}_{k+n}$ and $\mathbf{a}_k$ become uncorrelated for large n, eliminates line components except dc if we confine our attention to the fundamental frequency interval $|f| < \frac{1}{2}$, and so retain only the $n = 0$ term in (39).[‡] Consequently, the modulation has no periodic patterns.

IV. PSK AS A BASEBAND PULSE TRAIN

Our problem has been reduced to determining the spectral density of the complex wave $v(t)$:

$$v(t) \equiv e^{j\phi(t)}, \tag{41}$$

$$\phi(t) = \sum_{k=-\infty}^{\infty} \lfloor\mathbf{a}_k\rfloor\cdot\mathbf{g}(t-kT)], \tag{42}$$

where $\mathbf{a}$ and $\mathbf{g}$ are M-dimensional vectors. We show that if the signaling pulses are strictly time-limited to an interval KT, $v(t)$ may be written

$$v(t) = \sum_{k=-\infty}^{\infty} \lfloor\mathbf{b}_k\rfloor\cdot\mathbf{r}(t-kT)]. \tag{43}$$

[†] While this relation is defined for the entire range $-\infty < f < \infty$, $\tilde{\mathbf{P}}_{\mathbf{a}}(f)$ is periodic in f with unit period, and only the fundamental period $|f| < \frac{1}{2}$ is normally of interest. Thus, the inverse transform is

$$\tilde{\mathbf{\Phi}}_{\mathbf{a}}(n) = \mathbf{W}_n = \int_{-\frac{1}{2}}^{\frac{1}{2}} \tilde{\mathbf{P}}_{\mathbf{a}}(f)e^{+j2\pi fn}df.$$

[‡] We may thus write $\mathbf{a}_k = \mathbf{w} + \mathbf{a}_{ck}$ with (40) giving the spectral density of $\mathbf{a}_c$ and (39) the spectral density of $\mathbf{w}$.

Vectors $\mathbf{b}$ and $\mathbf{r}$ are M^K-dimensional vectors expressed in terms of $\mathbf{a}$ and $\mathbf{g}$, respectively. The different terms in (43) are strictly nonoverlapping:

$$\mathbf{r}(t)] = \mathbf{0}], \quad t \leqq 0, \quad t > T. \qquad (44)$$[†]

Each $\mathbf{b}_k$ is a unit basis vector, i.e., only one of its M^K components is unity, all others being zero. The spectral density of (43) is determined in Appendix B.

Figure 1 shows portions of $\phi(t)$ of (42) for four different maximum signal pulse durations; the terms $k = -1, 0, 1, 2$ of (42) are shown, and for convenience $\mathbf{a}_k$ has been taken the same for each of these time slots. The pulses are positioned along the time slots such that the limits of each signal pulse lie on the boundary between adjacent time slots (i.e., $t = \text{integer} \cdot T$); since symmetric pulses have been chosen for illustration in Fig. 1, their maxima are centered in the time slots for K odd, and lie on the time-slot boundaries for K even. Examine the $(0, T]$ time slot in Fig. 1 as typical; then with the above choice of pulse positions, the number of pulses contributing to $\phi(t)$ in each time slot equals K. Since each pulse can take on M different shapes, $\phi(t)$ can take on M^K different shapes in each time slot; the same is true for $v(t)$ of (41), thus demonstrating the representation of (43).

It remains for us to express the pulse shapes $\mathbf{r}(t)$ and coefficients $\mathbf{b}_k$ of (43) in terms of the signal pulses $\mathbf{g}(t)$ and coefficients $\mathbf{a}_k$ of (42). We give separate treatments for the cases $K = 1$ and $K = 2$, and extend these results to general K.

4.1 Nonoverlapping pulses: $K = 1$

The top portion of Fig. 1 shows digital phase modulation for which the signal pulses in different time slots never overlap; in this case,

$$\mathbf{g}(t)] = \mathbf{0}], \qquad t \leqq 0, \quad t > T, \qquad (45)$$

where $\mathbf{0}]$ represents the M-dimensional zero vector. Define

$$\underline{\mathbf{q}(t)} \equiv \begin{cases} [e^{jg_1(t)}\ e^{jg_2(t)}\ \cdots\ e^{jg_M(t)}], & 0 < t \leqq T. \\ \underline{\mathbf{0}}, & t \leqq 0, \quad t > T. \end{cases} \qquad (46)$$

Then (41) and (42) may be written

$$v(t) = \sum_{k=-\infty}^{\infty} \underline{\mathbf{a}_k} \cdot \mathbf{q}(t - kT)]. \qquad (47)$$

[†] $\mathbf{0}] = \underline{\mathbf{0}}'$ is a vector of appropriate dimension (here M^K) with all elements zero.

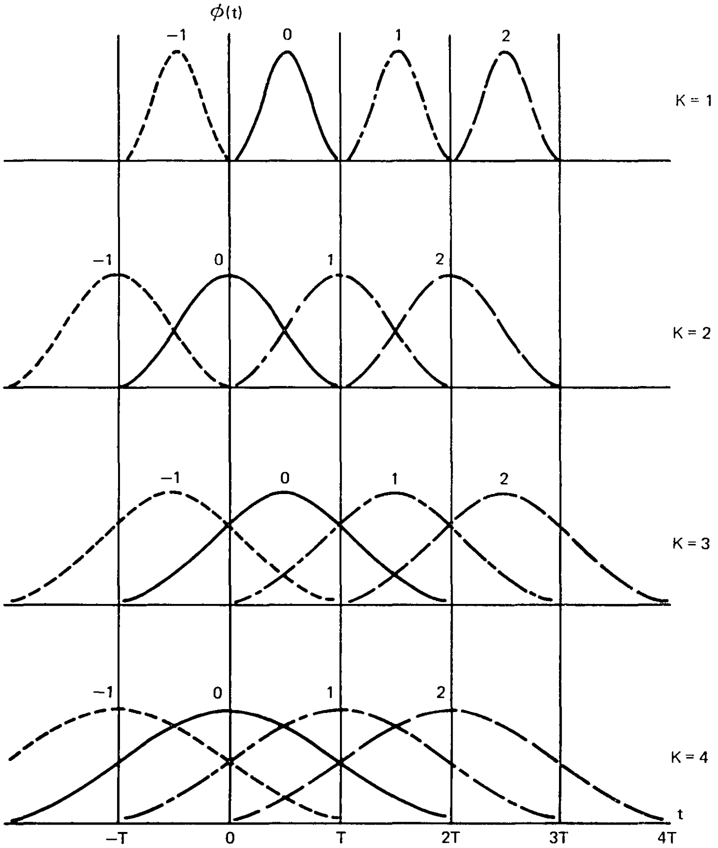


Fig. 1—Phase modulation for different signal pulse durations. Index k [eq. (42)] is shown near peak of each pulse. Also, for simplicity, same signal pulse is shown for each k. T = time slot duration or signaling period. KT = maximum signal pulse duration. Note different pulse center location for odd and even K.

Comparing (47) and (43), the parameters of the latter are given as follows for nonoverlapping signal pulses:

$$\lfloor \mathbf{b}_k \rfloor = \lfloor \mathbf{a}_k \rfloor, \quad \mathbf{r}(t)] = \mathbf{q}(t)]; \quad K = 1. \tag{48}$$

4.2 Overlapping pulses: $K = 2$

This case is illustrated in the second portion of Fig. 1. In the (0, T] time slot, the k = 0, 1 pulses contribute. We have

$$\mathbf{g}(t)] = 0], \qquad t \leqq -T, \quad t > T. \tag{49}$$

Define

$$\underline{\mathbf{q}(t)} \equiv \begin{cases} [e^{jg_1(t)}\ e^{jg_2(t)}\ \cdots\ e^{jg_M(t)}], & -T < t \leqq T. \\ \underline{0}, & t \leqq -T, \quad t > T. \end{cases} \qquad (50)^{\dagger}$$

Then

$$\begin{aligned} v(t) &= \sum_{k=-\infty}^{\infty} \{\underline{\mathbf{a}_k}\cdot \mathbf{q}(t-kT)]\}\{\underline{\mathbf{a}_{k+1}}\cdot \mathbf{q}(t-(k+1)T)]\} \\ &= \sum_{k=-\infty}^{\infty} \{\underline{\mathbf{a}_k}\cdot \mathbf{q}(t-kT)\ ]\} \times \{\underline{\mathbf{a}_{k+1}}\cdot \mathbf{q}(t-(k+1)T]\} \\ &= \sum_{k=-\infty}^{\infty} \{\underline{\mathbf{a}_k} \times \underline{\mathbf{a}_{k+1}}\}\cdot\{\mathbf{q}(t-kT)] \times \mathbf{q}(t-(k+1)T)]\}, \end{aligned} \qquad (51)$$

where × denotes the (right) Kronecker product[12,13] throughout.[‡] Comparing the last line of (51) with (43), the parameters of the latter are given as follows when no more than two signal pulses overlap:

$$\underline{\mathbf{b}_k} = \underline{\mathbf{a}_k} \times \underline{\mathbf{a}_{k+1}}, \quad \mathbf{r}(t)] = \mathbf{q}(t)] \times \mathbf{q}(t-T)]; \quad K = 2. \qquad (52)$$

Since the elements of $\mathbf{b}_k$ consist of all pairs of products of the elements of $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$, and since the $\mathbf{a}_k$ are M-dimensional unit-basis vectors

[†] Comparing (50) with (46), note that the definition of $\mathbf{q}(t)$ is different for different K; $\mathbf{q}(t) \not\equiv 0$ over the same interval in which $\mathbf{g}(t)$ may be nonzero.

[‡] The second line of (51) follows from the first by the observation that the two scalars may be regarded as 1-by-1 matrices; consequently, their scalar product and their Kronecker product are identical. The third line follows from the second by the well-known result connecting ordinary matrix and Kronecker products as follows: Consider two arbitrary matrices A and B with elements a_{ij}, b_{ij}. Define the (right) Kronecker product by (Refs. 12 and 13):

$$\mathbf{A} \times \mathbf{B} \equiv \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}$$

The following results may be found in Refs. 12 or 13.

For any matrices A, B, C, and D:

$$\mathbf{A} \times \mathbf{B} \times \mathbf{C} = (\mathbf{A} \times \mathbf{B}) \times \mathbf{C} = \mathbf{A} \times (\mathbf{B} \times \mathbf{C}), \qquad (\mathbf{A} \times \mathbf{B})' = \mathbf{A}' \times \mathbf{B}'.$$

For A and B the same size, and C and D the same size (possibly different from the size of A and B):

$$(\mathbf{A} + \mathbf{B}) \times (\mathbf{C} + \mathbf{D}) = \mathbf{A} \times \mathbf{C} + \mathbf{A} \times \mathbf{D} + \mathbf{B} \times \mathbf{C} + \mathbf{B} \times \mathbf{D}.$$

Assume the matrices $\mathbf{A}_1$, $\mathbf{B}_1$ and $\mathbf{A}_2$, $\mathbf{B}_2$ are dimensioned so that the ordinary matrix products $\mathbf{A}_1\cdot\mathbf{B}_1$ and $\mathbf{A}_2\cdot\mathbf{B}_2$ exist (i.e., there are λ columns of $\mathbf{A}_1$ and rows of $\mathbf{B}_1$, and μ columns of $\mathbf{A}_2$ and rows of $\mathbf{B}_2$). Then

$$(\mathbf{A}_1\cdot\mathbf{B}_1) \times (\mathbf{A}_2\cdot\mathbf{B}_2) = (\mathbf{A}_1 \times \mathbf{A}_2)\cdot(\mathbf{B}_1 \times \mathbf{B}_2);$$

this result generalizes to

$$\begin{aligned} &(\mathbf{A}_1\cdot\mathbf{B}_1\cdot\mathbf{C}_1) \times (\mathbf{A}_2\cdot\mathbf{B}_2\cdot\mathbf{C}_2) \times (\mathbf{A}_3\cdot\mathbf{B}_3\cdot\mathbf{C}_3) \\ &\qquad = (\mathbf{A}_1 \times \mathbf{A}_2 \times \mathbf{A}_3)\cdot(\mathbf{B}_1 \times \mathbf{B}_2 \times \mathbf{B}_3)\cdot(\mathbf{C}_1 \times \mathbf{C}_2 \times \mathbf{C}_3) \end{aligned}$$

etc.

by (14) and (15), the $\mathbf{b}_k$ are M^2-dimensional unit-basis vectors: that is, $\mathbf{b}_k$ has one element unity and the remainder $M^2 - 1$ elements zero. Note from (52) and (50) that

$$\mathbf{r}(t)] = \mathbf{0}], \quad t \leqq 0, \quad t > T. \tag{53}$$

We illustrate these relations for the binary case,[14] in which only two signal waveforms $g_1(t)$ and $g_2(t)$ are transmitted. Then

$$\mathbf{g}(t)] \equiv \begin{bmatrix} g_1(t) \\ g_2(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad t \leqq -T, \quad t > T. \tag{54}$$

$$\mathbf{q}(t)] = \begin{cases} \begin{bmatrix} e^{jg_1(t)} \\ e^{jg_2(t)} \end{bmatrix}, & -T < t \leqq T. \\ \begin{bmatrix} 0 \\ 0 \end{bmatrix} & t \leqq -T, \quad t > T. \end{cases} \tag{55}$$

$$\mathbf{r}(t)] = \begin{cases} \begin{bmatrix} e^{j[g_1(t)+g_1(t-T)]} \\ e^{j[g_1(t)+g_2(t-T)]} \\ e^{j[g_2(t)+g_1(t-T)]} \\ e^{j[g_2(t)+g_2(t-T)]} \end{bmatrix}, & 0 < t \leqq T. \\ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, & t \leqq 0, \quad t > T. \end{cases} \tag{56}$$

$$\begin{aligned} \underline{\mathbf{b}_k} &= [a_k^{(1)} \; a_k^{(2)}] \times [a_{k+1}^{(1)} \; a_{k+1}^{(2)}] \\ &= [a_k^{(1)} a_{k+1}^{(1)} \quad a_k^{(1)} a_{k+1}^{(2)} \quad a_k^{(2)} a_{k+1}^{(1)} \quad a_k^{(2)} a_{k+1}^{(2)}]. \end{aligned} \tag{57a}$$

	$\underline{\mathbf{a}_k}$ \ $\underline{\mathbf{a}_{k+1}}$	[1 0]	[0 1]
$\underline{\mathbf{b}_k}$:	[1 0]	[1 0 0 0]	[0 1 0 0]
	[0 1]	[0 0 1 0]	[0 0 0 1]

(57b)

The four possible waveshapes for $\phi(t)$ in the $(0, T]$ time slot are shown in Fig. 2.

4.3 Overlapping pulses: General K

The general case is treated by straightforward extension of the above method; Fig. 1 illustrates the phase modulation for $K = 3, 4$. The following expressions differ slightly for the even and odd cases; they

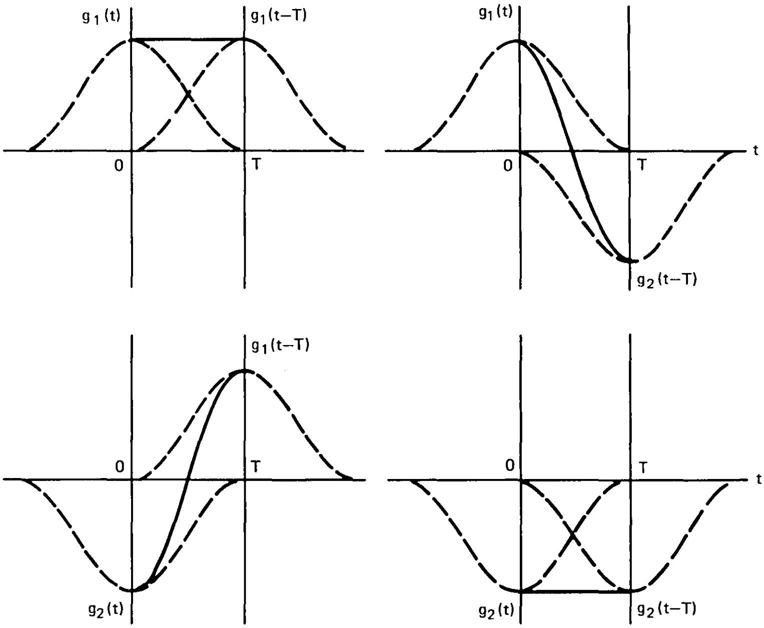


Fig. 2—$\phi(t)$ for $0 < t \leqq T$, binary signaling, and $K = 2$.

correctly reduce to the above results for $K = 1, 2$.

$$\mathbf{g}(t)] = \mathbf{0}], \quad \begin{array}{lll} t \leqq -\dfrac{K-1}{2}\,T, \quad t > \dfrac{K+1}{2}\,T; & & K \text{ odd.} \\ t \leqq -\dfrac{K}{2}\,T, \qquad t > \dfrac{K}{2}\,T; & & K \text{ even.} \end{array} \tag{58}$$

$$\underline{q(t)} \equiv \begin{cases} [e^{jg_1(t)}\; e^{jg_2(t)}\; \cdots\; e^{jg_M(t)}], & \\ \qquad -\dfrac{K-1}{2}\,T < t \leqq \dfrac{K+1}{2}\,T; & K \text{ odd.} \\ \qquad -\dfrac{K}{2}\,T < t \leqq \dfrac{K}{2}\,T; & K \text{ even.} \\ \underline{0}, & \\ \qquad t \leqq -\dfrac{K-1}{2}\,T, \quad t > \dfrac{K+1}{2}\,T; & K \text{ odd.} \\ \qquad t \leqq -\dfrac{K}{2}\,T, \quad t > \dfrac{K}{2}\,T; & K \text{ even.} \end{cases} \tag{59}$$

The parameters of (43) are:

$$\llcorner\mathbf{b}_k\lrcorner = \prod_{\times\; i=-(K-1)/2}^{(K-1)/2} \llcorner\mathbf{a}_{k+i}\lrcorner \qquad \mathbf{r}(t)] = \prod_{\times\; i=-(K-1)/2}^{(K-1)/2} \mathbf{q}(t-iT)]; \quad K \text{ odd.} \tag*{(60)†}$$

$$\llcorner\mathbf{b}_k\lrcorner = \prod_{\times\; i=-(K/2)+1}^{K/2} \llcorner\mathbf{a}_{k+1}\lrcorner \qquad \mathbf{r}(t)] = \prod_{\times\; i=-(K/2)+1}^{K/2} \mathbf{q}(t-iT)]; \quad K \text{ even.} \tag*{(61)†}$$

Note that

$$\mathbf{r}(t)] = \mathbf{0}], \qquad t \leqq 0, \quad t > T. \tag{62}$$

V. PSK WITH NONOVERLAPPING, CORRELATED SIGNAL PULSES: $K = 1$

Assume the signal pulses are confined to one time slot, as in (45):

$$\mathbf{g}(t)] = \mathbf{0}], \quad t \leqq 0, \quad t > T. \tag{63}$$

From (46) to (48) and (160) to (161),

$$\mathbf{R}(f)] = \int_0^T e^{-j2\pi ft} \begin{bmatrix} e^{jg_1(t)} \\ e^{jg_2(t)} \\ \vdots \\ e^{jg_M(t)} \end{bmatrix} dt. \tag{64}$$

We separate $v(t)$ of (2) and (3) or (41) and (42) into line and continuous components:

$$v(t) \equiv e^{j\phi(t)} = v_l(t) + v_c(t). \tag{65}$$

Using (48) and comparing (32) to (37) with (149) to (153), we have from (171)

$$v_l(t) = \frac{1}{T} \llcorner\mathbf{w}\lrcorner \cdot \sum_{n=-\infty}^{\infty} \mathbf{R}\left(\frac{n}{T}\right)\Bigg] e^{jn2\pi t/T}. \tag{66}$$

From (165),

$$\mathbf{P}_{vl}(f) = \frac{1}{T^2} |\llcorner\mathbf{w}\lrcorner \cdot \mathbf{R}(f)]|^2 \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right), \tag{67}$$

[†] The symbol $\Pi_\times$ used here and subsequently indicates a multiple Kronecker product:

$$\prod_{\times\; i=L}^{N} \mathbf{A}_i \equiv \mathbf{A}_L \times \mathbf{A}_{L+1} \times \cdots \times \mathbf{A}_{N-1} \times \mathbf{A}_N.$$

The products in (60) and (61) contain K factors.

which may be written

$$\mathbf{P}_{vl}(f) = \frac{1}{T^2} \left| w_1 R_1(f) + w_2 R_2(f) + \cdots + w_M R_M(f) \right|^2 \cdot \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right). \tag{68}$$

From (162) and (164)

$$\mathbf{P}_{vc}(f) = \frac{1}{T} \underline{\mathbf{R}(f)} \cdot \sum_{n=-\infty}^{\infty} e^{-j2\pi f T n} \left\{ \mathbf{W}_n - \mathbf{w}] \cdot \underline{\mathbf{w}} \right\} \cdot \mathbf{R}^*(f)]. \tag{69}$$

We illustrate these general results with two examples.

5.1 Independent, M-ary signal pulses

Equation (32) now holds for all $n \neq 0$; using (21),

$$\mathbf{W}_n = \mathbf{w}] \cdot \underline{\mathbf{w}} \equiv \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_M \end{bmatrix} \cdot [w_1 \, w_2 \cdots w_M], \quad n \neq 0. \tag{70}$$

From (27),

$$\mathbf{W}_0 = \mathbf{w}_d \equiv \begin{bmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_M \end{bmatrix}. \tag{71}$$

Then only the $n = 0$ term remains in (69);

$$\mathbf{P}_{vc}(f) = \frac{1}{T} \underline{\mathbf{R}(f)} \cdot \mathbf{w}_d \cdot \mathbf{R}^*(f)] - \frac{1}{T} \left| \underline{\mathbf{w}} \cdot \mathbf{R}(f)] \right|^2. \tag{72}$$

Writing out the matrix operations, (72) yields

$$\begin{aligned} \mathbf{P}_{vc}(f) &= \frac{1}{T} [w_1 |R_1(f)|^2 + w_2 |R_2(f)|^2 + \cdots + w_M |R_M(f)|^2] \\ &\qquad - \frac{1}{T} |w_1 R_1(f) + w_2 R_2(f) + \cdots + w_M R_M(f)|^2 \\ &= \frac{1}{2T} \sum_{i=1}^{M} \sum_{j=1}^{M} w_i \, w_j |R_i(f) - R_j(f)|^2. \end{aligned} \tag{73}$$

This result has been given in Ref. 6. The line component is given by (66), its spectrum by (67) and (68).

5.2 Correlated, binary signal pulses: M = 2

The matrix $\mathbf{W}_n$ consists of the four upper left-hand elements of (23). The constraints of (19) and (20) or (25) and (26) render three of these functions dependent on the fourth.

$$W_n = [W_n(1,1) - w_1^2]\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \mathbf{w}]\cdot\underline{\mathbf{w}}. \tag{74}$$

From (27)

$$W_0(1,1) = w_1, \tag{75}$$

from (30)

$$W_n(1,1) = W_{-n}(1,1), \tag{76}$$

and from (31)

$$\lim_{n\to\infty} W_n(1,1) = w_1^2. \tag{77}$$

Equation (69) becomes

$$\mathbf{P}_{v_c}(f) = \frac{1}{T}\,|R_1(f) - R_2(f)|^2 \sum_{n=-\infty}^{\infty} e^{-j2\pi fTn}[W_n(1,1) - w_1^2]. \tag{78}$$

$W_n(1,1)$ may be any discrete covariance function satisfying the constraints of (75) to (77). The line component and its spectrum are again given by (66), (67), and (68) with two-component vectors, e.g.,

$$\mathbf{P}_{v_l}(f) = \frac{1}{T^2}\,|w_1R_1(f) + w_2R_2(f)|^2 \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right). \tag{79}$$

The binary independent case is obtained by setting

$$W_n(1,1) = w_1^2, \quad n \neq 0. \tag{80}$$

Then (78) becomes

$$\mathbf{P}_{v_c}(f) = \frac{w_1w_2}{T}\,|R_1(f) - R_2(f)|^2. \tag{81}$$

This agrees with (73) for the binary case.

VI. PSK WITH INDEPENDENT, OVERLAPPING SIGNAL PULSES: *K* = 2

Assume that no more than two signal pulses overlap, as in (49):

$$\mathbf{g}(t)] = 0], \quad t \leqq -T, \quad t > T. \tag{82}$$

$\mathbf{R}(f)$ is determined from (50) to (53) and (160) to (161):

$$\mathbf{R}(f)] = \int_0^T e^{-j2\pi ft} \begin{bmatrix} e^{j\{\theta_1(t)+\theta_1(t-T)\}} \\ \vdots \\ e^{j\{\theta_1(t)+\theta_M(t-T)\}} \\ e^{j\{\theta_2(t)+\theta_1(t-T)\}} \\ \vdots \\ e^{j\{\theta_2(t)+\theta_M(t-T)\}} \\ e^{j\{\theta_3(t)+\theta_1(t-T)\}} \\ \vdots \\ e^{j\{\theta_{M-1}(t)+\theta_M(t-T)\}} \\ e^{j\{\theta_M(t)+\theta_1(t-T)\}} \\ \vdots \\ e^{j\{\theta_M(t)+\theta_M(t-T)\}} \end{bmatrix} dt. \tag{83}$$

From (52),

$$\underline{\mathbf{b}_k} = \underline{\mathbf{a}_k} \times \underline{\mathbf{a}_{k+1}}. \tag{84}$$

We determine the mean and covariance of $\mathbf{b}_k$ from (32) to (37) and (149) to (153). We assume throughout this section that signal pulses in different time slots are independent, as in (70) and (71):

$$\mathbf{W}_0 = \mathbf{w}_d; \quad \mathbf{W}_n = \mathbf{w}]\cdot\underline{\mathbf{w}}, \quad |n| \neq 0. \tag{85}$$

For the mean, since $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$ are independent,

$$\begin{aligned} \underline{\boldsymbol{\beta}} &\equiv \langle\underline{\mathbf{b}_k}\rangle = \underline{\mathbf{w}} \times \underline{\mathbf{w}} \\ &\equiv \underline{\mathbf{w}}^{[2]}. \end{aligned} \tag{86}$$

We introduce the notation that an integer exponent enclosed in square brackets denotes the Kronecker power,[12] i.e., the Kronecker product of the matrix (or vector) with itself the indicated number of times.

For the covariance (see footnote, p. 908):

$$\begin{aligned} \tilde{\boldsymbol{\Phi}}_{\mathrm{b}}(n) &\equiv \langle \mathbf{b}_{k+n}]\cdot\underline{\mathbf{b}_k}\rangle \\ &= \langle(\mathbf{a}_{k+n}] \times \mathbf{a}_{k+n+1}])\cdot(\underline{\mathbf{a}_k} \times \underline{\mathbf{a}_{k+1}})\rangle \\ &= \langle(\mathbf{a}_{k+n}]\cdot\underline{\mathbf{a}_k}) \times (\mathbf{a}_{k+n+1}]\cdot\underline{\mathbf{a}_{k+1}})\rangle. \end{aligned} \tag{87}$$

Since

$$\tilde{\boldsymbol{\Phi}}_{\mathrm{b}}(n) = \tilde{\boldsymbol{\Phi}}_{\mathrm{b}}'(-n), \tag{88}$$

we study only $n \geqq 0$. We treat three cases.

First assume $|n| \geqq 2$. All the a's in (87) are independent; from the second line,

$$\begin{aligned} \tilde{\boldsymbol{\Phi}}_{\mathrm{b}}(n) &= (\mathbf{w}] \times \mathbf{w}])\cdot(\underline{\mathbf{w}} \times \underline{\mathbf{w}}) \\ &= \mathbf{w}]^{[2]}\cdot\underline{\mathbf{w}}^{[2]}, \quad |n| \geqq 2. \end{aligned} \tag{89}$$

Next consider $n = 0$. From the third line of (87) and the fact that $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$ are independent,

$$\tilde{\mathbf{\Phi}}_{\mathrm{b}}(0) = \mathbf{w}_d^{[2]}. \tag{90}$$

Finally, for $|n| = 1$, from the third line of (87) we have[†] (see footnote, p. 908):

$$\begin{aligned} \tilde{\mathbf{\Phi}}_{\mathrm{b}}(1) &= \langle(\mathbf{a}_{k+1}] \times \underline{\mathbf{a}_k}) \times (\mathbf{a}_{k+2}] \times \underline{\mathbf{a}_{k+1}})\rangle \\ &= \langle\underline{\mathbf{a}_k} \times (\mathbf{a}_{k+1}] \times \underline{\mathbf{a}_{k+1}}) \times \mathbf{a}_{k+2}]\rangle. \end{aligned} \tag{91}$$

Since $\mathbf{a}_k$, $\mathbf{a}_{k+1}$, $\mathbf{a}_{k+2}$ are independent,

$$\tilde{\mathbf{\Phi}}_{\mathrm{b}}(1) = \underline{\mathbf{w}} \times \mathbf{w}_d \times \mathbf{w}]. \tag{92}$$

From (88),

$$\tilde{\mathbf{\Phi}}_{\mathrm{b}}(-1) = \mathbf{w}] \times \mathbf{w}_d \times \underline{\mathbf{w}}. \tag{93}$$

From (86) to (89), (153) is satisfied, i.e., $\mathbf{b}_k$ for widely separated time slots are uncorrelated.[‡] Therefore, from (86), (165), and (171):

$$v_l(t) = \frac{1}{T}\underline{\mathbf{w}}^{[2]} \cdot \sum_{n=-\infty}^{\infty} \mathbf{R}\left(\frac{n}{T}\right)\Big] e^{jn2\pi t/T}; \tag{94}$$

$$\mathbf{P}_{v_l}(f) = \frac{1}{T^2}\,|\underline{\mathbf{w}}^{[2]} \cdot \mathbf{R}(f)]|^2 \sum_{n=-\infty}^{\infty} \delta(f - nT). \tag{95}$$

In comparing these results with (66) and (67), note that $\mathbf{R}(f)$ of (83) differs from $\mathbf{R}(f)$ of (64).

Substituting (86), (89), (90), (92), and (93) into (162) and (164), the continuous spectrum is

$$\begin{aligned} \mathbf{P}_{v_c}(f) = \frac{1}{T}\underline{\mathbf{R}(f)} \cdot \{&(\mathbf{w}_d^{[2]} - \mathbf{w}]^{[2]} \cdot \underline{\mathbf{w}}^{[2]}) \\ &+ (\underline{\mathbf{w}} \times \mathbf{w}_d \times \mathbf{w}] - \mathbf{w}]^{[2]} \cdot \underline{\mathbf{w}}^{[2]})e^{-j2\pi fT} \\ &+ (\mathbf{w}] \times \mathbf{w}_d \times \underline{\mathbf{w}} - \mathbf{w}]^{[2]} \cdot \underline{\mathbf{w}}^{[2]})e^{+j2\pi fT}\} \cdot \mathbf{R}^*(f)]. \end{aligned} \tag{96}$$

We illustrate these results for the binary case.

[†] The following vector relations may be established by inspection:

$$\mathbf{a}] \cdot \underline{\mathbf{b}} = \mathbf{a}] \times \underline{\mathbf{b}} = \underline{\mathbf{b}} \times \mathbf{a}].$$

[‡] $\mathbf{b}_{k+n}$ and $\mathbf{b}_k$ also become independent as $n \to \infty$, because the $\mathbf{b}_k$ are unit basis vectors (Ref. 10).

6.1 Independent, binary signal pulses: M = 1

R(f) in (83) now contains four components, the Fourier transforms of the components of (56). We have

$$\underline{\mathbf{w}} = [w_1 \;\; w_2], \quad \mathbf{w}_d = \begin{bmatrix} w_1 & 0 \\ 0 & w_2 \end{bmatrix} \tag{97}$$

$$\underline{\mathbf{w}}^{[2]} = [w_1^2 \;\; w_1w_2 \;\; w_1w_2 \;\; w_2^2] \tag{98}$$

$$\mathbf{w}_d^{[2]} = \begin{bmatrix} w_1^2 & 0 & 0 & 0 \\ 0 & w_1w_2 & 0 & 0 \\ 0 & 0 & w_1w_2 & 0 \\ 0 & 0 & 0 & w_2^2 \end{bmatrix} \tag{99}$$

$$\mathbf{w}]^{[2]} \cdot \underline{\mathbf{w}}^{[2]} = \begin{bmatrix} w_1^4 & w_1^3w_2 & w_1^3w_2 & w_1^2w_2^2 \\ w_1^3w_2 & w_1^2w_2^2 & w_1^2w_2^2 & w_1w_2^3 \\ w_1^3w_2 & w_1^2w_2^2 & w_1^2w_2^2 & w_1w_2^3 \\ w_1^2w_2^2 & w_1w_2^3 & w_1w_2^3 & w_2^4 \end{bmatrix} \tag{100}$$

$$\underline{\mathbf{w}} \times \mathbf{w}_d \times \mathbf{w}] = \begin{bmatrix} w_1^3 & 0 & w_1^2w_2 & 0 \\ w_1^2w_2 & 0 & w_1w_2^2 & 0 \\ 0 & w_1^2w_2 & 0 & w_1w_2^2 \\ 0 & w_1w_2^2 & 0 & w_2^3 \end{bmatrix} \tag{101}$$

$$\mathbf{w}] \times \mathbf{w}_d \times \underline{\mathbf{w}} = (\underline{\mathbf{w}} \times w_d \times \mathbf{w}])'. \tag{102}^{\dagger}$$

Equations (98) to (102) substituted in (94) to (96) yield the final results for independent, binary signal pulses.

To illustrate, we write out a few terms of the matrix contained in { } in (96), for the continuous spectrum in the binary case:[15]

$$\begin{aligned} \{\;\;\}_{11} &= w_1^2(1 - w_1^2)(1 + 2w_1 \cos 2\pi fT), \\ \{\;\;\}_{12} &= w_1^2w_2e^{+j2\pi fT} - w_1^3w_2(1 + 2\cos 2\pi fT), \\ &\vdots \\ \{\;\;\}_{44} &= \cdots. \end{aligned} \tag{103}$$

Writing out even the present simplest overlapping case produces an awful mess. Consequently, in the examples presented in this paper the digital computer is programmed to work directly with (96) or its generalization (116), performing matrix operations (both ordinary and Kronecker matrix multiplication) directly, rather than entering expressions such as (103). In this way, quite complicated cases involving multilevel signal pulses overlapping several time slots may be simply treated.

[†] See footnote, p. 908.

The general results for $K = 2$, (94) to (96), require the signal pulses to be less than two time slots in duration (82). Therefore, they must apply when the signal pulses are restricted to a single time slot, i.e., when the stronger condition (63) is satisfied, and must then reduce to the results for $K = 1$, (66), (67), and (72). This reduction is demonstrated in Appendix C.

VII. PSK WITH INDEPENDENT OVERLAPPING SIGNAL PULSES: GENERAL K

Finally, assume that the signal pulses occupy at most K time slots, so no more than K signal pulses overlap; $\mathbf{g}(t)$ is now given by (58). $\mathbf{r}(t)$ is given by (59) to (62), and $\mathbf{R}(f)$ by (160) and (161).

The $\mathbf{b}_k$ are given by (60) or (61); the mean and covariance of $\mathbf{b}_k$ are found from (32) to (37) and (149) to (153), assuming throughout that signal pulses in different time slots are independent, as in (70) to (71).

For the mean,

$$\underbracket{\boldsymbol{\beta}} = \langle \underbracket{\mathbf{b}_k} \rangle = \underset{i}{\textstyle\prod_{\times}} \langle \underbracket{\mathbf{a}_{k+i}} \rangle \tag{104}$$

by the independence of the $\mathbf{a}_k$. The limits over the $\prod_{\times}$, given in (60) or (61) for K odd or even respectively, extended over K factors in either case. Thus,

$$\underbracket{\boldsymbol{\beta}} = \underbracket{\mathbf{w}}^{[K]}. \tag{105}$$

For the covariance (see footnote, p. 908):

$$\begin{aligned} \tilde{\boldsymbol{\Phi}}_{\mathbf{b}}(n) &\equiv \langle \mathbf{b}_{k+n}\rbrack \cdot \underbracket{\mathbf{b}_k} \rangle \\ &= \langle (\underset{i}{\textstyle\prod_{\times}} \mathbf{a}_{k+n+i}\rbrack) \cdot (\underset{i}{\textstyle\prod_{\times}} \underbracket{\mathbf{a}_{k+i}}) \rangle \\ &= \langle \underset{i}{\textstyle\prod_{\times}} (\mathbf{a}_{k+n+i}\rbrack \cdot \underbracket{\mathbf{a}_{k+i}}) \rangle, \end{aligned} \tag{106}$$

the $\prod_{\times}$ being taken over K factors as given by either (60) or (61). By stationarity, (106) is independent of k; consequently, for both even and odd cases

$$\begin{aligned} \tilde{\boldsymbol{\Phi}}_{\mathbf{b}}(n) &\equiv \langle \mathbf{b}_{k+n}\rbrack \cdot \underbracket{\mathbf{b}_k} \rangle \\ &= \left\langle \left(\prod_{i=1}^{K}{}_{\times}\, \mathbf{a}_{n+i}\rbrack \right) \cdot \left(\prod_{i=1}^{K}{}_{\times}\, \underbracket{\mathbf{a}_i} \right) \right\rangle \\ &= \left\langle \prod_{i=1}^{K}{}_{\times}\, (\mathbf{a}_{n+1}\rbrack \cdot \underbracket{\mathbf{a}_i}) \right\rangle \cdot \end{aligned} \tag{107}$$

First, for $n = 0$, from the third line of (107) and the independence of the $\mathbf{a}_k$,

$$\tilde{\mathbf{\Phi}}_{\mathrm{b}}(0) = \prod_{i=1}^{K}{}_{\times} \langle \mathbf{a}_i] \cdot \underbracket{\mathbf{a}_i} \rangle = \mathbf{w}_d^{[K]}. \tag{108}$$

Next consider $n = 1$. From the third line of (107) (see footnote, p. 915),

$$\begin{aligned}\tilde{\mathbf{\Phi}}_{\mathrm{b}}(1) &= \left\langle \prod_{i=1}^{K}{}_{\times} (\underbracket{\mathbf{a}_i} \times \mathbf{a}_{i+1}]) \right\rangle \\ &= \langle \underbracket{\mathbf{a}_1} \rangle \times \left\{ \prod_{i=1}^{K}{}_{\times} \langle \mathbf{a}_i] \cdot \underbracket{\mathbf{a}_i} \rangle \right\} \times \langle \mathbf{a}_{K+1}] \rangle \\ &= \underbracket{\mathbf{w}} \times \mathbf{w}_d^{[K-1]} \times \mathbf{w}], \end{aligned} \tag{109}$$

the second line again following from the independence of the $\mathbf{a}_k$.

For $n = 2$, proceeding as above:

$$\begin{aligned}\tilde{\mathbf{\Phi}}_{\mathrm{b}}(2) &= \left\langle \prod_{i=1}^{K}{}_{\times} (\underbracket{\mathbf{a}_i} \times \mathbf{a}_{i+2}]) \right\rangle \\ &= \langle \underbracket{\mathbf{a}_1} \rangle \times \left\langle \prod_{i=2}^{K}{}_{\times} (\mathbf{a}_{i+1}] \times \underbracket{\mathbf{a}_i}) \right\rangle \times \langle \mathbf{a}_{K+2}] \rangle \\ &= \langle \underbracket{\mathbf{a}_1} \rangle \times \left\langle \prod_{i=2}^{K}{}_{\times} (\underbracket{\mathbf{a}_i} \times \mathbf{a}_{i+1}]) \right\rangle \times \langle \mathbf{a}_{K+2}] \rangle \\ &= \langle \underbracket{\mathbf{a}_1} \rangle \times \langle \underbracket{\mathbf{a}_2} \rangle \times \left\{ \prod_{i=3}^{K}{}_{\times} \langle \mathbf{a}_i] \cdot \underbracket{\mathbf{a}_i} \rangle \right\} \times \langle \mathbf{a}_{K+1}] \rangle \times \langle \mathbf{a}_{K+2}] \rangle \\ &= \underbracket{\mathbf{w}}^{[2]} \times \mathbf{w}_d^{[K-2]} \times \mathbf{w}]^{[2]}. \end{aligned} \tag{110}$$

By induction:

$$\begin{aligned}\tilde{\mathbf{\Phi}}_{\mathrm{b}}(3) &= \underbracket{\mathbf{w}}^{[3]} \times \mathbf{w}_d^{[K-3]} \times \mathbf{w}]^{[3]} \\ &\;\;\vdots \\ \tilde{\mathbf{\Phi}}_{\mathrm{b}}(K-1) &= \underbracket{\mathbf{w}}^{[K-1]} \times \mathbf{w}_d \times \mathbf{w}]^{[K-1]} \\ \tilde{\mathbf{\Phi}}_{\mathrm{b}}(K) &= \underbracket{\mathbf{w}}^{[K]} \times \mathbf{w}]^{[K]}. \end{aligned} \tag{111}$$

Next, from the second line of (107) and the independence of the $\mathbf{a}_K$,

$$\tilde{\mathbf{\Phi}}_{\mathrm{b}}(n) = \mathbf{w}]^{[K]} \cdot \underbracket{\mathbf{w}}^{[K]}, \quad n \geqq K. \tag{112}$$

This result is of course consistent with the final line of (111) (see footnote, p. 915). Finally,

$$\tilde{\mathbf{\Phi}}_{\mathrm{b}}(n) = \tilde{\mathbf{\Phi}}'_{\mathrm{b}}(-n). \tag{113}$$

From these results the line component is:

$$v_l(t) = \frac{1}{T}\underbracket{\mathbf{w}}^{[K]} \cdot \sum_{n=-\infty}^{\infty} \mathbf{R}\left(\frac{n}{T}\right)\Bigg] e^{jn2\pi t/T}; \tag{114}$$

$$\mathbf{P}_{v_l}(f) = \frac{1}{T^2}\left|\underbracket{\mathbf{w}}^{[K]} \cdot \mathbf{R}(f)]\right|^2 \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right). \tag{115}$$

The continuous spectrum is:

$$\begin{aligned}\mathbf{P}_{v_c}(f) = \frac{1}{T}\underbracket{\mathbf{R}(f)} \cdot \{(&\mathbf{w}_d^{[K]} - \mathbf{w}]^{[K]} \cdot \underbracket{\mathbf{w}}^{[K]}) \\ &+ \sum_{n=1}^{K-1} (\underbracket{\mathbf{w}}^{[n]} \times \mathbf{w}_d^{[K-n]} \times \mathbf{w}]^{[n]} - \mathbf{w}]^{[K]} \cdot \underbracket{\mathbf{w}}^{[K]}) e^{-jn2\pi fT} \\ &+ \sum_{n=1}^{K-1} (\mathbf{w}^{[n]} \times \mathbf{w}_d^{[K-n]} \times \underbracket{\mathbf{w}}^{[n]} - \mathbf{w}]^{[K]} \cdot \underbracket{\mathbf{w}}^{[K]}) e^{+jn2\pi fT}\} \\ &\cdot \mathbf{R}^*(f)].\end{aligned} \tag{116}$$

If we set $K = 2$ in (114) to (116), these results agree with those of Section VI. If the signal pulses are restricted to a smaller number of time slots $\tilde{K} < K$, the present results reduce correctly to those appropriate for $\tilde{K}$; this reduction is shown for $\tilde{K} = 1$, $K = 2$ in Appendix C, but the general case is left as an exercise for the reader.

VIII. EXAMPLES

We now consider a number of examples of computing spectra of multiphase PSK systems, subject to the following assumptions:

(*i*) The number of phase levels is a power of 2,

$$M = 2^N, \quad N \text{ an integer.} \tag{117}$$

(*ii*) The M signaling pulses have a common shape, and the M phase levels (peak phase modulations) are equally spaced.

$$\underbracket{\mathbf{g}(t)} = \frac{\pi}{M}[1 \;\; 3 \cdots (2M-1)]g(t), \tag{118}$$

$$g(t)_{\max} = 1. \tag{119}$$

Figure 3 shows vector diagrams of the peak phase modulations for $M = 2$, 4, and 8. We also assume that $g(t)$ is symmetric,

$$\begin{aligned} g(t) &= g(-t), \quad K \text{ even}, \\ g(t + T/2) &= g(-t + T/2), \quad K \text{ odd}, \end{aligned} \tag{120}$$

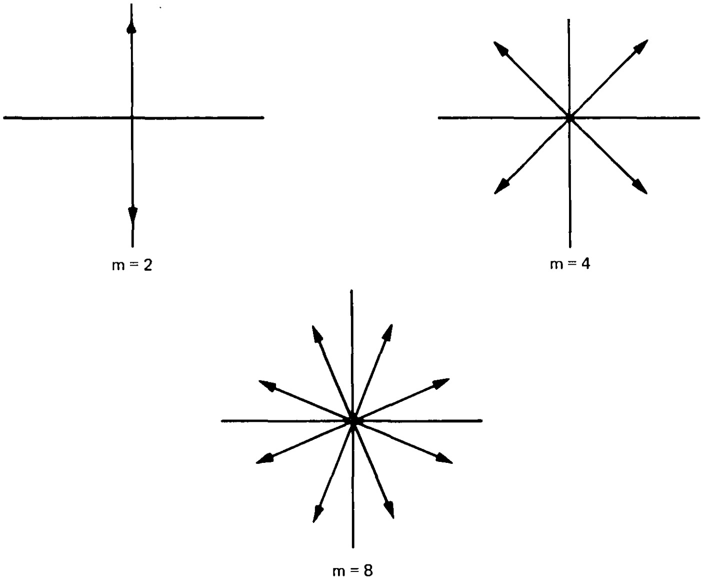


Fig. 3—Peak phase modulation for PSK with equally spaced levels.

with maximum at

$$\begin{aligned} t &= 0, \quad K \text{ even}, \\ t &= \frac{T}{2}, \quad K \text{ odd}. \end{aligned} \tag{121}$$

(*iii*) Signal pulses in different time slots are statistically independent, and all signal pulses are equally likely;

$$\mathrm{w}] = \frac{1}{M}\mathbf{1}], \qquad \mathrm{w}_d = \frac{1}{M}\mathbf{I}, \tag{122}$$

where **I** is the identity matrix of order M.

As a consequence of (118) and (122), we can show that $\mathbf{P}_v(f)$ is symmetric, that is

$$\mathbf{P}_v(f) = \mathbf{P}_v(-f). \tag{123}$$

8.1 Rectangular nonoverlapping signal pulses

If $g(t)$ is a rectangular pulse (see Fig. 4) of duration $\eta \leqq T$,

$$g(t) = \begin{matrix} 1, & \dfrac{T-\eta}{2} < t \leqq \dfrac{T+\eta}{2}, \quad 0 < \eta \leqq T, \\ 0, & \text{otherwise}. \end{matrix} \tag{124}$$

From (118) and (64),

$$R_i(f) = \left\{ \exp\{ j \frac{\pi}{M} (2i - 1) \} - 1 \right\} \frac{\sin \pi f \eta}{\pi f} e^{-j\pi f T}$$
$$+ \frac{\sin \pi f T}{\pi f} e^{-j\pi f T}, \quad i = 1, 2, \cdots, M. \quad (125)$$

The spectral density $\mathbf{P}_v(f)$ can be computed from (68), (73), and (122) with $\mathbf{R}_i(f)$ given by (125). The results can be shown to agree with those given in Refs. 4 and 6. Since this case can be treated as amplitude modulation and has been discussed extensively in Refs. 4 and 6, we shall not discuss it further in this paper.

8.2 Raised-cosine nonoverlapping signal pulses: $K = 1$

If $g(t)$ is a raised-cosine pulse (see Fig. 5) of duration T, the pulse just fills the time slot, and

$$g(t) = \begin{cases} \frac{1}{2}\left[1 - \cos\frac{2\pi t}{T}\right], & 0 < t \leqq T \\ 0, & \text{otherwise.} \end{cases} \quad (126)$$

In this case, $\mathbf{R}(f)]$ may be evaluated either numerically or using Bessel function expansion (which again requires the use of a computer).

For $M = 2$, 4, 8, and 16, we have evaluated as above

$$\mathbf{P}_v(f) = \mathbf{P}_{vl}(f) + \mathbf{P}_{v_c}(f), \quad (127)$$

with results shown in Fig. 6. We note that there is very little variation in either the discrete or the continuous spectrum when M is increased from 2 to 16. The tails of the spectrum are not shown in Fig. 6, although they are easily calculated down to the −60-dB level.

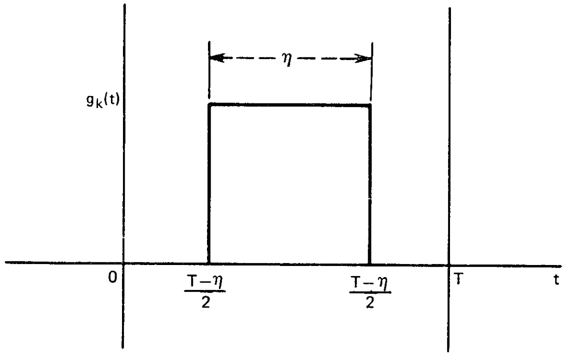


Fig. 4—Square-wave signaling of duration $\eta \leqq T$. $K = 1$.

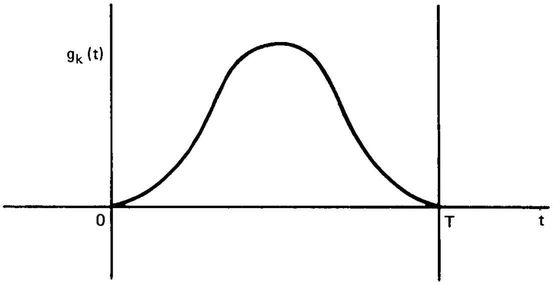


Fig. 5—Raised-cosine signaling with pulse duration T. $K = 1$.

8.3 Raised-cosine overlapping signal pulses: $K = 2$

If a raised-cosine signal pulse (see Fig. 7) just fills up two time slots,

$$g(t) = \begin{cases} \frac{1}{2}\left[1 + \cos\frac{\pi t}{T}\right], & -T < t \leqq T, \\ 0, & \text{otherwise,} \end{cases} \tag{128}$$

$K = 2$, and two signal pulses overlap.

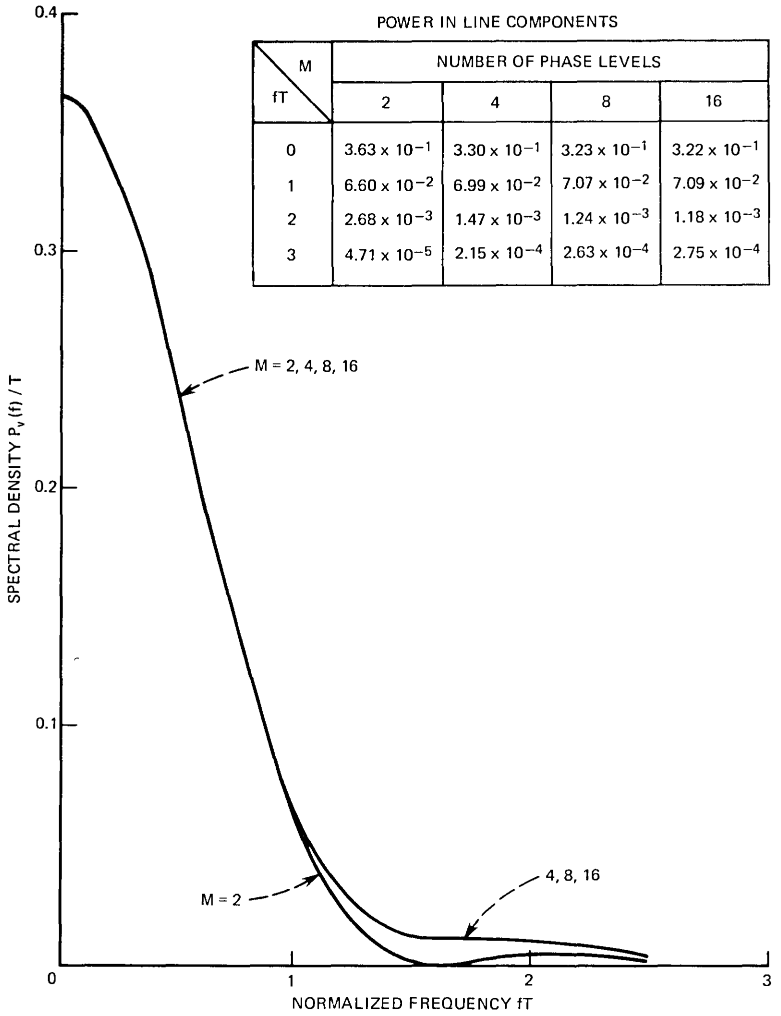
In this case, $\mathbf{P}_{vl}(f)$ and $\mathbf{P}_{vc}(f)$ are given by (95), (96), (122). Using a digital computer, $\mathbf{P}_v(f)$ has been determined for $M = 2$, 4, 8, and 16, and the results plotted in Fig. 8. Again, it may be noted that there is not very much variation in either the discrete or the continuous spectra when the number of phase levels M is increased from 2 to 16.

Comparing Figs. 6 and 8, we observe that the power in the line components in a PSK system with overlapping signal pulses is smaller than the power in the lines with nonoverlapping pulses. Also, for the same signaling rates, the principal portion of the continuous part of the spectrum with $K = 2$ is narrower than that with $K = 1$. The tails of $\mathbf{P}_v(f)$ were calculated down to -60 dB even though they are not shown in Fig. 8.

In Fig. 9, for $M = 4$, and raised-cosine wave signaling, we plot the spectral density $\mathbf{P}_v(f)$ when the amount of overlap is increased from zero to one time slot. Specifically, we assume that

$$g(t) = \begin{cases} \frac{1}{2}\left[1 + \cos\frac{\pi t}{\beta T}\right], & -\beta T < t \leqq \beta T, \quad 0.5 \leqq \beta \leqq 1, \\ 0, & \text{otherwise.} \end{cases} \tag{129}$$

Note that $K = 1$ when $\beta = 0.5$, and $K = 2$ when $0.5 < \beta \leqq 1$. Figure 9 shows that there is a gradual reduction of power in the line



POWER IN LINE COMPONENTS

fT \ M	NUMBER OF PHASE LEVELS			
	2	4	8	16
0	3.63×10^{-1}	3.30×10^{-1}	3.23×10^{-1}	3.22×10^{-1}
1	6.60×10^{-2}	6.99×10^{-2}	7.07×10^{-2}	7.09×10^{-2}
2	2.68×10^{-3}	1.47×10^{-3}	1.24×10^{-3}	1.18×10^{-3}
3	4.71×10^{-5}	2.15×10^{-4}	2.63×10^{-4}	2.75×10^{-4}

Fig. 6—Spectral density of binary, quaternary, octonary, and 16-ary PSK systems with raised-cosine signaling and pulse duration T. $K = 1$. Although the values of the spectral density for $M = 2$, 4, 8, and 16 are shown to be the same over certain segments of the range of f, they are usually different from each other, but this difference is not large enough to be shown on the linear scale in Fig. 6.

components when the overlap is increased from zero to one time slot. There is also the narrowing of the principal portion of the spectrum.

Decrease of the carrier component [$n = 0$ term in (68) or (95)] when β is increased may result from the fact that the phasors in Fig. 3 spend

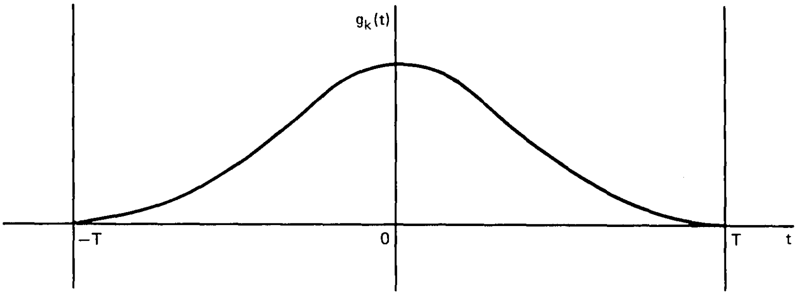


Fig. 7—Raised-cosine signaling with pulse duration $2T$. $K = 2$.

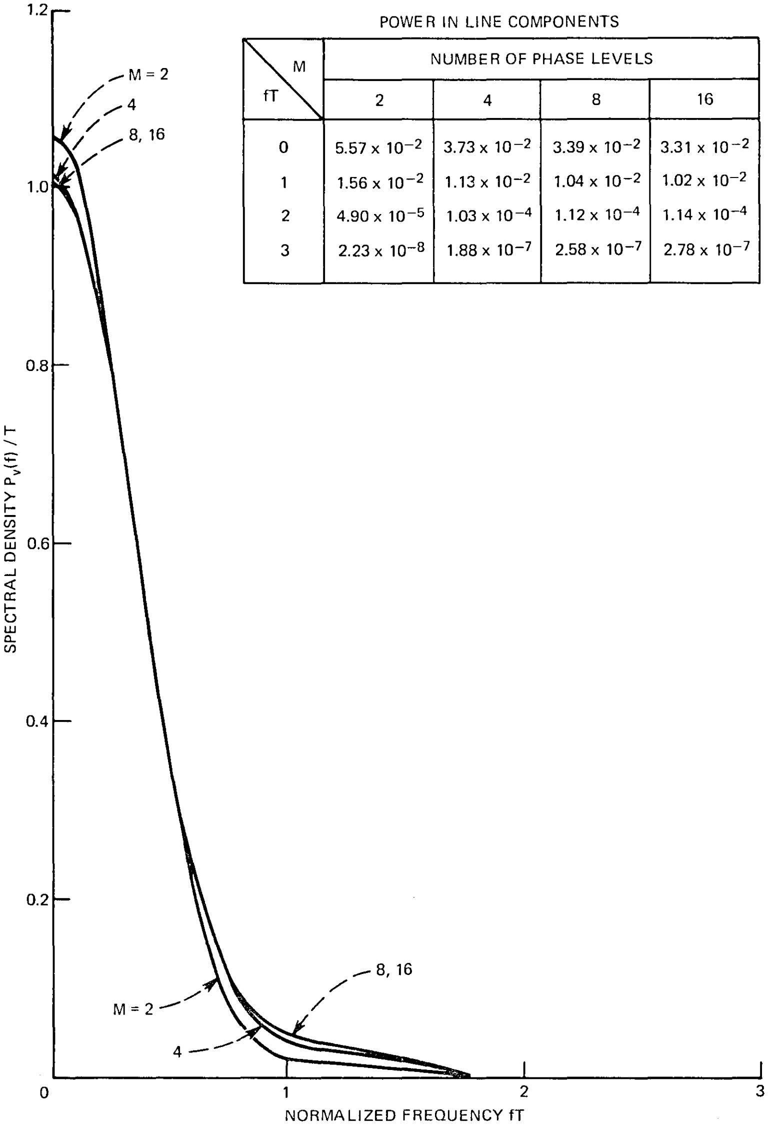
increasingly less time in the neighborhood of 0°. There are also smoother transitions between the phasors with increasing β; this probably explains the decrease in width of the main part of continuous spectrum. The narrowing of the spectral density and the reduction of the discrete spectral lines may or may not indicate better interchannel and intersymbol performance, but this remains a subject for future investigation.

IX. SUMMARY AND CONCLUSIONS

Matrix methods have been used to express the spectral density of a sinusoidal carrier phase modulated by a pulse train with a finite pulse duration. Arbitrary pulse shapes may be used for M-ary digital signaling and they may overlap over a finite number of signaling intervals.

If the pulse duration is one signaling interval, our results give the spectral density even though successive symbols are correlated. If the pulse duration is more than one signaling interval, we assume that symbols transmitted during different time slots are statistically independent; the case of correlated symbols may be considered by extension of the present work, but the required statistical description of the modulation $\mathbf{a}_k$ will be more complicated. For example, if $K = 2$, in addition to $\langle \mathbf{a}_k] \rangle$, $\langle \mathbf{a}_k] \times \lfloor \mathbf{a}_l \rfloor \rangle$, we need to know $\langle \mathbf{a}_i] \times \mathbf{a}_k] \times \lfloor \mathbf{a}_l \rfloor \times \lfloor \mathbf{a}_m \rfloor \rangle$, the fourth order statistics of $\mathbf{a}_k$, to determine the spectral density from our methods.

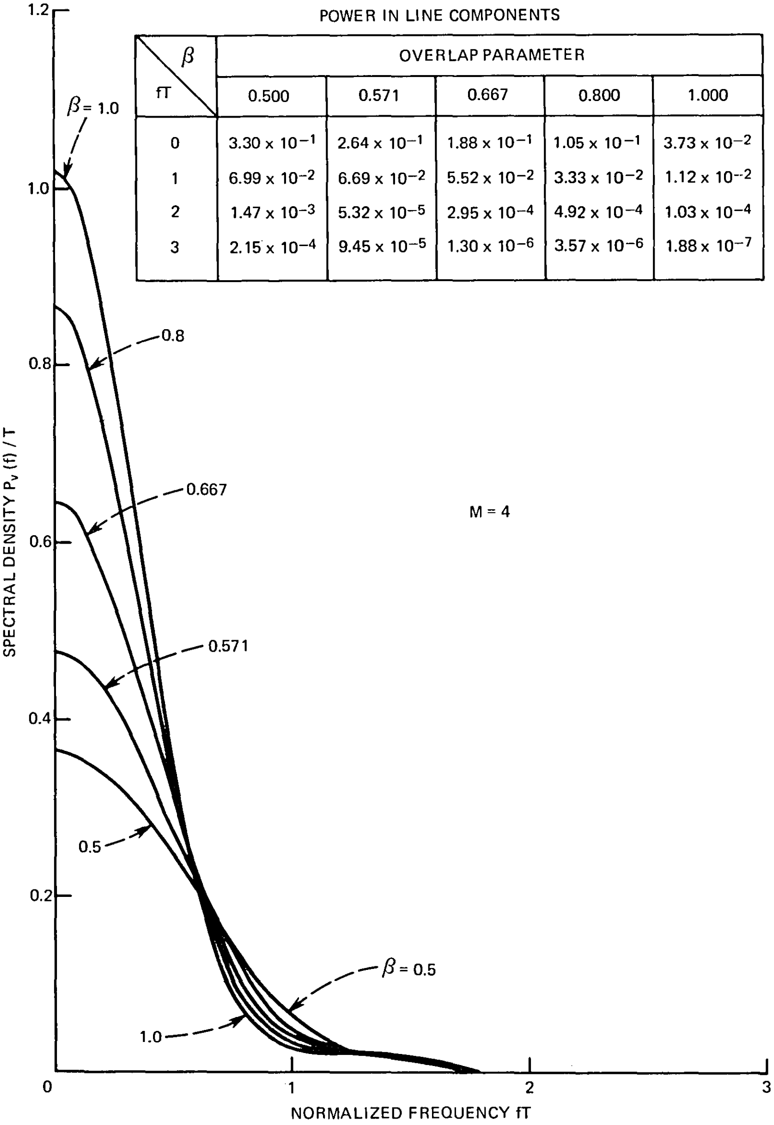
The spectral density has a compact Hermitian form suitable for numerical computation by a digital computer. The associated Hermitian matrix is a function of only the symbol probabilities, and the column matrix associated with the Hermitian form is the Fourier transform of certain time functions related to the signaling pulses. The computations presented in this paper for binary, quaternary, octonary,



POWER IN LINE COMPONENTS

M / fT	NUMBER OF PHASE LEVELS			
	2	4	8	16
0	5.57×10^{-2}	3.73×10^{-2}	3.39×10^{-2}	3.31×10^{-2}
1	1.56×10^{-2}	1.13×10^{-2}	1.04×10^{-2}	1.02×10^{-2}
2	4.90×10^{-5}	1.03×10^{-4}	1.12×10^{-4}	1.14×10^{-4}
3	2.23×10^{-8}	1.88×10^{-7}	2.58×10^{-7}	2.78×10^{-7}

Fig. 8—Spectral density of binary, quaternary, octonary, and 16-ary PSK systems with raised-cosine signaling and pulse duration $2T$. $K = 2$. Although the values of the spectral density for $M = 2$, 4, 8, and 16 are shown to be the same over certain segments of the range of f, they are usually different from each other, but this difference is not large enough to be shown on the linear scale in Fig. 8.



POWER IN LINE COMPONENTS

fT \ β	OVERLAP PARAMETER				
	0.500	0.571	0.667	0.800	1.000
0	3.30×10^{-1}	2.64×10^{-1}	1.88×10^{-1}	1.05×10^{-1}	3.73×10^{-2}
1	6.99×10^{-2}	6.69×10^{-2}	5.52×10^{-2}	3.33×10^{-2}	1.12×10^{-2}
2	1.47×10^{-3}	5.32×10^{-5}	2.95×10^{-4}	4.92×10^{-4}	1.03×10^{-4}
3	2.15×10^{-4}	9.45×10^{-5}	1.30×10^{-6}	3.57×10^{-6}	1.88×10^{-7}

Fig. 9—Spectral density of quaternary PSK system with raised-cosine signaling and pulse duration $2\beta T$, $0.5 \leqq \beta \leqq 1$. When $\beta = 0.5$, $K = 1$, the pulse fills up just one time slot and the spectral density is as shown in Fig. 6. When $\beta = 1$, $K = 2$, the pulse fills up just two time slots, and the spectral density is as shown in Fig. 8. Although the values of the spectral density for $\beta = 0.5$, 0.571, 0.667, 0.8, and 1.0 are shown to be the same over certain segments of the range of f, they are usually different from each other, but this difference is not large enough to be shown on the linear scale in Fig. 9. Note that when $\beta = 0.5$, 0.571, 0.667, 0.8, and 1.0, $1/\beta = 2$, 1.75, 1.5, 1.25, and 1.0.

and 16-ary PSK systems do not consume very much computer time and are quite inexpensive.

Extraction of a timing wave is often essential in the detection and regeneration of digital signals. In a self-timed digital system not containing a timing wave, the wave is extracted from an incoming pulse stream by a nonlinear processing of the signal. For example, the incoming pulse stream may be passed through a square-law rectifier and a linear narrow-band filter to get the timing wave. We would like to note here that the methods given in this paper can be extended to determine the spectral density of the output of a nonlinear device whose input is a random time pulse train of the form (16) or (42). The results presented here apply to the particular nonlinearity exp $j(\cdot)$; other nonlinear functions are treated in a similar manner.

X. ACKNOWLEDGMENT

We would like to thank Wanda L. Mammel for carrying out the computations reported in this paper, and also for her help in determining the numerical accuracy of the results.

APPENDIX A

Spectra of Complex and Real PSK Waves

From (1) written in complex form, the covariance of the real wave $x(t)$ is

$$\begin{aligned}\Phi_x(\tau) &= \overline{\Phi_x(t,\tau)}\\ &= \tfrac{1}{4}\{e^{j2\pi f_c\tau}\Phi_v(\tau)+e^{-j2\pi f_c\tau}\Phi_v^*(\tau)+e^{j2\pi f_c\tau}\overline{e^{j4\pi f_ct}\Phi_{vv^*}(t,\tau)}\\ &\qquad +e^{-j2\pi f_c\tau}\overline{e^{-j4\pi f_ct}\Phi_{vv^*}^*(t,\tau)}\},\end{aligned} \tag{130}$$

where the cross-variance is given by

$$\Phi_{vv^*}(t,\tau) = \langle v(t+\tau)v(t)\rangle = \langle e^{j[\phi(t+\tau)+\phi(t)]}\rangle. \tag{131}$$

Now $\Phi_{vv^*}(t,\tau)$, with τ regarded as a parameter, is periodic in t with period T;

$$\Phi_{vv^*}(t+T,\tau) = \Phi_{vv^*}(t,\tau). \tag{132}$$

This follows from the assumed strict stationarity of the s_k of (2). Then we may write

$$\Phi_{vv^*}(t,\tau) = \sum_{n=-\infty}^{\infty} \varphi_n(\tau)e^{jn2\pi t/T}. \tag{133}$$

Substituting in the time average quantity of (130) and interchanging the order of summation and integration,

$$\begin{aligned}\overline{e^{j4\pi f_c t}\Phi_{vv^*}(t,\,\tau)} &= \sum_{n=-\infty}^{\infty} \varphi_n(\tau) \lim_{A\to\infty} \frac{1}{2A}\int_{-A}^{A} e^{j2\pi[2f_c+(n/T)]t}dt \\ &= \sum_{n=-\infty}^{\infty} \varphi_n(\tau) \lim_{A\to\infty} \frac{\sin 2\pi[2f_c+(n/T)]A}{2\pi[2f_c+(n/T)]A}.\end{aligned} \tag{134}$$

The $\lim\limits_{A\to\infty} \to 0$ if $2f_c + (n/T) \neq 0$ for every integer n, i.e., if twice the carrier frequency is not a precise multiple of the modulation frequency $1/T$. Under this condition, (130) and (134) yield (5).

Next, assume that $\mathbf{P}_v(f)$ is strictly band-limited;

$$\mathbf{P}_v(f) = 0, \quad |f| \geqq f_c. \tag{135}$$

Then $\mathbf{P}_v(f - f_c)$ and $\mathbf{P}_v(-f - f_c)$, the two terms appearing in (5), do not overlap. Now define

$$\nu(t) \equiv e^{j2\pi f_c t}v(t). \tag{136}$$

Then

$$\begin{gathered}\Phi_\nu(t,\,\tau) = e^{j2\pi f_c\tau}\Phi_v(t,\,\tau); \quad \Phi_\nu(\tau) = e^{j2\pi f_c\tau}\Phi_v(\tau). \\ \mathbf{P}_\nu(f) = \mathbf{P}_v(f - f_c); \quad \mathbf{P}_{\nu^*}(f) = \mathbf{P}_v(-f - f_c).\end{gathered} \tag{137}$$

Therefore, subject to (135)

$$\begin{array}{ll} & f \leqq 0. \\ \mathbf{P}_\nu\,(f) = 0, & \\ & f \geqq 2f_c. \\ & \\ & f \leqq -\,2f_c. \\ \mathbf{P}_{\nu^*}(f) = 0, & \\ & f \geqq 0. \end{array} \tag{138}$$

Now the cross-variance of $\nu(t)$ and $\nu^*(t)$ is

$$\Phi_{\nu\nu^*}(\tau) = \overline{\Phi_{\nu\nu^*}(t,\,\tau)} = e^{j2\pi f_c\tau}\,\overline{e^{j4\pi f_c t}\Phi_{vv^*}(t,\,\tau)}, \tag{139}$$

where Φ_{vv^*} is given by (131). The Fourier transform of (139) is the corresponding cross-spectrum $\mathbf{P}_{\nu\nu^*}(f)$. Since the two spectra $\mathbf{P}_\nu(f)$ and $\mathbf{P}_{\nu^*}(f)$ do not overlap by (138), the cross-spectrum $\mathbf{P}_{\nu\nu^*}(f)$ must be identically zero,[16] and hence also the cross-variance;

$$\Phi_{\nu\nu^*}(\tau) = 0, \quad \mathbf{P}_v(f) = 0 \quad \text{for} \quad |f| \geqq f_c. \tag{140}$$

Substituting (139) and (140) into (130) and taking the Fourier transform,

$$\mathbf{P}_x(f) = \tfrac{1}{4}\mathbf{P}_v(f - f_c) + \tfrac{1}{4}\mathbf{P}_v(-f - f_c),$$
$$\mathbf{P}_v(f) = 0 \quad \text{for} \quad |f| \geqq f_c. \tag{141}$$

While $\mathbf{P}_v(f)$ will never be strictly zero, it will eventually fall off so rapidly with increasing $|f|$ that the spectral relation of (5) will be a good approximation when f_c is large enough so that the two terms do not overlap appreciably, even if $2f_c$ is an integral multiple of $1/T$.

Consider now the exceptional case of a low carrier frequency that is an integral multiple of half the modulation frequency:

$$2f_c + \frac{n_0}{T} = 0. \tag{142}$$

Then in (134) the $\lim\limits_{A \to \infty} \to 1$ for the n_0 term, and $\to 0$ for all other terms as before. Therefore, (130) and (134) yield

$$\Phi_x(\tau) = \tfrac{1}{4}e^{j2\pi f_c\tau}[\Phi_v(\tau) + \varphi_{-2f_cT}(\tau)]$$
$$+ \tfrac{1}{4}e^{-j2\pi f_c\tau}[\Phi_v^*(\tau) + \overset{*}{\varphi}_{-2f_cT}\,(\tau)], \quad 2f_cT = \text{integer}. \tag{143}$$

The exceptional case requires the evaluation of the additional quantity $\varphi_{-2f_cT}(\tau)$, $2f_cT$ = integer, where φ is defined in (133). This may be done similarly to the evaluation of $\Phi_v(\tau)$ performed in the text, but it is not undertaken here.

Finally, consider the wave of (9), with $\phi(t)$ given by (2), with the carrier and modulation phases ϕ_0 and t_0 independent random variables, independent of the modulation. Equation (130) is replaced by

$$\Phi_x(\tau) = \tfrac{1}{4}\{e^{j2\pi f_c\tau}\Phi_v(\tau) + e^{-j2\pi f_c\tau}\Phi_v^*(\tau)$$
$$+e^{j2\pi f_c\tau}\langle e^{j2\phi_0}\rangle\langle e^{j4\pi f_c t_0}\rangle\overline{e^{j4\pi f_c t}\Phi_{vv^*}(t,\ \tau)}$$
$$+e^{-j2\pi f_c\tau}\langle e^{-j2\phi_0}\rangle\langle e^{-j4\pi f_c t_0}\rangle\overline{e^{-j4\pi f_c t}\Phi_{vv^*}^*(t,\ \tau)\}}, \tag{144}$$

where v, Φ_v, and Φ_{vv^*} remain as given in (3), (7), (8), and (131) for $t_0 = 0$. The last two lines of (144) vanish if any of the three quantities

$$\overline{e^{j4\pi f_c t}\Phi_{vv^*}(t,\ \tau)}, \quad \langle e^{j2\phi_0}\rangle, \quad \langle e^{j4\pi f_c t_0}\rangle \tag{145}$$

vanish. Therefore,

$$\mathbf{P}_x(f) = \tfrac{1}{4}\mathbf{P}_v(f - f_c) + \tfrac{1}{4}\mathbf{P}_v(-f - f_c) \tag{146}$$

if any of the following conditions are satisfied:

$$
\begin{aligned}
&2f_cT \neq \text{integer.}\\
&2f_cT = \text{integer, } \phi_0 \text{ is uniformly distributed over an interval}\\
&\qquad\qquad \pi,\ 2\pi,\ 3\pi,\ \cdots. \qquad\qquad (147)\\
&2f_cT = \text{integer, } t_0 \text{ is uniformly distributed over an interval}\\
&\qquad\qquad \frac{T}{2f_cT},\ 2\frac{T}{2f_cT},\ 3\frac{T}{2f_cT},\ \cdots.
\end{aligned}
$$

The first condition is, of course, that of (5). The last two conditions show that suitably randomizing the phase of either the carrier or the modulation suffices to yield the simple spectral result of (5) when $2f_cT$ = integer.

APPENDIX B

Spectrum of a Baseband Pulse Train With Different Pulse Shapes

We determine the spectral density of (43),

$$v(t) = \sum_{k=-\infty}^{\infty} \underbracket{\mathbf{b}_k}\cdot\mathbf{r}(t-kT)], \tag{148}$$

by the vector analog of the corresponding derivation for a train of equally spaced pulses with similar shapes. In this appendix, the waveforms $\mathbf{r}$ for different time slots (i.e., different k) may overlap, although they do not overlap in the applications of this paper. Also, $\mathbf{b}_k$ may be complex in this appendix. It is real in the main part of the paper.

Let $\mathbf{b}_k$ in (43) and (148) be wide-sense stationary. Then, using the notation of (35) to (40):

$$\underbracket{\boldsymbol{\beta}} \equiv \langle \underbracket{\mathbf{b}_k} \rangle. \tag{149}$$

$$\tilde{\mathbf{\Phi}}_{\mathrm{b}}(k-l) \equiv \langle \mathbf{b}_k]\cdot\underbracket{\mathbf{b}_l^*} \rangle. \tag{150}$$

$$\tilde{\mathbf{P}}_{\mathrm{b}}(f) = \sum_{n=-\infty}^{\infty} e^{-j2\pi fn}\tilde{\mathbf{\Phi}}_{\mathrm{b}}(n). \tag{151}$$

$$\tilde{\mathbf{\Phi}}_{\mathrm{b}}(n) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \tilde{\mathbf{P}}_{\mathrm{b}}(f)e^{+j2\pi fn}df. \tag{152}$$

Further, assume that $\mathbf{b}_{k+n}$ and $\mathbf{b}_k$ become uncorrelated as $n \to \infty$:

$$\lim_{n\to\infty} \tilde{\mathbf{\Phi}}_{\mathrm{b}}(n) \equiv \tilde{\mathbf{\Phi}}_{\mathrm{b}}(\infty) = \boldsymbol{\beta}]\cdot\underbracket{\boldsymbol{\beta}^*}. \tag{153}$$

From (148),

$$\Phi_v(t,\tau) = \langle v(t+\tau)v^*(t)\rangle$$

$$= \sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} \underline{\mathbf{r}(t+\tau-kT)}\cdot\tilde{\boldsymbol{\Phi}}_{\mathrm{b}}(k-l)\cdot\mathbf{r}^*(t-lT)]. \tag{154}$$

Since

$$\Phi_v(t+T,\tau) = \Phi_v(t,\tau), \tag{155}$$

we have

$$\Phi_v(\tau) = \overline{\Phi_v(t,\tau)} = \frac{1}{T}\int_{-\mathrm{T}/2}^{T/2}\Phi_v(t,\tau)dt. \tag{156}$$

Substituting (154) in (156) and making the substitutions $k - l = n$ and $t - lT = t'$,

$$\Phi_v(\tau) = \frac{1}{T}\sum_{n=-\infty}^{\infty}\sum_{l=-\infty}^{\infty}\int_{-(l+\frac{1}{2})T}^{-(l-\frac{1}{2})T} \underline{\mathbf{r}(t+\tau-nT)}\cdot\tilde{\boldsymbol{\Phi}}_{\mathrm{b}}(n)\cdot\mathbf{r}^*(t)]dt$$

$$= \frac{1}{T}\sum_{n=-\infty}^{\infty}\int_{-\infty}^{\infty} \underline{\mathbf{r}(t+\tau-nT)}\cdot\tilde{\boldsymbol{\Phi}}_{\mathrm{b}}(n)\cdot\mathbf{r}^*(t)]dt. \tag{157}$$

The spectral density of (148) is now the Fourier transform of (157):

$$\mathbf{P}_v(f) = \int_{-\infty}^{\infty}\Phi_v(\tau)e^{-j2\pi f\tau}d\tau$$

$$= \frac{1}{T}\sum_{n=-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-j2\pi f\tau}\,\underline{\mathbf{r}(t+\tau-nT)}$$

$$\cdot\tilde{\boldsymbol{\Phi}}_{\mathrm{b}}(n)\cdot\mathbf{r}^*(t)]dtd\tau. \tag{158}$$

The treatment differs slightly from that for the scalar case, because the order of the factors in the matrix products may not be changed. Setting $t' = t + \tau$, (158) becomes

$$\mathbf{P}_v(f) = \frac{1}{T}\sum_{n=-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} e^{-j2\pi ft'}\,\underline{\mathbf{r}(t'-nT)}dt'\right\}\cdot\tilde{\boldsymbol{\Phi}}_{\mathrm{b}}(n)$$

$$\cdot\left\{\int_{-\infty}^{\infty} e^{+j2\pi ft}\,\mathbf{r}^*(t)]dt\right\}\cdot \tag{159}$$

Define

$$\mathbf{R}(f)] = \underline{\mathbf{R}(f)}' \equiv \int_{-\infty}^{\infty} e^{-j2\pi ft}\,\mathbf{r}(t)]dt; \tag{160}$$

i.e., the elements of $\mathbf{r}(t)$ and $\mathbf{R}(f)$ are the pulse shapes and their

Fourier transforms, respectively. In applications throughout the remainder of this paper,

$$\mathbf{r}(t)] = 0], \quad t \leqq 0, \quad t > T, \tag{161}$$

and the integral in (160) therefore takes the limits $\int_0^T$. However, the restriction of (161) is not necessary in this appendix. Substituting (160) into (159) and using (151),

$$\tilde{\mathbf{P}}_v(f) = \frac{1}{T} \underbracket{\mathbf{R}(f)} \cdot \tilde{\mathbf{P}}_{\mathrm{b}}(fT) \cdot \mathbf{R}^*(f)], \tag{162}$$

our desired result. Eq. (162) reduces correctly to the scalar case.[17]

It is convenient to separate out the line component of $v(t)$. As in (39) and (40), the line and continuous components of (151) are, using (153),

$$\tilde{\mathbf{P}}_{\mathrm{b}l}(f) = \boldsymbol{\beta}] \cdot \underbracket{\boldsymbol{\beta}^*} \sum_{n=-\infty}^{\infty} \delta(f - n). \tag{163}$$

$$\tilde{\mathbf{P}}_{\mathrm{b}c}(f) = \sum_{n=-\infty}^{\infty} e^{-j2\pi f n} \{ \tilde{\mathbf{\Phi}}_{\mathrm{b}}(n) - \boldsymbol{\beta}] \cdot \underbracket{\boldsymbol{\beta}^*} \}. \tag{164}$$

The line and continuous spectral components of (148) are found by substituting (163) and (164) into (162). For the line component

$$\begin{aligned} \mathbf{P}_{v_l}(f) &= \frac{1}{T^2} \underbracket{\mathbf{R}(f)} \cdot \boldsymbol{\beta}] \cdot \underbracket{\boldsymbol{\beta}^*} \cdot \mathbf{R}^*(f)] \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right) \\ &= \frac{1}{T^2} \left| \underbracket{\boldsymbol{\beta}} \cdot \mathbf{R}(f)] \right|^2 \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right). \end{aligned} \tag{165}$$

The line component is composed of dc and sine waves at the signaling rate and its harmonics.

Finally, taking the expected value of (148) and using (149),

$$\langle v(t) \rangle = \underbracket{\boldsymbol{\beta}} \cdot \sum_{k=-\infty}^{\infty} \mathbf{r}(t - kT)]. \tag{166}$$

Then

$$\begin{aligned} \Phi_{\langle v \rangle}(t, \tau) &= \langle v(t + \tau) \rangle \langle v^*(t) \rangle \\ &= \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \underbracket{\mathbf{r}(t + \tau - kT)} \cdot \boldsymbol{\beta}] \cdot \underbracket{\boldsymbol{\beta}^*} \cdot \mathbf{r}^*(t - lT)]. \end{aligned} \tag{167}$$

Comparing with (154) and proceeding as above,

$$\Phi_{\langle v \rangle}(\tau) = \frac{1}{T} \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} \underline{\mathbf{r}(t + \tau - nT)} \cdot \boldsymbol{\beta}] \cdot \underline{\boldsymbol{\beta}}^{*} \cdot \mathbf{r}^{*}(t)] dt. \qquad (168)$$

Consequently, (168) is a periodic function of τ with period T. Comparing (157) and (168) as $|\tau| \rightarrow \infty$, using (153), and assuming that $\mathbf{r}(t)$ eventually falls off for large $|t|$ [the stronger assumption (161) applies throughout the rest of this paper],

$$\Phi_{v}(\tau) \rightarrow \Phi_{\langle v \rangle}(\tau) \quad \text{as} \quad |\tau| \rightarrow \infty. \qquad (169)$$

Therefore, we may write (148) as

$$v(t) = v_{l}(t) + v_{c}(t), \qquad (170)$$

where, from (166),

$$\begin{aligned} v_{l}(t) &= \underline{\boldsymbol{\beta}} \cdot \sum_{k=-\infty}^{\infty} \mathbf{r}(t - kT)] \\ &= \frac{1}{T} \underline{\boldsymbol{\beta}} \cdot \sum_{n=-\infty}^{\infty} \mathbf{R}\left(\frac{n}{T}\right)] e^{jn2\pi t/T}, \end{aligned} \qquad (171)$$

and the spectral densities of $v_{l}(t)$ and $v_{c}(t)$ are given by (165) and by substituting (164) into (162), respectively. Equation (171) contains phase information lost in (165).

The assumption of (153) follows if the vector coefficients $\mathbf{b}_{k}$ become independent for widely separated time slots. For the applications of this paper, the $\mathbf{b}_{k}$ are unit basis vectors, i.e., each $\mathbf{b}_{k}$ has a single component equal to unity and all other components zero. In this special case, the assumption of (153) that $\mathbf{b}_{k}$ in widely separated time slots are uncorrelated also guarantees independence.[10]

APPENDIX C

Simplification of PSK Spectra for K = 2 for Nonoverlapping Signal Pulses

We demonstrate that the $K = 2$ results, (94) to (96), which apply subject to condition (82), specialize to the $K = 1$ results (66), (67), and (72) when the stronger condition (63) is satisfied.

We distinguish the different **R**'s in Sections V and VI by appending subscripts K, denoting (64) and (83) as $\mathbf{R}_{1}$ and $\mathbf{R}_{2}$, respectively. Then if, in Section VI, (82) is replaced by the stronger condition (63), (83) becomes

$$\mathbf{R}_{2}(f)] = \mathbf{R}_{1}(f)] \times \mathbf{1}], \qquad (172)$$

where $\mathbf{R}_1$ is given by (64) and $\mathbf{1}$ is the M-dimensional vector with all components unity defined in (24).

Considering first (94) and (95), using (26) and (172) (see footnote, p. 908):

$$\begin{aligned}\underline{\mathbf{w}}^{[2]}\cdot\mathbf{R}_2(f)] &= (\underline{\mathbf{w}}\times\underline{\mathbf{w}})\cdot(\mathbf{R}_1(f)]\times\mathbf{1}])\\ &= (\underline{\mathbf{w}}\cdot\mathbf{R}_1(f)])\times(\underline{\mathbf{w}}\cdot\mathbf{1}]) = \underline{\mathbf{w}}\cdot\mathbf{R}_1(f)].\end{aligned}\tag{173}$$

Substituting (173) into (94) and (95) yields (66) and (67).

Next, substituting (172) into (96), we evaluate the following resulting products using (26) and (28) (see footnote, p. 908):

$$\begin{aligned}&\underline{\mathbf{R}_2(f)}\cdot\mathbf{w}_d^{[2]}\cdot\mathbf{R}_2^*(f)]\\ &\qquad= (\underline{\mathbf{R}_1(f)}\times\underline{\mathbf{1}})\cdot(\mathbf{w}_d\times\mathbf{w}_d)\cdot(\mathbf{R}_1^*(f)]\times\mathbf{1}])\\ &\qquad= (\underline{\mathbf{R}_1(f)}\cdot\mathbf{w}_d\cdot\mathbf{R}_1^*(f)])\times(\underline{\mathbf{1}}\cdot\mathbf{w}_d\cdot\mathbf{1}])\\ &\qquad= \underline{\mathbf{R}_1(f)}\cdot\mathbf{w}_d\cdot\mathbf{R}_1^*(f)].\end{aligned}\tag{174}$$

$$\begin{aligned}&\underline{\mathbf{R}_2(f)}\cdot\mathbf{w}]^{[2]}\cdot\underline{\mathbf{w}}^{[2]}\cdot\mathbf{R}_2^*(f)]\\ &\qquad= (\underline{\mathbf{R}_1(f)}\times\underline{\mathbf{1}})\cdot(\mathbf{w}]\times\mathbf{w}])\cdot(\underline{\mathbf{w}}\times\underline{\mathbf{w}})\cdot(\mathbf{R}_1^*(f)]\times\mathbf{1}])\\ &\qquad= (\underline{\mathbf{R}_1(f)}\cdot\mathbf{w}]\cdot\underline{\mathbf{w}}\cdot\mathbf{R}_1^*(f)])\times(\underline{\mathbf{1}}\cdot\mathbf{w}]\cdot\underline{\mathbf{w}}\cdot\mathbf{1}])\\ &\qquad= (\underline{\mathbf{R}_1(f)}\cdot\mathbf{w}])(\underline{\mathbf{w}}\cdot\mathbf{R}_1^*(f)])\\ &\qquad= |\underline{\mathbf{w}}\cdot\mathbf{R}_1(f)]|^2.\end{aligned}\tag{175}$$

$$\begin{aligned}&\underline{\mathbf{R}_2(f)}\cdot(\underline{\mathbf{w}}\times\mathbf{w}_d\times\mathbf{w}])\cdot\mathbf{R}_2^*(f)]\\ &\qquad= (1\times\underline{\mathbf{R}_1(f)}\times\underline{\mathbf{1}})\cdot(\underline{\mathbf{w}}\times\mathbf{w}_d\times\mathbf{w}])\cdot(\mathbf{R}_1^*(f)]\times\mathbf{1}]\times 1)\\ &\qquad= (1\cdot\underline{\mathbf{w}}\cdot\mathbf{R}_1^*(f)])\times(\underline{\mathbf{R}_1(f)}\cdot\mathbf{w}_d\cdot\mathbf{1}])\times(\underline{\mathbf{1}}\cdot\mathbf{w}]\cdot 1)\\ &\qquad= (\underline{\mathbf{w}}\cdot\mathbf{R}_1^*(f)])(\underline{\mathbf{R}_1(f)}\cdot\mathbf{w}])\\ &\qquad= |\underline{\mathbf{w}}\cdot\mathbf{R}_1(f)]|^2.\end{aligned}\tag{176}$$

Substituting (174) to (176) into (96) yields (72).

REFERENCES

1. W. R. Bennett, "Statistics of Regenerative Digital Transmission," B.S.T.J., *37*, No. 6 (November 1958), pp. 1501–1542.
2. W. R. Bennett and S. O. Rice, "Spectral Density and Autocorrelation Functions Associated with Binary Frequency Shift Keying," B.S.T.J., *42*, No. 5 (September 1963), pp. 2355–2385.
3. R. R. Anderson and J. Salz, "Spectra of Digital FM," B.S.T.J., *44*, No. 6 (July–August 1965), pp. 1165–1189.
4. L. Lundquist, "Digital PM Spectra by Transform Techniques," B.S.T.J., *48*, No. 2 (February 1969), pp. 397–411.

5. H. C. Van Den Elzen, "Calculating Power Spectral Densities for Data Signals," Proc. IEEE, *58*, No. 6 (June 1970), pp. 942–943.
6. B. Glance, "Power Spectra of Multilevel Digital Phase-Modulated Signals," B.S.T.J., *50*, No. 9 (November 1971), pp. 2857–2878.
7. J. E. Mazo and J. Salz, "Spectra of Frequency Modulation with Random Waveforms," Information and Control, *9*, No. 4 (August 1966), pp. 414–422.
8. O. Shimbo, "General Formula for Power Spectra of Digital FM Signals," Proc. IEE(GB), *113*, No. 11 (November 1966), pp. 1783–1789.
9. H. E. Rowe, *Signals and Noise in Communication Systems*, New York: D. Van Nostrand, 1965, pp. 98–119.
10. W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd edition, New York: John Wiley, 1957, p. 224, Prob. 16.
11. Ref. 9, pp. 22–25, 50–52.
12. R. Bellman, *Introduction to Matrix Analysis*, New York: McGraw-Hill, 1970, Ch. 12.
13. F. A. Graybill, *Introduction to Matrices with Applications in Statistics*, New York: Wadsworth, 1969, Section 8.8.
14. S. O. Rice, "Phase Jitter Produced by Overlapping Pulses in a Pulse Communication System," unpublished work, 1967.
15. B. Glance, "Power Spectrum of a Carrier Phase-Modulated by Polar Overlapping Pulses," unpublished work.
16. Ref. 9, pp. 49–50.
17. Ref. 9, pp. 236–237.

Copyright © 1974 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 5, May–June 1974
Printed in U.S.A.

On the Correlation Between Bit Sequences in Consecutive Delta Modulations of a Speech Signal

By N. S. JAYANT

(Manuscript received November 16, 1973)

We consider a communication link in which a band-limited speech signal is delta-modulated, detected, and filtered by a low-pass filter, and the analog output is delta-modulated again with an identical encoder. We are concerned with the correlation C between equal-length bit sequences, designated $\{b\}$ and $\{B\}$, that result from the two stages of delta modulation. We study C as a function of the sequence length W; the starting sample T in $\{b\}$; the time shift L between $\{b\}$ and $\{B\}$; the signal-sampling frequency F; and a parameter $P(\geqq 1)$ which specifies the speed of step-size adaptations in the delta modulators. ($P = 1$ provides nonadaptive, or linear, delta modulation.)

Computer simulations have confirmed that for small time shifts L and for statistically adequate window lengths W, C is a strong positive number (0.4, for example). Moreover, the C function tends to exhibit a maximum $C_{\max}$ at a small nonzero value of L (between 1 and 5, say) reflecting a delay introduced by the low-pass filter preceding the second delta modulator; and when W is on the order of 100 or more, the dependence of $C_{\max}$ on the starting sample T is surprisingly weak. Also, in the range of F and P values included in our simulation, $C_{\max}$ increased with F and decreased with P. Finally, the positive C values for small L are retained even when the delta modulators are out of synchronization in amplitude level and step size, as long as the delta modulators incorporate leaky integrators and finite, nonzero values for maximum and minimum step size.

With a given T, the $C(L)$ function can exhibit significant nonzero values even for large L. However, these values are both positive and negative; and if correlations are averaged over several values of T, the average $C(L)$ function tends to be essentially zero for sufficiently large L ($L \geqq 100$, say), while still preserving the strong positive peaks at a predictable small value of L. This observation is the basis of an interesting application

where the value of C is used to determine whether or not two digital codes, appearing at different points in a speech communication system, carry identical speech information.

I. THE PROBLEM

Consider a speech signal subjected to two successive stages of delta modulation, with an intermediate stage of low-pass filtering, as shown in Fig. 1. A previous paper[1] has studied how signal quality degrades as a function of the number of delta modulations. The present paper is concerned with the amount of correlation that exists between the bit sequences $\{b\}$ and $\{B\}$ from the two (identical) delta modulators. Specifically, we employ computer simulations to study the correlation

$$C = \frac{1}{W} \sum_{i=T}^{T+W} b_i B_{i+L}.$$

It is assumed that $\{b\}$ and $\{B\}$ are zero-mean sequences with equiprobable ± 1 entries. Apart from being a function of the window duration W and time shift L, the correlation C will also depend on the signal-sampling frequency F and a parameter P specifying the step-size logic used in the delta modulators. The delta-modulator simulations are described in Section II and the properties of C are described in succeeding sections.

The studies reported in this paper were prompted by an interesting potential application where the value of the correlation C would be used to determine whether or not two digital codes (appearing at different points in a speech communication network) carry the same speech information. More specifically, we were considering a telephonic system that incorporated digital and analog signal terminals capable of being interconnected via a common switching network. The problem was to determine whether digital terminals communicating with each other (in other words, handling the same speech information) could be detected by digitally correlating the signals of each digital terminal with the signals at other digital terminals in the system.[2,3] The digital coding under consideration was delta modulation, and the results of this paper indeed suggest that the detection of communicating terminals should be possible on the basis of appropriate bit correlations.

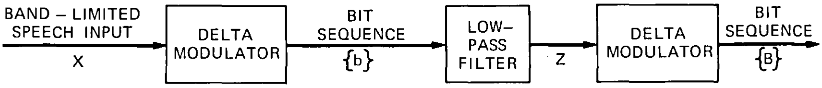


Fig. 1—Block diagram of the simulated speech communication system.

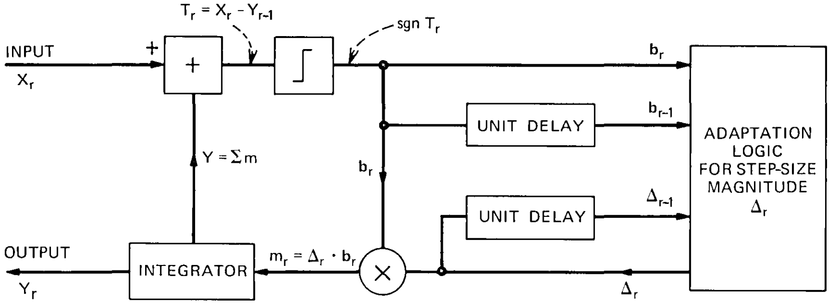


Fig. 2—Schematic diagram of an adaptive delta modulator.

II. SIMULATION DETAILS

The delta modulator utilized in our simulations is schematized in Fig. 2 and is the same instantaneously adaptive delta modulator (ADM) discussed in Ref. 4. Basically, it is described by the equations

$$b_r = \text{sgn}\,(X_r - Y_{r-1}),$$
$$Y_r = Y_{r-1} + \Delta_r \cdot b_r,$$

and

$$\Delta_r = \Delta_{r-1} \cdot P^{b_r \cdot b_{r-1}},$$

where X_r is the amplitude of the input sample r, and Y_{r-1} is the amplitude of the latest staircase approximation to it. The parameter P ($\geqq 1$) automatically increases step size when Y is not tracking X fast enough ($b_r = b_{r-1}$), and decreases it when Y is hunting around X ($b_r = -b_{r-1}$). Nonadaptive or linear delta modulation (LDM) corresponds to the special case of $P = 1$.

The speech signal is a 1.5-second male utterance of "Have you seen Bill?" that is band-limited to 3.3 kHz. The sampling rate, unless otherwise noted, is 60 kHz. A plot of the speech waveform appears in Fig. 3, where a number at the right of a line represents the last 60-kHz sample in that line. The original signal samples are quantized to a 12-bit accuracy, and have integer amplitudes in the range -2^{11} to $+2^{11}$. Finally, the low-pass filter is a programmed recursive filter with an 18-dB/octave roll-off. This seems to represent adequate filtering for toll-quality speech reproduction using ADM at 60 kHz.

III. DEPENDENCE OF CORRELATION ON TIME SHIFT

Figure 4 shows the dependence of C on the time-shift L for two different values of starting sample T. It is interesting to observe that both the functions show a maximum at $L = L_{\max} = 4$. Even more

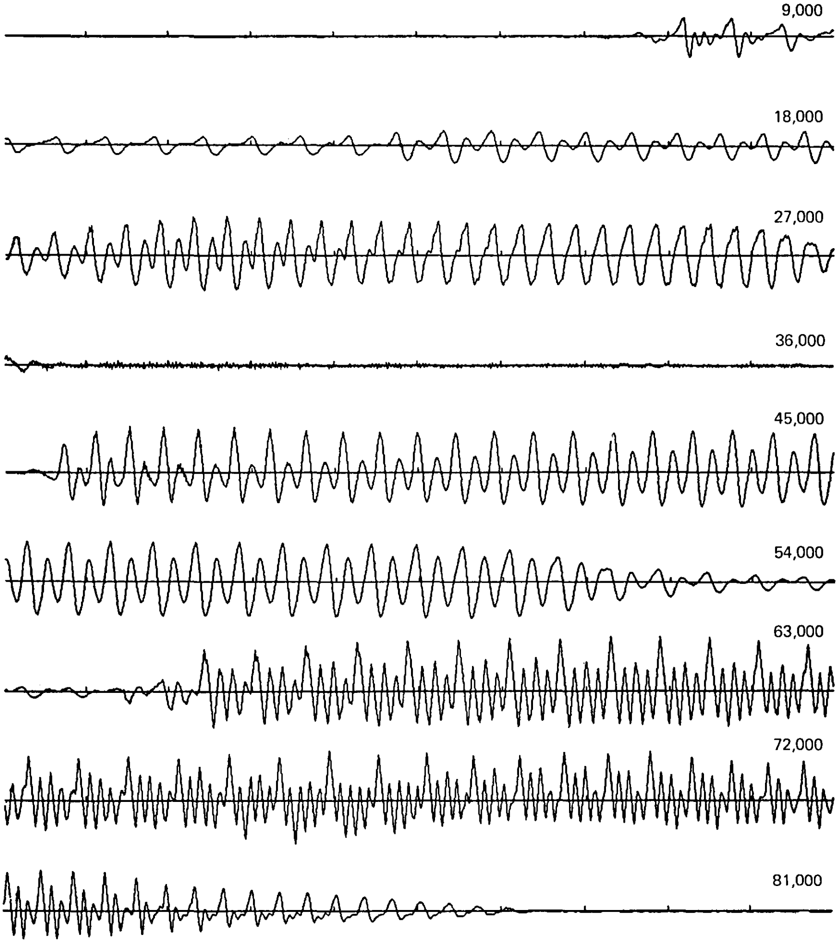


Fig. 3—The speech waveform of "Have you seen Bill?"

interesting is the fact that respective values of $CXZ(L)$, the correlation between the speech input X and the low-pass filter output Z, are also maximized (as determined in a separate simulation) at $L = 4$. It would seem that the nonzero value of $L_{\max}$ in Fig. 4 is to be attributed to the delay introduced by the low-pass filter. Actual values of $C_{\max}$ and $L_{\max}$ depend on the short-term speech spectrum and the nature of low-pass filtering, as determined by the parameters T and F (see Tables I and II). It is a general result, however, that the $C(L)$ function always shows a unique, strongly positive maximum value at a small

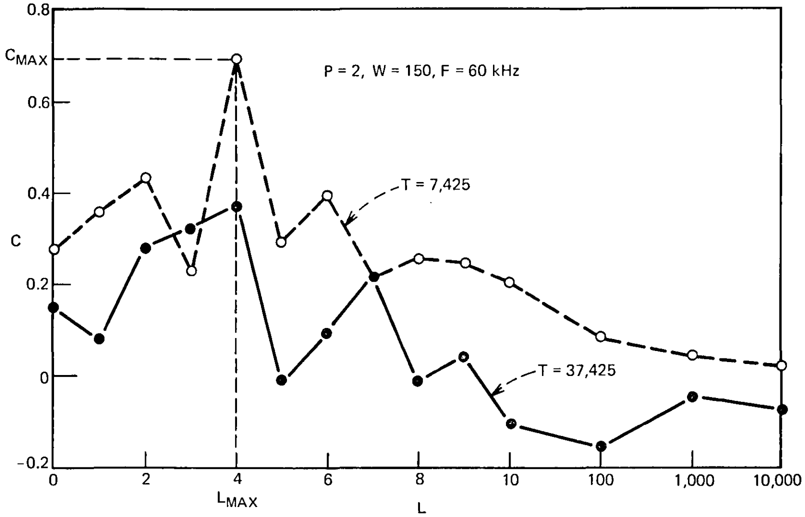


Fig. 4—Dependence of correlation C on time shift L.

value of L. Secondary peaks at large values of L tend to be less unique, and they tend to be randomly positive and negative depending on the part of the speech utterance being considered, as determined by T.

IV. DEPENDENCE OF MAXIMUM CORRELATION ON SAMPLING FREQUENCY

Table I indicates a tendency for $C_{\max}$ to decrease with decreasing sampling frequency. This may be ascribed to the fact that, at a lower sampling rate, delta modulation provides a cruder approximation to the input signal. The bits, therefore, carry more signal-independent noise information, and they have corresponding random properties that cause a decorrelation between $\{b\}$ and $\{B\}$.

Table I — Dependence of maximum correlation $C_{\max}$ on sampling frequency F ($P = 2$, $W = 1000$)

F	60 kHz		40 kHz	
T	$L_{\max}$	$C_{\max}$	$L_{\max}$	$C_{\max}$
17425	4	0.46	3	0.32
37425	4	0.37	1	0.28
57425	4	0.48	2	0.33

Table II — Dependence of correlation C on starting sample T and time shift L ($W = 150$, $P = 2$, $F = 60$ kHz; numbers in parentheses are values of $L_{\max}$)

L T	0	$L_{\max}$	10	100	1000	10000
7425 (4)	0.27	0.69	0.20	0.08	0.04	0.01
17425 (4)	0.35	0.48	0.24	−0.09	−0.04	−0.11
27425 (5)	0.23	0.47	0.11	−0.03	−0.03	0.03
37425 (4)	0.15	0.37	−0.11	−0.16	−0.08	−0.08
47425 (2)	0.29	0.33	0.31	−0.13	−0.11	−0.08
57425 (4)	0.37	0.43	0.11	−0.04	0.21	0.13
67425 (1)	0.39	0.44	0.37	0.27	0.22	−0.03
Average of C values (over T)	0.29	0.46	0.18	−0.01	0.03	−0.02

V. DEPENDENCE OF MAXIMUM CORRELATION ON STEP-SIZE MULTIPLIER P

Table III demonstrates how $C_{\max}$ tends to decrease with increasing P. Larger values of P increase the high-frequency excursions of the staircase function Y. These are filtered out by the low-pass filter. This leads to lesser correlation between the filter output Z and the bit sequence $\{b\}$ and, thence, to a decorrelation of $\{B\}$ and $\{b\}$.

VI. DEPENDENCE OF MAXIMUM CORRELATION ON WINDOW LENGTH

Our results so far have tacitly assumed window length values that represent bit sequences whose durations are of the order of a few milliseconds. Figure 5 shows C explicitly as a function of W. It is seen that very stable indications result with W in the order of 1000, although values close to a respective asymptote are sometimes reached for W values in the order of 100. In fact, a window length of $W = 10$ is seen to be sufficient, for all values of T in Fig. 5, to bring out the strong positive nature of $C_{\max}$. The convergence of the three curves in Fig. 5

Table III — Dependence of maximum correlation $C_{\max}$ on step-size multiplier P ($F = 60$ kHz, $W = 1000$)

T	37425		37000	
P	$L_{\max}$	$C_{\max}$	$L_{\max}$	$C_{\max}$
1.0	4	0.91	4	0.89
1.5	4	0.66	4	0.62
2.0	4	0.34	4	0.44

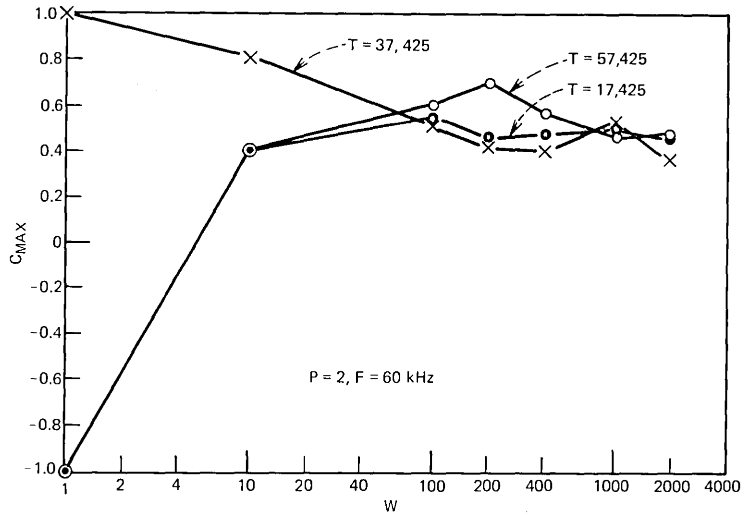


Fig. 5—Dependence of correlation C on window length W.

is not at all surprising. Note that, by definition, C should indeed be independent of T for $W \rightarrow \infty$. The results of Fig. 5 were based on a search for $C_{\max}$ in the range $0 \leqq L \leqq 10$. Except for $W = 1$, unique maxima were noted at nonzero values of L. For $W = 1$, the value of $C_{\max}$ was surprisingly constant in the range $0 \leqq L \leqq 10$, the constant value being $+1$ for one value of T, and -1 for the other two.

VII. DEPENDENCE OF MAXIMUM CORRELATION ON WINDOW LOCATION

As seen in the last section, $C_{\max}$ is a significant function of the starting sample for finite W. Table II shows the values of $C_{\max}$ for seven equally spaced values of T. The average value of $C_{\max}$ is 0.46 and the standard deviation is only 0.10. Note also that C values for large L are smaller in general, and the effect is more noticeable when correlations are averaged over T. This is because the positive $C_{\max}$ values always add up, while C values for large L, being randomly positive or negative, tend to average out to values close to zero.

At least one interesting application of the preceding observations has been suggested.[2,3] Suppose the second delta modulator has several potential speech inputs including the input Z resulting from X. The function C would then assume the strong positive values of Table II only when the input to the second delta modulator is indeed the DM version of the speech X; and it would show values of $C \rightarrow 0$ if the input

Table IV—The effect of unsynchronized delta modulators ($T = 37425$, $P = 2$, $F = 60$ kHz, $W = 1000$. Values in parentheses are for $W = 150$)

Case	Initial Conditions				Integrators	Limits Step Size	$L_{\max}$	$C_{\max}$
	Y_1	Y_2	$\vec{\Delta}_1$	$\vec{\Delta}_2$				
I	0	0	1	1	Perfect	(0, ∞)	4 (4)	0.34 (0.37)
II	0	0	1	1	Leaky	(25, 250)	2 (3)	0.48 (0.83)
III	0	−50	1	−10	Leaky	(25, 250)	2 (3)	0.47 (0.76)
IV	0	−50	1	−10	Perfect	(0, ∞)	5 (1)	0.11 (0.16)

was an entirely different speech signal* (possibly due to a different speaker). This effect will be more pronounced if the averaging process indicated in Table II is carried out. We are suggesting, in other words, a means of determining whether or not two digital DM codes, appearing at different points in a speech communication network, carry the same segment of speech information. The basic recipe is a DM bit correlator with a window of 0.1 to 1 ms, and a window location T that seems to be quite uncritical, especially when time diversity (averaging over T) is possible.

VIII. EFFECT OF UNSYNCHRONIZED DELTA MODULATORS

In practice, the two delta modulators in Fig. 1 can be unsynchronized in amplitude Y and step size Δ when either or both of them are in some kind of a transient state of operation. It is an interesting result of our study that the strong positive values of $C_{\max}$ are retained even during such asynchronous periods, provided the delta modulators operate with a leaky integrator, and with finite and nonzero limits on step size. Leaky integration decreases the effect of amplitude history and, hence, the effect of amplitude asynchrony. Finite and nonzero limits on step size provide potential meeting points for the two step-size sequences, although they may begin with a different starting value.

In Table IV, Y_1 and Y_2 represent initial amplitudes for the two delta modulators, while $\vec{\Delta}_1$ and $\vec{\Delta}_2$ are the initial (signed) step sizes. The step-size limits, 250 (maximum) and 25 (minimum), include a significant range of step sizes that are called for in the adaptive delta modulation of speech (with $F = 60$ kHz, and with signal amplitudes in the range -2^{11} to $+2^{11}$).[4] Finally, the leaky integrators of Table IV leaked 5 percent of signal amplitude in a sampling period.

* This situation is hypothesized to be equivalent to the case of large L.

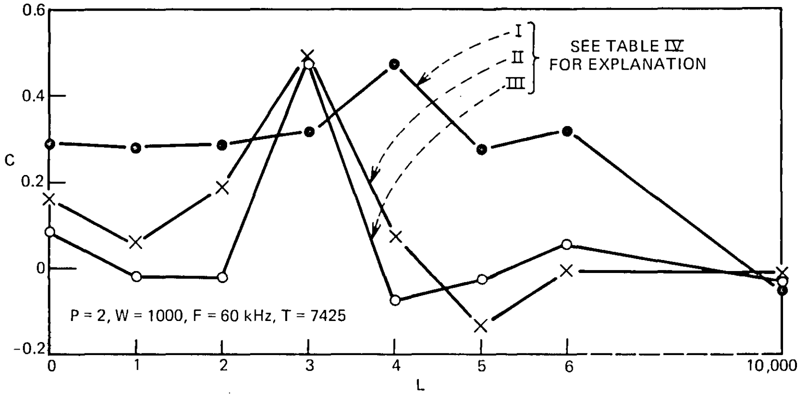


Fig. 6—Dependence of correlation C on time shift $L \leqq$—example of unsynchronized delta modulators.

Note that Table IV shows that leaky integration and finite, nonzero step-size limits are imperative in the asynchronous case (rows III and IV, Table IV) to preserve a strong positive $C_{\max}$; they are also desirable to boost the value of $C_{\max}$ in the synchronous case (rows I and II, Table IV). (The boost is quite significant for $W = 150$). A separate simulation showed that finite (nonzero) step-size limits and leaky integrators were effective only when employed in unison; and in one study of C as a function of L, they also sharpened the peak at $L_{\max}$ (see Fig. 6).

Finally, Table V is a counterpart of Table II for the case of unsynchronized encoders. The step-size limits are 25 and 250, the leak is 1 percent in a sample duration, and $P = 1.5$. (The last two numbers are probably more representative than the corresponding values in

Table V — Dependence of correlation C on starting time T and time shift L with unsynchronized delta modulators ($W = 10$, $P = 1.5$, $F = 60$ kHz)

L T	0	$L_{\max}$	10	100	1000	10000
7425	−0.4	0.0	0.0	0.0	0.0	0.0
27425	−0.4	0.4	−0.4	−0.4	−0.2	0.2
47425	0.4	0.4	0.4	−0.4	−0.2	−0.2
67425	0.6	0.6	0.6	0.6	0.6	0.0
Average of C values (over T)	0.05	0.35	0.15	−0.05	0.05	0.00

Table IV.) Finally, we have reduced the window duration to $W = 10$. This results in obviously crude $C(L)$ functions for a given beginning sample T. But, as in Table II, when $C(L)$ values are averaged over T, the resulting C function shows a clear tendency to decay to near-zero values for $L \geqq 100$. The values of $C_{\max}$ in Table V represent largest values as seen in a finite search ($0 \leqq L \leqq 5$). None of these was a unique maximum, which is possibly due to the insufficient duration (0.16 ms) of the window, $W = 10$.

REFERENCES

1. N. S. Jayant and K. Shipley, "Multiple Delta Modulation of a Speech Signal," Proc. IEEE (Letters), September 1971, p. 1382.
2. J. L. Flanagan and N. S. Jayant, "Digital Detection of Intra-Office Calling," unpublished work.
3. J. L. Flanagan and N. S. Jayant, "Digital Signal Detection in Telephonic Communication Systems," U. S. Patent Application, December 12, 1973.
4. N. S. Jayant, "Adaptive Delta Modulation with a One-Bit Memory," B.S.T.J., *49*, No. 3 (March 1970), pp. 321–342.

Contributors to This Issue

Lane H. Brandenburg, B. S., 1962, M. S., 1963, Ph.D., 1968, Columbia University; Bell Laboratories, 1968—. Mr. Brandenburg has worked on various analytical problems associated with communication theory. Member, IEEE, Sigma XI, Tau Beta Pi.

Yo-Sung Cho, B.S.E.E., 1962, Seoul (Korea) National University; M.S., 1966, and Ph.D., 1968, Yale University; Honeywell E.D.P. Division, 1964–1965 and 1967–1969; Bell Laboratories, 1969—. Mr. Cho has made equalization studies of the L5 Coaxial Transmission System employing manual and automatic equalizers. He was also engaged in the development of the equalizer adjustment system which is used for the optimal equalization of the L5 line. His subsequent work includes exploratory repeater amplifiers for the next generation coaxial transmission system. He is presently supervisor of the group developing terminal multiplexing equipment for coaxial and radio transmission systems. Member, IEEE.

Lewis M. Goodrich, B.S.E.E., 1950, University of Rochester; M.S.E.E., 1952, Ohio State University; Bell Laboratories, 1952—. Mr. Goodrich was first engaged in systems studies and testing in underwater sound. In 1961, he began work on test and analysis of No. 5 crossbar circuits. In 1972, he started a 20-month assignment in system test planning for electronic switching systems. He is currently assigned to the No. 5 Crossbar Maintenance Circuit Design Group where he is working on minicomputer applications to central office maintenance. He has recently been an instructor in an in-hours course in "Logic Design of Switching Systems" and is currently teaching an INCEP course in "Electronic Switching Systems." Member, Tau Beta Pi.

Wayne S. Holden, Electronics Technology, 1970, RCA Institutes; Bell Laboratories, 1970—. Mr. Holden has been involved in the evaluation of optical fiber parameters and the design of electronic circuitry for optical fiber communication systems.

Nuggehally S. Jayant, B.Sc., 1962, University of Mysore (India); B.E. (Distinction), 1965, and Ph.D., 1970, Indian Institute of Science, Bangalore; Research Associate, Stanford Electronics Laboratories, 1967–68; Visiting Scientist, Indian Institute of Science, January–March, 1972; Bell Laboratories, 1968—. Mr. Jayant has worked on digital communication in the presence of burst-noise, on the detection of fading signals, on pattern discrimination problems, and on adaptive quantizers for waveform encoding.

Stanley Kaufman, B.E., 1948, and M.S., 1957, Johns Hopkins University; Bellcomm, 1968–1971; Bell Laboratories, 1972—. Mr. Kaufman has worked principally in structures and structural dynamics. At Bellcomm, Mr. Kaufman developed a model for the stability and performance of the Lunar Roving Vehicle. Associate Fellow, AIAA.

Debasis Mitra, B.Sc. (E.E.), 1964, and Ph.D. (E.E.), 1967, University of London; United Kingdom Atomic Energy Authority Research Fellow, 1965–1967; University of Manchester, U.K., 1967–1968; Bell Laboratories, 1968—. Mr. Mitra, a member of the Mathematics of Physics and Networks Department, is interested in the application of mathematical methods to physical problems.

S. D. Personick, B.E.E., 1967, City College of New York; S. M., 1968, E.E., 1969, and Sc.D., 1969, Massachusetts Institute of Technology; Bell Laboratories, 1967—. Mr. Personick is engaged in studies of optical communication systems.

Vasant K. Prabhu, B.E. (Dist.), 1962, Indian Institute of Science, Bangalore, India; S.M., 1963, Sc.D., 1966, Massachusetts Institute of Technology; Bell Laboratories, 1966—. Mr. Prabhu has been concerned with various theoretical problems in solid-state microwave devices, noise, and optical communication systems. Member, IEEE, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and Commission 6 of URSI.

Harrison E. Rowe, B. S., 1948, M.S., 1950, and Sc.D., 1952, Massachusetts Institute of Technology; Bell Laboratories, 1952—. Mr. Rowe's fields of interest have included parametric amplifier theory, noise and communication theory, modulation theory, propagation in random media, and related problems in waveguide, radio, and optical systems. Fellow, IEEE; member, Sigma XI, Tau Beta Pi, Eta Kappa Nu, and Commission 6 of URSI.

R. D. Standley, B.S., 1957, University of Illinois; M.S., 1960, Rutgers University; Ph.D., 1966, Illinois Institute of Technology; USASRDL, Ft. Monmouth, N.J., 1957–1960; IIT Research Institute, Chicago, 1960–1966; Bell Laboratories, 1966—. Mr. Standley has been engaged in research projects involving microwave, millimeter wave, and optical components. He is presently concerned with electron beam lithography as applied to fabrication of integrated optic devices. Member, IEEE, Sigma Tau, Sigma Xi.

Aaron D. Wyner, B. S., 1960, Queens College; B.S.E.E., 1960, M.S., 1961, and Ph.D., 1963, Columbia University; Bell Laboratories, 1963—. Mr. Wyner has worked on information and communication theory and related mathematical problems. From 1969 to 1970 he visited the Department of Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel, and the Faculty of Electrical Engineering, the Technion, Haifa, Israel, on a Guggenheim Foundation Fellowship. He has been a member of the faculty of Columbia University and the Polytechnic Institute of Brooklyn. He has been chairman of the Metropolitan New York Chapter of the IEEE Information Theory Group, and has served as an associate editor of the Group's *Transactions* and as cochairman of two international symposia. He is presently Second Vice-President of the IEEE Information Theory Group. Member, IEEE, AAS, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 5, May–June, 1974
Printed in U.S.A.

B.S.T.J. BRIEF

Optical Waveguides With Very Low Losses

By W. G. FRENCH, J. B. MACCHESNEY,
P. B. O'CONNOR, and G. W. TASKER

(Manuscript received April 25, 1974)

Low-loss optical fibers may be necessary for economical optical transmission systems. We have developed fibers that exhibit losses of less than 2 dB/km, at 1.06 μm. The fibers were made by a chemical vapor deposition (CVD) technique that employs simultaneous reaction and fusion to a clear glassy core material.

Two fiber compositions have been used. In the first fiber, a GeO_2-doped fused-silica core is deposited inside a fused-quartz tube that acts as the cladding after the tube is collapsed into a rod and pulled into a fiber.

Figure 1 shows the loss spectrum of a fiber made in this manner. The fiber is 723 m long and has a core approximately 35 μm in diameter. The numerical aperture is 0.235. The loss decreases by approximately λ^{-4}, the expected Rayleigh scattering dependence, to a minimum just under 2 dB/km at 1.06 μm. Hydroxyl-ion-related absorptions at 0.72, 0.88, and 0.95 μm are low, amounting to less than 10 dB/km at 0.95 μm. We believe that the OH impurities causing these absorptions are due to siloxane present in the $SiCl_4$ starting material. This can be removed by fractional distillation, and loss peaks due to the hydroxyl-ion-related absorptions as low as 2 dB/km above background at 0.95 μm have been observed in similar fibers. This process has been used to produce GeO_2-doped fibers with numerical apertures as high as 0.35,

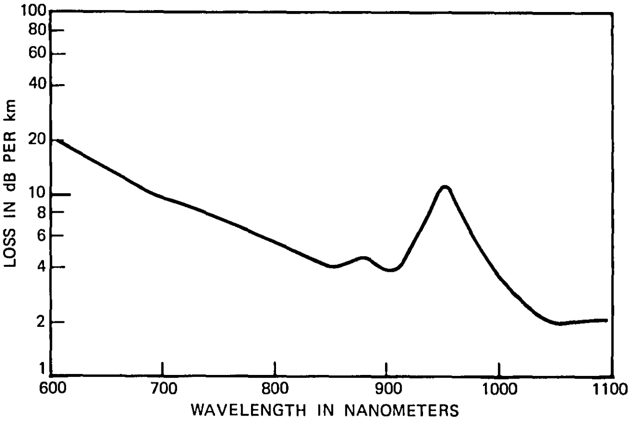


Fig. 1—Loss spectrum of a fiber waveguide with a GeO_2–SiO_2 core and a SiO_2 cladding.

and lengths up to 1.2 km. The length is presently limited by the available fiber-drawing facilities.

Figure 2 illustrates the loss spectrum of a second type of fiber consisting of a pure fused-silica core and borosilicate cladding. In this

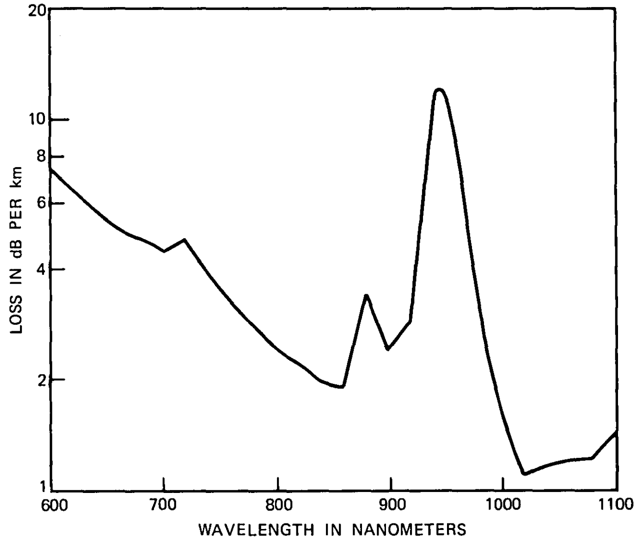


Fig. 2—Loss spectrum of a fiber waveguide with a SiO_2 core and B_2O_3–SiO_2 cladding.

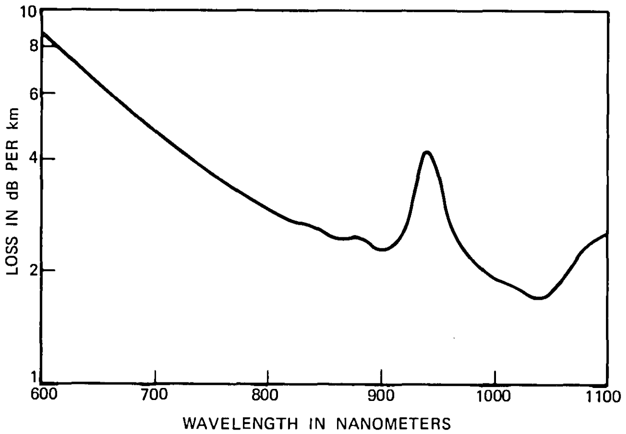


Fig. 3—Loss spectrum of a fiber waveguide with a graded-refractive-index core (B_2O_3–SiO_2) and borosilicate cladding.

case, both core and cladding were formed by CVD with simultaneous fusion inside a fused-quartz tube, which will be described in an article to be published in the near future. This fiber was over 0.5 km long and was characterized by an 18-μm-diameter core, a 15-μm cladding thickness, a 100-μm overall diameter, and a numerical aperture of 0.17. Loss minima occurred at 0.86, 0.90, and 1.02 μm. The average losses at these wavelengths were 1.9, 2.4, and 1.1 dB/km, respectively. The loss at 1.06 μm was 1.2 dB/km.

In addition to low loss, optical waveguides should exhibit low pulse dispersion so that high data rates can be achieved. One way to accomplish this is through the use of graded-refractive-index cores.[1] By gradually changing the concentration of reactive gases as the film thickness is built up, graded-refractive-index profiles can be achieved. This has been accomplished in the GeO_2-doped-core system by varying the concentration of $GeCl_4$ in the gas stream and in the silica-core, borosilicate-clad system by varying the concentration of BCl_3 during the deposition. The loss spectrum of a fiber in which the B_2O_3 concentration gradually changes from 0 at the center to about 20 percent at the core-cladding interface is presented in Fig. 3. This fiber had a 22-μm core diameter, a 15-μm cladding thickness, and a 0.17 numerical aperture. Minima occur in the loss spectrum at 0.90 and 1.04 μm. The losses at these wavelengths were 2.3 and 1.7 dB/km, respectively.

ACKNOWLEDGMENTS

The authors are grateful to J. R. Simpson and P. D. Lazay for performing the loss measurements for this work.

REFERENCES

1. D. Gloge and E. A. J. Marcatili, "Multimode Theory of Graded-Core Fibers," B.S.T.J., *52*, No. 9 (November 1973), pp. 1563–1578.

THE BELL SYSTEM TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering, Applied Mechanics Review, Applied Science & Technology Index, Chemical Abstracts, Computer Abstracts, Computer & Control Abstracts, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electrical & Electronic Abstracts, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Mathematical Reviews, Metals Abstracts, Science Abstracts,* and *Solid State Abstracts Journal.* Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.

Bell System